Radosław Adamczak

The entropy method and concentration of measure in product spaces

Master's thesis

Amsterdam, July 2003

This paper was written during the academic year 2002/2003. Its slightly modified versions served as the author's master's theses at the Warsaw University (under the title Koncentracja miary) and the Vrije University of Amsterdam within the Joint Master's Programme. The advisor in Poland was dr hab. Rafał Latała, the advisor in the Netherlands - prof. Aad van der Vaart.

Preface

Deviation inequalities, i.e. inequalities providing upper bounds on the quantities of the type $\mathbb{P}(|X - a| \geq t)$ (or $\mathbb{P}(X - a \geq t)$), where X is a random variable and a stands for the mean, median or some other parameter of X, have always been among the main tools of probability theory. The first, classical examples are the Markov and Chebyshev inequality. Although general, they are quite weak and mathematicians quite soon realized the need to provide stronger, exponential inequalities for special classes of random variables, of particular interest, first of all for sums of independent random variables. Inequalities such as the Bernstein inequality or its improved version by Bennett proved very useful for asymptotic analysis of sums of i.i.d. random variables, for instance in the proof of the law of the iterated logarithm. On the other hand in a more analytical setting some strong results for particular measures have been obtained, that link the deviation inequalities with isoperimetric issues and yield exponential inequalities for Lipschitz functions defined on special measure metric spaces. The most important example is probably the isoperimetric theorem for the uniform measure on the n-dimensional sphere, and as its consequence - the Gaussian isoperimetry. Such inequalities have been successfully used outside the classical probability theory, for example in the local theory of Banach spaces, where one of the most impressing results is the proof of Dvoretzky's theorem by V. Milman.

Those ideas have been recently replanted into the setting of general product spaces by M. Talagrand. Since the only natural distance in such spaces is the Hamming distance, which is not always useful, some other measures of distance between a point and a set have been introduced and the isoperimetrical theorems for such distances allowed to obtain deviation inequalities for much larger classes of functions of independent random variables, for example for convex functions. However, the method of proof, relying on the induction with respect to the number of coordinates, is quite technical and not always intuitive. An alternate method has been proposed by M. Ledoux and has been further developed among others by P. Massart. At its core there are estimations of entropy of a random variable (called sometimes a (modified) logarithmic Sobolev inequality), which together with a simple tensorization procedure lead to a useful bound on the entropy of a function of independent random variables, which can be interpreted as a differential inequality for the moment generating function and in some cases, via integration, yields an upper bound on the moment generating function, which can be in turn transformed into a deviation inequality. In the following chapters of the thesis, the author will introduce this method and present its various applications.

Chapter 1 contains the basic facts about Φ-entropies to be used in the sequel, e.g. their variational characterization and tensorization property. In Chapter 2 the basic entropy estimates are presented. The author presents also a refinement of these estimates in the special case of the discrete cube. The entropy bounds are used to derive many concentration inequalities from various branches of probability theory, for instance the bounded difference inequality, Talagrand's inequality for Rademacher chaos of order 2, tail inequalities for configuration functions and convex functions. As an application of the last inequality the author presents some inequalities for the eigenvalues of random matrices and Rademacher averages. Although the presented results have been already known, the use of the entropy method and the improved entropy bound for the discrete cube allow to simplify the proofs and/or improve the numerical constants. Chapter 3 is devoted to connections between discrete Sobolev inequalities, moment estimates and concentration of measure. The author introduces the moment method by deriving tail inequalities for some special functions of independent random variables with sub-Gaussian tails, e.g. suprema of empirical processes. The final part of the chapter presents the recent powerful moment inequality by Boucheron, Bosquet, Lugosi and Massart, which proof relies on tensorization property of Φ-entropy for some particular functions Φ. As an application, moment and tail inequalities for U-statistics in Banach spaces are proven.

The author would like to express his gratitude to the advisor at the Warsaw University, dr hab. Rafal Latała and the advisor at the Vrije Universiteit van Amsterdam, prof. Aad van der Vaart for introducing him to the subject, all the support and useful conversations.

Chapter 1

Entropy and tensorization

1.1. Basic assumptions and definitions

For a smooth convex function $\Phi\colon I \to \mathbb{R}$ (I - a closed interval of the real line) and a probability space $(\Omega, \mathcal{F}, \mu)$ let us consider a functional, defined on $\{X \in L^1(\mu)\colon X \in I \ a.e.,\ \mathbb{E}\Phi(X) < \infty\}$ with the formula

$$E_{\Phi,\mu}(X) := \mathbb{E}\Phi(X) - \Phi(\mathbb{E}X).$$

Let us notice that from the convexity of Φ it follows that $E_{\Phi,\mu}(X)$ is non-negative for every X from the domain of $E_{\Phi,\mu}$. Moreover, the domain is a convex subset of $L^1(\mu)$.
In the following part of this chapter we will restrict our attention to functions Φ, such that $E_{\Phi,\mu}$ is a convex functional for every probability space $(\Omega, \mathbb{F}, \mu)$, i.e.

$$E_{\Phi,\mu}(pX + (1-p)Y) \leq pE_{\Phi,\mu}(X) + (1-p)E_{\Phi,\mu}(Y) \tag{1.1}$$

for every $p \in [0,1]$.

1.1.1. Entropy

The most important example of a functional obtained by the above definition is the so-called entropy functional, corresponding to the function $\Phi(x) = x\log x$. As $\lim_{x\to 0} x\log x = 0$, we can consider here Φ as a function defined on $[0,\infty)$. We will denote $E_{\Phi,\mu}X$ by $\mathrm{Ent}_\mu X$. Entropy satisfies the condition (1.1), since

$$\mathrm{Ent}X := \mathbb{E}X\log X - \mathbb{E}X\log\mathbb{E}X = \sup\{\mathbb{E}XY\colon \mathbb{E}e^Y \leq 1\}. \tag{1.2}$$

Indeed, consider a random variable Y, satisfying $\mathbb{E}e^Y \leq 1$. We will show at first that $\mathrm{Ent}X \geq \mathbb{E}XY$. We can assume that $\mathrm{Ent}X < \infty$. Let us also assume for a while that $X > 0$. We have

$$\begin{aligned}
\mathbb{E}XY - \mathrm{Ent}X &= \mathbb{E}X\log(e^Y) - \mathbb{E}X\log\frac{X}{\mathbb{E}X}\\
&= \mathbb{E}X\log(e^Y\frac{\mathbb{E}X}{X})\\
&= \mathbb{E}X\cdot\mathbb{E}\log(e^Y\frac{\mathbb{E}X}{X})\frac{X}{\mathbb{E}X}\\
&\leq \mathbb{E}X\cdot\log\mathbb{E}e^Y \leq 0,
\end{aligned}$$

where the first inequality follows from Jensen's inequality applied to the probability measure with density $X/\mathbb{E}X$ and the function log.
To obtain the same inequality for arbitrary random variable X we approximate X by $X_n = X+\frac{1}{n}\mathbf{1}_{\{X=0\}}$. Random variables X_n are strictly positive, so they satisfy the desired inequality. Moreover, it is not hard to notice that $\lim_{n\to\infty} \mathrm{Ent}X_n = \mathrm{Ent}X$ and $\lim_{n\to\infty} \mathbb{E}X_nY = \mathbb{E}XY$. The inequality has thus been proved.
To complete the proof of (1.2) it suffices to find a sequence of random variables Y_n with $\mathbb{E}e^{Y_n} \leq 1$ and $\lim_{n\to\infty} \mathbb{E}XY_n = \mathrm{Ent}X$. Define

$$Y_n = \begin{cases} \log(\frac{X}{\mathbb{E}X}) - \frac{1}{n} & \text{if} \quad X > 0 \\ M_n & \text{if} \quad X = 0, \end{cases}$$

where M_n is a number such that $e^{-\frac{1}{n}} + e^{M_n} \leq 1$. We have

$$\mathbb{E}e^{Y_n} = \mathbb{E}e^{-\frac{1}{n}}\frac{X}{\mathbb{E}X}\mathbf{1}_{\{X>0\}} + e^{M_n}\mathbb{P}(X=0) \leq e^{-\frac{1}{n}} + e^{M_n} \leq 1.$$

Moreover

$$\begin{aligned} \mathbb{E}XY_n &= \mathbb{E}X\log(\frac{X}{\mathbb{E}X})\mathbf{1}_{\{X>0\}} - \frac{1}{n}\mathbb{E}X\mathbf{1}_{\{X>0\}} + \mathbb{E}XM_n\mathbf{1}_{\{X=0\}} \\ &= \mathbb{E}X\log(X)\mathbf{1}_{\{X>0\}} - \mathbb{E}X\log(\mathbb{E}X)\mathbf{1}_{\{X>0\}} - \frac{1}{n}\mathbb{E}X \\ &= \mathbb{E}X\log X - \mathbb{E}X\log\mathbb{E}X - \frac{1}{n}\mathbb{E}X \\ &= \mathrm{Ent}X - \frac{1}{n}\mathbb{E}X, \end{aligned}$$

so indeed $\lim_{n\to\infty} \mathbb{E}XY_n = \mathrm{Ent}X$.

□

There is also another variational characterization of entropy, we will use in the sequel, namely

$$\mathrm{Ent}X = \inf_{u>0} \mathbb{E}(X(\log X - \log u) - (X-u)) \tag{1.3}$$

for any nonnegative random variable X, such that $\mathbb{E}X\log X < \infty$.
To prove (1.3) it suffices to notice that for $x = \mathbb{E}X \geq 0$ the function $f(u) = -x\log u - x + u$ attains its minimum on $\mathbb{R}^+$ at $u = x$.

□

1.1.2. Variance

Another important and well-known example is the variance of a random variable, which corresponds to the function $\Phi(x) = x^2$. It is easy to check that the condition (1.1) is satisfied:

$$\begin{aligned} \mathrm{Var}(pX+(1-p)Y) &= \mathbb{E}(p(X-\mathbb{E}X)+(1-p)(Y-\mathbb{E}Y))^2 \\ &\leq \mathbb{E}(p(X-\mathbb{E}X)^2+(1-p)(Y-\mathbb{E}Y)^2) \\ &= p\mathrm{Var}X + (1-p)\mathrm{Var}Y, \end{aligned}$$

where to get the inequality in the second line we used Jensen's inequality.

1.1.3. Further examples

The following Theorem from [12] generalizes the above examples.

Theorem 1 *If $\Phi\colon I \to \mathbb{R}$ is a twice differentiable function, such that Φ'' is strictly positive in* $\mathrm{int} I$ *and $1/\Phi''$ is concave, then Φ satisfies the condition (1.1).*

Proof. For $p \in [0,1]$, let us define the function $F_p\colon I^2 \to \mathbb{R}$ with the formula

$$F_p(x,y) = p\Phi(x) + (1-p)\Phi(y) - \Phi(px + (1-p)y).$$

From the convexity of Φ it follows that F_p is nonnegative. We claim that F_p is convex on I^2. Since F_p is continuous on I^2 and twice differentiable in $\mathrm{int} I^2$, it is enough to show that the matrix of second order derivatives is positively definite. We have

$$\frac{\partial^2 F_p}{\partial x^2}(x,y) = p\Phi''(x) - p^2\Phi''(px + (1-p)y) \geq 0,$$

since by the concavity of $1/\Phi''$ we have

$$\frac{1}{\Phi''(px + (1-p)y)} \geq \frac{p}{\Phi''(x)} + \frac{1-p}{\Phi''(y)} \geq \frac{p}{\Phi''(x)}.$$

Similarly $\frac{\partial^2 F_p}{\partial y^2}(x,y) \geq 0$.
To complete the proof of convexity of F_p, it is enough to show that $\det \mathrm{Hess}(F_p) \geq 0$, or equivalently

$$\frac{\partial^2 F_p}{\partial x^2}(x,y) \cdot \frac{\partial^2 F_p}{\partial y^2}(x,y) \geq \left(\frac{\partial^2 F_p}{\partial x \partial y}(x,y)\right)^2.$$

After computing the mixed derivative, we see that the above inequality is equivalent to

$$(p\Phi''(x) - p^2\Phi''(px+(1-p)y))((1-p)\Phi''(y) - (1-p)^2\Phi''(px+(1-p)y)) \geq (p(1-p)\Phi''(px+(1-p)y))^2$$

or

$$\Phi''(x)\Phi''(y) \geq p\Phi''(px + (1-p)y)\Phi''(y) + (1-p)\Phi''(px + (1-p)y)\Phi''(x).$$

But since Φ'' is strictly positive, this is equivalent to the concavity of $1/\Phi''$, which shows the convexity of F_p.
Let now X, Y be two random variables in the domain of $E_{\Phi,\mu}$. Define $x_0 = \mathbb{E}X$, $y_0 = \mathbb{E}Y$. From the convexity of F_p it follows that there exists $a, b, c \in \mathbb{R}$ (depending on p), such that

$$\begin{aligned} F_p(x_0, y_0) &= ax_0 + by_0 + c, \\ F_p(x,y) &\geq ax + by + c \end{aligned}$$

for all $x, y \in I$. Thus

$$\mathbb{E}F_p(X,Y) \geq \mathbb{E}(aX + bY + c) = ax_0 + by_0 + c = F_p(x_0, y_0),$$

which is equivalent to (1.1).

□

Example. From the above theorem, it follows that for all $\alpha \in (1,2]$, the function $\Phi_\alpha(x) = x^\alpha$, defined on $I = [0,\infty)$, satisfies the condition (1.1).

1.2. Properties of $E_{\Phi,\mu}$

The condition (1.1) implies the following generalization of the formula (1.2).

Theorem 2 *Let $\Phi\colon I \to \mathbb{R}$ be a differentiable, convex function, satisfying the condition (1.1). Assume that X is an integrable random variable, such that $\mathbb{E}\Phi(X) < \infty$.*

$$E_{\Phi,\mu}(X) = \sup_{\substack{Y\colon \Omega\to intI\\ Y\in L^1, \mathbb{E}\Phi(Y)<\infty}} \{\mathbb{E}(\Phi'(Y) - \Phi'(\mathbb{E}Y))(X-Y) + E_{\Phi,\mu}(Y)\}. \tag{1.4}$$

Before we proceed with the proof of the above theorem, let us state the following

Lemma 1 *Let $\varphi\colon [x, x+\varepsilon) \to \mathbb{R}$ be a smooth convex function. Then*

$$\lim_{h\to 0+} h\varphi'(x+h) = 0.$$

Proof. For every $h \in (0,\varepsilon)$ we have

$$\frac{\varphi(x+2h)-\varphi(x+h)}{h} \geq \varphi'(x+h) \geq \frac{\varphi(x+h)-\varphi(x)}{h}$$

or equivalently

$$\varphi(x+2h)-\varphi(x+h) \geq h\varphi(x+h) \geq \varphi(x+h)-\varphi(x).$$

The lemma thus follows by the continuouity of φ.

□

Proof of Theorem 2. First we will prove that

$$E_{\Phi,\mu}(X) \geq \mathbb{E}(\Phi'(Y) - \Phi'(\mathbb{E}Y))(X-Y) + E_{\Phi,\mu}(Y). \tag{1.5}$$

Assume temporarily that the values of X and Y are separated from the ends of the interval I. By (1.1) the function $\varphi\colon [0,1] \to \mathbb{R}$, defined as

$$\varphi(t) = E_{\Phi,\mu}(X + t(Y-X))$$

is convex. Thus

$$E_{\Phi,\mu}(X) = \varphi(0) \geq \varphi(1) - \varphi'(1).$$

But

$$\begin{aligned}\varphi'(t) &= \mathbb{E}\Phi'(X+t(Y-X))\cdot(Y-X) - \Phi'(\mathbb{E}X + t\mathbb{E}(Y-X))\cdot\mathbb{E}(Y-X)\\ &= \mathbb{E}(\Phi'(X+t(Y-X)) - \Phi'(\mathbb{E}X+t\mathbb{E}(Y-X)))(Y-X)\end{aligned}$$

and thus

$$E_{\Phi,\mu}(X) \geq \mathbb{E}(\Phi'(Y) - \Phi'(\mathbb{E}Y))(X-Y) + E_{\Phi,\mu}(Y). \tag{1.6}$$

Let now $a_n, b_n \in \text{int}I$ be monotone sequences converging respectively to the left and right end of I, with $a_1 = b_1$. Define $X_n = \min(\max(X,a_n),b_n)$, $Y_n = \min(\max(Y,a_n),b_n)$. By (1.6) we have

$$E_{\Phi,\mu}(X_n) \geq \mathbb{E}(\Phi'(Y_k) - \Phi'(\mathbb{E}Y_k))(X_n - Y_k) + E_{\Phi,\mu}(Y_k)$$

or equivalently

$$\mathbb{E}(\Phi(X_n) - \Phi(Y_k) - \Phi'(Y_k)(X_n - Y_k)) \geq -\Phi'(\mathbb{E}Y_k)\mathbb{E}(X_n - Y_k) - \Phi(\mathbb{E}Y_k) + \Phi(\mathbb{E}X_n). \quad (1.7)$$

Let us consider the left-hand side of (1.7). It is of the form $\mathbb{E}\Psi(X_n, Y_k)$ with $\Psi(x, y) = \Phi(x) - \Phi(y) - (x - y)\Phi'(y)$. Note that by the convexity of Φ, we have $\Psi \geq 0$. We will prove that

$$\lim_{k\to\infty} \lim_{n\to\infty} \mathbb{E}\Psi(X_n, Y_k) = \mathbb{E}\Psi(X, Y), \quad (1.8)$$

provided that $\Psi(X, Y)$ is integrable. It will finish the proof of the desired inequality, since the analogous limit of the right-hand side of (1.7) equals $-\Phi'(\mathbb{E}Y)\mathbb{E}(X - Y) - \Phi(\mathbb{E}Y) + \Phi(\mathbb{E}X)$ and

$$\mathbb{E}\Psi(X, Y) \geq -\Phi'(\mathbb{E}Y)\mathbb{E}(X - Y) - \Phi(\mathbb{E}Y) + \Phi(\mathbb{E}X)$$

is equivalent to (1.5) (in the case $\mathbb{E}\Psi(X, Y) = \infty$ the above inequality is obvious). Let us now notice that

$$\begin{aligned} \frac{\partial}{\partial x}\Psi(x, y) &= \Phi'(x) - \Phi'(y) \\ \frac{\partial}{\partial y}\Psi(x, y) &= -x\Phi''(y) + y\Phi''(y), \end{aligned}$$

so (since, by the convexity of Φ, the function Φ' is nondecreasing and Φ'' is nonnegative) we see that

- for any fixed $x \in I$ the function $y \mapsto \Psi(x, y)$ is decreasing for $y \leq x$ and increasing for $y \geq x$,
- for any fixed $y \in I$ the function $x \mapsto \Psi(x, y)$ is decreasing for $x \leq y$ and increasing for $x \geq y$.

The first property implies that for every x, $\Psi(x, Y_k) \leq \Psi(x, a_1) + \Psi(x, Y)$. Indeed, consider the case $x \geq a_1$. If $Y \leq a_1$ then $Y_k = \max(Y, a_k) \geq Y$, so $\Psi(x, Y) \geq \Psi(x, Y_k)$. If $Y \in (a_1, x)$ then $Y_k = \min(Y, b_k) \geq a_1$, so $\Psi(x, a_1) \geq \Psi(x, Y_k)$. If $Y \geq x$ then $Y_k = \min(Y, b_k) \leq Y$, so $\Psi(x, Y) \geq \Psi(x, Y_k)$. The case $x < a_1$ is similar.
By analogy, from the second property of the function Ψ it follows that $\Psi(X_n, y) \leq \Psi(a_1, y) + \Psi(X, y)$ for every $y \in I$. Thus, for fixed k, we have for every n

$$\Psi(X_n, Y_k) \leq \Psi(a_1, Y_k) + \Psi(X, Y_k) \leq \Psi(a_1, Y_k) + \Psi(X, a_1) + \Psi(X, Y).$$

Now, since Y_k is separated from the boundary of I and $\Psi(X, Y)$,$\Psi(X, a_1)$ are integrable, by the Lebesgue dominated convergence theorem we obtain that

$$\lim_{n\to\infty} \mathbb{E}\Psi(X_n, Y_k) = \mathbb{E}\Psi(X, Y_k).$$

But

$$\Psi(X, Y_k) \leq \Psi(X, a_1) + \Psi(X, Y)$$

and (as by assumption $\mathbb{E}\Phi(X) < \infty$) the right hand side is integrable, so again

$$\lim_{k\to\infty} \mathbb{E}\Psi(X, Y_k) = \mathbb{E}\Psi(X, Y),$$

which proves (1.5).

It remains to show that $E_{\Phi,\mu}$ is indeed the supremum of expressions considered at the right hand side of (1.5). It is obvious if the random variable X takes values in the interior of I, since the supremum is then obtained for $Y = X$. In the general case we construct a sequence Y_n of random variables such that

$$\lim_{n\to\infty} E_{\Phi,\mu}(Y_n) = E_{\Phi,\mu}(X) \quad \text{and} \quad \lim_{n\to\infty} \mathbb{E}(\Phi'(Y_n) - \Phi'(\mathbb{E}Y_n))(X - Y_n) = 0.$$

Let a and b denote respectively the left and right end of the interval I. Define

$$Y_n = X + \frac{1}{n}\mathbf{1}_{\{X=a\}} - \frac{1}{n}\mathbf{1}_{\{X=b\}}.$$

Let us notice that in the case $a = -\infty$ (resp. $b = \infty$) we have $\{X = a\} = \emptyset$ (resp. $\{X = b\} = \emptyset$). The sequence Y_n converges uniformly to X and $\Phi(Y_n)$ converges uniformly to $\Phi(X)$. Thus indeed $\lim_{n\to\infty} E_{\Phi,\mu}(Y_n) = E_{\Phi,\mu}(X)$. Moreover

$$\begin{aligned}\mathbb{E}(\Phi'(Y_n) - \Phi'(\mathbb{E}Y_n))(X - Y_n) &= \Phi'\left(a + \frac{1}{n}\right) \cdot \frac{1}{n}\Pr(X = a) + \Phi'\left(b - \frac{1}{n}\right) \cdot \frac{1}{n}\Pr(X = b) \\ &- \Phi'(\mathbb{E}Y_n)(\mathbb{E}X - \mathbb{E}Y_n)\end{aligned}$$

and by Lemma 1 the right-hand side converges to 0 as $n \to \infty$.

□

Corollary 1 *Let $\Omega = \Omega_1 \times \Omega_2$ be a product space equipped with a product probability measure $\mu = \mu_1 \otimes \mu_2$. For every integrable $X\colon \Omega \to I$ with $\mathbb{E}\Phi(X) < \infty$ we have*

$$E_{\Phi,\mu_2}(\mathbb{E}_{\mu_1}X) \le \mathbb{E}_{\mu_1}E_{\Phi,\mu_2}(X),$$

where $E_{\Phi,\mu_2}(X)$ denotes the value of the functional E_{Φ,μ_2} at the function $\omega_2 \mapsto X(\omega_1, \omega_2)$ with the first coordinate fixed.

Proof. By Theorem 2 we have

$$\begin{aligned}E_{\Phi,\mu_2}(\mathbb{E}_{\mu_1}X) &= \sup_{\substack{Y\colon \Omega_2\to \mathrm{int} I \\ Y\in L^1, \mathbb{E}\Phi(Y)<\infty}} \left\{\mathbb{E}(\Phi'(Y) - \Phi'(\mathbb{E}_{\mu_2}Y))(\mathbb{E}_{\mu_1}X - Y) + E_{\Phi,\mu_2}(Y)\right\} \\ &= \sup_{\substack{Y\colon \Omega_2\to \mathrm{int} I \\ Y\in L^1, \mathbb{E}\Phi(Y)<\infty}} \left\{\mathbb{E}_{\mu_1}\mathbb{E}(\Phi'(Y) - \Phi'(\mathbb{E}_{\mu_2}Y))(X - Y) + E_{\Phi,\mu_2}(Y)\right\} \\ &\le \mathbb{E}_{\mu_1} \sup_{\substack{Y\colon \Omega_2\to \mathrm{int} I \\ Y\in L^1, \mathbb{E}\Phi(Y)<\infty}} \left\{\mathbb{E}_{\mu_2}(\Phi'(Y) - \Phi'(\mathbb{E}_{\mu_2}Y))(X - Y) + E_{\Phi,\mu_2}(Y)\right\} \\ &= \mathbb{E}_{\mu_1}E_{\Phi,\mu_2}(X).\end{aligned}$$

□

The following theorem describes the basic property of functionals $E_{\Phi,\mu}$, which we will call *the tensorization property*.

Theorem 3 *Consider a product probability space (Ω, μ), where $\Omega = \Omega_1 \times \Omega_2 \times \ldots \times \Omega_n$ and $\mu = \mu_1 \otimes \mu_2 \otimes \ldots \otimes \mu_n$. Then for every function X in the domain of $E_{\Phi,\mu}$ we have*

$$E_{\Phi,\mu}(X) \le \sum_{i=1}^{n} \mathbb{E}\, E_{\Phi,\mu_i}(X),$$

where $E_{\Phi,\mu_i}(X)$ denotes the value of the functional E_{Φ,μ_i} at the function X, considered as a function of ω_i, with the other coordinates fixed.

Proof. We will proceed by the induction with respect to n. For $n = 1$ the theorem is trivial. Assume it is true for some n and consider $\mu = \mu_1 \otimes \ldots \otimes \mu_{n+1}$ and a random variable X in the domain of $E_{\Phi,\mu}$. We have by the induction assumption

$$\mathbb{E}\Phi(X) = \mathbb{E}_{\mu_{n+1}}\mathbb{E}_{\mu_1 \otimes \ldots \otimes \mu_n}\Phi(X) \leq \mathbb{E}_{\mu_{n+1}}\left(\Phi(\mathbb{E}_{\mu_1 \otimes \ldots \otimes \mu_n}X) + \mathbb{E}_{\mu_1 \otimes \ldots \otimes \mu_n}\sum_{i=1}^{n} E_{\Phi,\mu_i}(X)\right).$$

Thus it is enough to show that

$$\mathbb{E}_{\mu_{n+1}}\Phi(\mathbb{E}_{\mu_1 \otimes \ldots \otimes \mu_n}X) \leq \Phi(\mathbb{E}X) + \mathbb{E}E_{\Phi,\mu_{n+1}}X$$

or equivalently

$$E_{\Phi,\mu_{n+1}}(\mathbb{E}_{\mu_1 \otimes \ldots \otimes \mu_n}X) \leq \mathbb{E}_{\mu_1 \otimes \ldots \otimes \mu_n}E_{\Phi,\mu_{n+1}}X.$$

But this is true due to Corollary 1.

□

Chapter 2

Logarithmic Sobolev inequalities

2.1. Basic inequalities

Let us start this chapter with the following theorem

Theorem 4 *Let $X_1, \ldots, X_n$ be independent random variables taking values in a measurable space $(\Sigma, \mathcal{F})$ and $f\colon \Sigma^n \to \mathbb{R}$ a measurable function (with respect to the product σ-field). Denote $S = f(X_1, \ldots, X_n)$, $S_i = f(X_1, \ldots, X_{i-1}, \tilde{X}_i, X_{i+1}, \ldots, X_n)$, where $(X_1, \ldots, X_n)$ and $(\tilde{X}_1, \ldots, \tilde{X}_n)$ are independent random vectors, equal in distribution. Let us also assume that $\mathbb{E}Se^S < \infty$. Then the following inequality holds*

$$\mathrm{Ent}\ e^S \leq \mathbb{E}(e^S \sum_{i=1}^{n} (S - S_i)_+^2). \tag{2.1}$$

Proof. Consider X, Y - i.i.d. real random variables. From Jensen's inequality we have

$$\log \mathbb{E}e^X \geq \mathbb{E} \log e^X = \mathbb{E}X. \tag{2.2}$$

Applying this inequality in the definition of entropy we immediately get

$$\begin{aligned} \mathrm{Ent}\ e^X &= \mathbb{E}e^X(X - \log \mathbb{E}e^X) && (2.3)\\ &\leq \mathbb{E}e^X(X - \mathbb{E}X) && (2.4)\\ &= \mathbb{E}e^X(X - Y) && (2.5)\\ &= \frac{1}{2}\mathbb{E}(e^X - e^Y)(X - Y). && (2.6) \end{aligned}$$

But for $x, y \in \mathbb{R}$ we have

$$(x - y)(e^x - e^y) \leq (x - y)_+^2 e^x + (y - x)_+^2 e^y, \tag{2.7}$$

so

$$\mathrm{Ent}\ e^X \leq \frac{1}{2}\mathbb{E}(e^X(X - Y)_+^2 + e^Y(Y - X)_+^2) = \mathbb{E}e^X(X - Y)_+^2, \tag{2.8}$$

which is exactly (2.1) in dimension 1. A direct use of Theorem 3 allows us to finish the proof.

□

2.1.1. Deviation inequalities

Theorem 4 can be applied to derive in an easy way concentration inequalities for a wide class of random variables, via the so called *Herbst argument.* The main idea is to transform (2.1) into a differential inequality for the Laplace transform of a random variable. To show how it works in practice we will prove the following fact:

Lemma 2 (Herbst argument) *Let S be a random variable, such that for every $\lambda \geq 0$ $\mathbb{E}e^{\lambda S} < \infty$. If $c \in \mathbb{R}$ is such that*

$$\mathrm{Ent}e^{\lambda S} \leq c\lambda^2 \mathbb{E}e^{\lambda S}$$

for all $\lambda \geq 0$, then for all $t \geq 0$ we have

$$\mathbb{P}(S - \mathbb{E}S \geq t) \leq e^{-\frac{t^2}{4c}}.$$

Proof. Define $F(\lambda) = \mathbb{E}e^{\lambda S}$ and $\psi(\lambda) = \log F(\lambda)$. Notice that $F(0) = 1$, $\psi(0) = 0$ and $F'(\lambda) = \mathbb{E}Se^{\lambda S}$. Thus, according to the assumption, we have

$$\lambda F'(\lambda) - F(\lambda)\log F(\lambda) \leq c\lambda^2 F(\lambda)$$

or, taking advantage of the fact that $F(\lambda) > 0$

$$\frac{\lambda\psi'(\lambda) - \psi}{\lambda^2} \leq c,$$

that is

$$\left(\frac{\psi(\lambda)}{\lambda}\right)' \leq c.$$

We also have

$$\lim_{\lambda \to 0+} \frac{\psi(\lambda)}{\lambda} = \psi'(0) = \frac{F'(0)}{F(0)} = \frac{\mathbb{E}S}{1} = \mathbb{E}S,$$

and so

$$\frac{\psi(\lambda)}{\lambda} \leq \mathbb{E}S + c\lambda,$$

for all $\lambda > 0$, which can be reformulated as

$$\log \mathbb{E}e^{\lambda(S - \mathbb{E}S)} \leq c\lambda^2.$$

Now we can apply Markov inequality to obtain

$$\mathbb{P}(S - \mathbb{E}S \geq t) \leq \inf_{\lambda > 0} \frac{\mathbb{E}e^{\lambda(S-\mathbb{E}S)}}{e^{\lambda t}} \leq \inf_{\lambda > 0} e^{c\lambda^2 - \lambda t} = e^{-\frac{t^2}{4c}}.$$

□

Corollary 2 *Let S, S_i be defined as in* Theorem 4*. Denote $c = ||\sum_{i=1}^n (S - S_i)_+^2||_\infty$. Then for every $t > 0$ we have*

$$\mathbb{P}(S - \mathbb{E}S \geq t) \leq e^{-\frac{t^2}{4c}}. \tag{2.9}$$

Proof. Denote $V_+ = \sum_{i=1}^{n}(S - S_i)_+^2$ and assume that V_+ is bounded (otherwise the statement is obvious). Inequality (2.1) for a random variable λS ($\lambda \geq 0$) may be rewritten as

$$\mathrm{Ent}\, e^{\lambda S} \leq \mathbb{E}\lambda^2 V_+ e^{\lambda S}, \tag{2.10}$$

which implies

$$\mathrm{Ent}\, e^{\lambda S} \leq c\lambda^2 \mathbb{E} e^{\lambda S}.$$

The statement to be proven follows now from Lemma 2.

□

2.1.2. Bounded difference inequality

Corollary 2 allows us to derive (up to constants) the well-known *bounded difference inequality* due to McDiarmid (cf. [17]).

Corollary 3 *With the notation of Theorem 4, if there exist constants c_i such that*

$$|S - S_i| \leq c_i \quad i = 1, \ldots, n,$$

then for all $t \geq 0$

$$\mathbb{P}(|S - \mathbb{E}S| \geq t) \leq 2e^{-\frac{t^2}{4\sum_{i=1}^{n} c_i^2}}.$$

Remark Actually the constant 4 in the exponent may be replaced by 1/2, as it may be proved with the so-called *martingale method* (comp. [17]).

2.2. Discrete cube

The probability space we will consider in the following part is the discrete cube $\Omega = \{-1, +1\}^n$ with the uniform probability measure. Our purpose is to obtain an improvement of Theorem 4 for this particular probability space. On the way we will also prove the Gross' logarithmic Sobolev inequality and use it to derive concentration inequalities for Gaussian measures.

Lemma 3 *For all $x \geq y > 0$*

$$\log(\frac{x^2 + y^2}{2x^2}) \geq \frac{y - x}{x}.$$

Proof. The function $f(t) = \log(\frac{1+t^2}{2}) - t + 1$ satisfies $f(1) = 0$ and for all t

$$f'(t) = \frac{2t}{1 + t^2} - 1 = -\frac{(1 - t)^2}{1 + t^2} \leq 0.$$

Thus f is nonincreasing and in consequence $f(t) \geq 0$ for $t \leq 1$. In particular $f(y/x) \geq 0$, which means

$$\log(\frac{x^2 + y^2}{2x^2}) \geq \frac{y - x}{x}.$$

□

Lemma 4 *For every $x, y \in \mathbb{R}$*

$$x^2 \log x^2 + y^2 \log y^2 - (x^2 + y^2) \log(\frac{x^2 + y^2}{2}) \leq (x - y)^2 \tag{2.11}$$

Proof. Without loss of generality we can assume that $x \geq y \geq 0$. For a fixed y let $f(x)$, $g(x)$ denote respectively the left and the right hand side of (4) as a function of x. Since $f(y) = g(y) = 0$, to prove (4) it is enough to show that $f'(x) \leq g'(x)$ for all $x \geq y$. But

$$\begin{aligned} f'(x) &= 2x \log(\frac{2x^2}{x^2 + y^2}) \\ g'(x) &= 2(x - y) \end{aligned}$$

so the desired claim follows from Lemma 3.

□

Lemma 5 *For every $x \geq y > 0$,*

$$\log x - \log y \geq 2 \cdot \frac{x - y}{x + y}$$

Proof. Consider the function $f(t) = \log t - 2\frac{t-1}{t+1}$, $t > 0$. We have

$$f'(t) = \frac{1}{t} - \frac{4}{(t+1)^2} = \frac{(t-1)^2}{t(t+1)^2} \geq 0.$$

Moreover $f(1) = 0$. Thus for every $t > 1$, $f(t) > 0$ and therefore

$$\log t \geq \frac{2(t-1)}{t+1}.$$

Now it is enough to substitute $t = x/y$.

□

Definition 1 *For $f \colon \Omega \to \mathbb{R}$ and $x = (x_1, \ldots, x_n) \in \Omega$ let us define the* discrete gradient *of f in x along the i-th coordinate as*

$$D_i f(x) = f(x) - f(s_i(x)),$$

where $s_i(x) = (x_1, \ldots, x_{i-1}, -x_i, x_{i+1}, \ldots, x_n)$.

Theorem 5 (Gross' logarithmic Sobolev inequality) *For every $f \colon \Omega \to \mathbb{R}$ the following inequalities hold*

(i)

$$\mathrm{Ent} f^2 \leq \frac{1}{2} \sum_{i=1}^{n} \mathbb{E}|D_i f|^2,$$

(ii)

$$\mathrm{Ent}\ e^f \leq \frac{1}{8} \sum_{i=1}^{n} \mathbb{E} e^f |D_i f|^2.$$

Proof. From the tensorization property of entropy, it is enough to prove the theorem for $n = 1$, which corresponds to $\Omega = \{-1, +1\}$.
Denote $f(1) = x$, $f(-1) = y$. Then

$$\mathrm{Ent} f^2 = \mathbb{E} f^2(\log f^2 - \mathbb{E}\log f^2) = \frac{1}{2}(x^2 \log x^2 + y^2 \log y^2 - (x^2 + y^2)\log(\frac{x^2+y^2}{2})).$$

On the other hand

$$\mathbb{E}|D_1 f|^2 = (x-y)^2,$$

so the part *(i)* follows from Lemma 4.
To prove the second part of the theorem, let us denote $g = e^{f/2}$. From part *(i)* we have

$$\mathrm{Ent} e^f = \mathrm{Ent} g^2 \leq \frac{1}{2}\mathbb{E}|D_1 g|^2 = \frac{1}{2}(e^{x/2} - e^{y/2})^2. \tag{2.12}$$

We can assume that $x > y$. From Lemma 5 we have

$$e^{x/2} - e^{y/2} \leq \frac{e^{x/2} + e^{y/2}}{2}(\frac{x}{2} - \frac{y}{2}) \leq \frac{1}{2} \cdot \sqrt{\frac{e^x + e^y}{2}}(x - y).$$

Hence

$$\frac{1}{2}\mathbb{E}|D_1 g|^2 \leq \frac{1}{8}\frac{e^x + e^y}{2}(x-y)^2 = \frac{1}{8}\mathbb{E}e^f|D_1 f|^2,$$

which together with (2.12) proves the one-dimensional version of part *(ii)*.

□

Corollary 4 *Let γ_d be the standard Gaussian measure on $\mathbb{R}^d$ i.e. the measure with density $g(x) = \frac{1}{(2\pi)^{d/2}}e^{-(x_1^2+\ldots+x_d^2)/2}$. Then for every smooth enough (e.g. Lipschitz continuous) function $f\colon \mathbb{R}^d \to \mathbb{R}$ the following statements are satisfied*

$$\mathrm{Ent}_{\gamma_d} f^2 \leq 2\int_{\mathbb{R}^d} |\nabla f|^2 d\gamma_d \tag{2.13}$$

$$\mathrm{Ent}_{\gamma_d} e^f \leq \frac{1}{2}\int_{\mathbb{R}^n} e^f|\nabla f|^2 d\gamma_d. \tag{2.14}$$

In consequence for every 1-Lipschitz function f and every $t \geq 0$

$$\gamma_d\left(\left\{f \geq \int_{\mathbb{R}^d} f d\gamma_d + t\right\}\right) \leq e^{-t^2/2}. \tag{2.15}$$

Proof. It is enough to prove the Corollary for C^∞ functions with compact support. Then, using a standard approximation technique, one can extend it to some more general classes of functions, e.g. for Lipschitz functions, which by the Rademacher Theorem are almost everywhere differentiable .
Let us also notice that we can focus on the first inequality, since (2.14) follows easily from (2.13) by substituting $e^{f/2}$ as the "new" function f. Moreover, the tensorization property of entropy allows us to restrict the proof to $d = 1$.
Consider a sequence of independent Rademacher variables $(\varepsilon_i)_{i=1}^\infty$ and random variables

$$S_n = f(\frac{\varepsilon_1 + \ldots + \varepsilon_n}{\sqrt{n}}).$$

By the Central Limit Theorem $\lim_{n\to\infty} \mathrm{Ent} S_n^2 = \mathrm{Ent}_{\gamma_1} f^2$ and $\lim_{n\to\infty} \mathbb{E} f'(\frac{\varepsilon_1+\ldots+\varepsilon_n}{\sqrt{n}})^2 = \mathbb{E}_{\gamma_1}(f')^2$. Theorem 5 implies that

$$\mathrm{Ent} S_n^2 \leq \frac{1}{2}\mathbb{E}\sum_{i=1}^n \left(f\left(\frac{\varepsilon_1+\ldots+\varepsilon_n}{\sqrt{n}}\right) - f\left(\frac{\varepsilon_1+\ldots+\varepsilon_n}{\sqrt{n}} - 2\frac{\varepsilon_i}{\sqrt{n}}\right)\right)^2 .$$

But each component of the sum at the right hand side is equal to $4n^{-1} f'(\frac{\varepsilon_1+\ldots+\varepsilon_n}{\sqrt{n}})^2 + \mathcal{O}(n^{-3/2})$ and thus taking the limits with $n\to\infty$ yields exactly (2.13).
It remains to show the deviation inequality (2.15). But it is a direct consequence of the inequality (2.14) and Lemma 2, since the Lipschitz condition guarantees that $\mathbb{E}e^{\lambda f} < \infty$ for every $\lambda \geq 0$.

□

Another consequence of Theorem 5 is a refinement of Theorem 4 in the special case of Rademacher variables.

Corollary 5 *Consider independent Rademacher variables $\varepsilon_1,\ldots,\varepsilon_n,\tilde{\varepsilon}_1,\ldots,\tilde{\varepsilon}_n$ and a function $f\colon \{-1,+1\}^n \to \mathbb{R}$. Denote $S = f(\varepsilon_1,\ldots,\varepsilon_n)$, $S_i = f(\varepsilon_1,\ldots,\varepsilon_{i-1},\tilde{\varepsilon}_i,\varepsilon_{i+1},\ldots,\varepsilon_n)$. Then*

$$\mathrm{Ent}\; e^S \leq \frac{1}{2}\mathbb{E}(e^S \sum_{i=1}^n (S-S_i)_+^2). \tag{2.16}$$

Proof. Let us consider the crucial case $n=1$. Denote $f(-1)=a$, $f(1)=b$ and assume that $a\geq b$. From Theorem 5 we have

$$\mathrm{Ent}\; e^S \leq \frac{1}{8}\frac{e^a+e^b}{2}(a-b)^2 \leq \frac{1}{8}e^a(a-b)^2 = \frac{1}{2}\cdot\frac{1}{4}e^a(a-b)^2 = \frac{1}{2}\mathbb{E}(S-S_1)_+^2 e^S$$

□

As an example of application of Corollary 5 we will consider *Rademacher chaos* of order 2, i.e. a random variable defined as

$$S = \sup_{M\in\mathcal{F}} \sum_{i,j=1}^n \varepsilon_i\varepsilon_j M_{ij},$$

where $\mathcal{F}$ is a countable set of real symmetric matrices with zeros on the diagonal such that

$$\sup_{M\in\mathcal{F}} \sup_{\substack{\alpha,\beta\in\mathbb{R}^n \\ ||\alpha||_2=||\beta||_2=1}} \sum_{i,j=1}^n M_{ij}\alpha_i\beta_j = K < \infty, \tag{2.17}$$

where $||\cdot||_2$ stands for the euclidean norm in $\mathbb{R}^n$. Let us define a random variable Y by

$$Y = \sup_{M\in\mathcal{F}} \left(\sum_{i=1}^n \left(\sum_{j=1}^n \varepsilon_j M_{ij}\right)^2\right)^{1/2}.$$

We are interested in obtaining an upper bound on $\mathbb{P}(S-\mathbb{E}S\geq t)$ in terms of $\mathbb{E}Y^2$. Using Corollary 5 we will prove the following theorem, which was first obtained by M. Talagrand in [23].

Theorem 6 *For all $t \geq 0$*

$$\mathbb{P}(S - \mathbb{E}S \geq t) \leq e^{-\frac{t^2}{16\mathbb{E}Y^2 + 16Kt}}. \tag{2.18}$$

Proof. The proof will basically follow the arguments from [4] and Corollary 5 will just allow us to slightly improve the constants.

Without loss of generality we can assume that $\mathcal{F}$ is finite, since when we take limits with $\#\mathcal{F} \to \infty$, inequality (2.18) will be preserved. We will also assume that $K = 1$ (the whole generality of the theorem may be obtained from this special case by applying it to the random variable S/K).

For fixed $\varepsilon_1, \ldots, \varepsilon_n$ let M be the element of $\mathcal{F}$ for which the supremum in the definition of S is obtained. Then, since M is symmetric and $M_{ii} = 0$, we have

$$S - S_i \leq \left(2\sum_{j=1}^n M_{ij}\varepsilon_j\right)(\varepsilon_i - \tilde{\varepsilon}_i)$$

and thus

$$\sum_{i=1}^n \mathbb{E}_{\tilde{\varepsilon}_i}(S - S_i)_+^2 \leq 4\sum_{i=1}^n \left(\sum_{j=1}^n M_{ij}\varepsilon_j\right)^2 \mathbb{E}_{\tilde{\varepsilon}_i}(\varepsilon_i - \tilde{\varepsilon}_i)^2 = 8\sum_{i=1}^n \left(\sum_{j=1}^n M_{ij}\varepsilon_j\right)^2 \leq 8Y^2.$$

Thus, from Corollary 5 we get

$$\text{Ent } e^{\lambda S} \leq 4\lambda^2 \mathbb{E}Y^2 e^{\lambda S}. \tag{2.19}$$

But from Jensen's inequality it follows that

$$\mathbb{E}\lambda(Y^2 - S)\frac{e^{\lambda S}}{\mathbb{E}e^{\lambda S}} = \mathbb{E}\log\left(e^{\lambda(Y^2 - S)}\right)\frac{e^{\lambda S}}{\mathbb{E}e^{\lambda S}} \leq \log\frac{\mathbb{E}e^{\lambda Y^2}}{\mathbb{E}e^{\lambda S}},$$

so $\mathbb{E}\lambda Y^2 e^{\lambda S} \leq \text{Ent } e^{\lambda S} + \mathbb{E}e^{\lambda S}\log\mathbb{E}e^{\lambda Y^2}$, which combined with (2.19) gives us

$$\text{Ent}e^{\lambda S} \leq \frac{4\lambda}{1 - 4\lambda}\mathbb{E}e^{\lambda S}\log\mathbb{E}e^{\lambda Y^2} \tag{2.20}$$

for all $\lambda \in [0, 1/4)$.

It remains to find an upper bound on $\log\mathbb{E}e^{\lambda Y^2}$. Let us notice that

$$Y = \sup_{M\in\mathcal{F}} \sup_{\substack{\alpha\in\mathbb{R}^n \\ ||\alpha||_2\leq 1}} \sum_{i=1}^n\sum_{j=1}^n \varepsilon_j\alpha_i M_{ij} = \sum_{i=1}^n\sum_{j=1}^n \varepsilon_j\alpha_i M_{ij}$$

for some M, α, depending on the sample $\varepsilon_1, \ldots, \varepsilon_n$. Thus

$$Y - Y_j \leq \left(\sum_{i=1}^n M_{ij}\alpha_i\right)(\varepsilon_j - \tilde{\varepsilon}_j)$$

hence

$$\mathbb{E}_{\tilde{\varepsilon}_j}(Y - Y_j)_+^2 \leq 2\left(\sum_{i=1}^n M_{ij}\alpha_i\right)^2$$

and

$$\sum_{j=1}^{n} \mathbb{E}_{\tilde{\varepsilon}_j}(Y-Y_j)_+^2 \le 2 \sup_{||\alpha||_2=1} \sum_{j=1}^{n} \left(\sum_{i=1}^{n} M_{ij}\alpha_i \right)^2$$

But

$$\sup_{||\alpha||=1} \sum_{j=1}^{n} \left(\sum_{i=1}^{n} M_{ij}\alpha_i \right)^2 \le 1$$

by (2.17) and our assumption $K = 1$. So finally we have (since $(a^2-b^2)_+ \le 2a(a-b)_+^2$ for all $a, b \ge 0$)

$$\sum_{j=1}^{n} \mathbb{E}_{\tilde{\varepsilon}_j}(Y^2-Y_j^2)_+^2 \le 4Y^2 \sum_{j=1}^{n} \mathbb{E}_{\tilde{\varepsilon}_j}(Y-Y_j)_+^2 \le 8Y^2$$

and thus by Corollary 5

$$\mathrm{Ent} e^{\lambda Y^2} \le 4\lambda^2 \mathbb{E} Y^2 e^{\lambda Y^2}$$

or denoting $\psi(\lambda) = \log \mathbb{E} e^{\lambda Y^2}$.

$$\left(\frac{\psi(\lambda)}{\lambda} \right)' \le 4\psi'(\lambda).$$

Since $\psi(0) = 0$ and $\lim_{\lambda \to 0+} \frac{\psi(\lambda)}{\lambda} = \mathbb{E}Y^2$, integration of the above inequality yields

$$\frac{\psi(\lambda)}{\lambda} - \mathbb{E}Y^2 \le 4\psi(\lambda),$$

and thus

$$\log \mathbb{E} e^{\lambda Y^2} \le \frac{\lambda}{1-4\lambda} \mathbb{E} Y^2$$

for all $\lambda \in [0, 1/4)$, which combined with (2.20) gives

$$\mathrm{Ent} e^{\lambda S} \le \frac{4\lambda^2}{(1-4\lambda)^2} \mathbb{E} Y^2 \mathbb{E} e^{\lambda S}.$$

Again if we denote $\psi(\lambda) = \log \mathbb{E} e^{\lambda S}$, the last inequality reads as

$$\left(\frac{\psi(\lambda)}{\lambda} \right)' \le \frac{4\mathbb{E}Y^2}{(1-4\lambda)^2}.$$

Thus (since $\lim_{\lambda \to 0+} \frac{\psi(\lambda)}{\lambda} = \mathbb{E}S$)

$$\frac{1}{\lambda} \log \mathbb{E} e^{\lambda(S-\mathbb{E}S)} = \frac{\psi(\lambda)}{\lambda} - \mathbb{E}S \le \int_0^\lambda \frac{4\mathbb{E}Y^2}{(1-4s)^2} ds = \frac{4\lambda}{1-4\lambda} \mathbb{E}Y^2$$

or equivalently

$$\log \mathbb{E} e^{\lambda(S-\mathbb{E}S)} \le \frac{4\lambda^2}{1-4\lambda} \mathbb{E} Y^2$$

for $\lambda \in [0, 1/4)$. Now by Markov inequality for $t \ge 0$ and $\lambda \in [0, 1/4)$

$$\mathbb{P}(S - \mathbb{E}S \ge t) \le e^{\frac{4\lambda^2}{1-4\lambda}\mathbb{E}Y^2 - \lambda t}.$$

Substituting $\lambda = (1 - \frac{1}{\sqrt{(t/\mathbb{E}Y^2)+1}})/4$ gives

$$\mathbb{P}(S - \mathbb{E}S \ge t) \le e^{-\mathbb{E}Y^2 h(\frac{t}{\mathbb{E}Y^2})/4},$$

where

$$h(u) = (\sqrt{u+1} - 1)^2 = \left(\frac{u}{\sqrt{u+1}+1}\right)^2 \geq \frac{u^2}{4(u+1)}.$$

Thus for all $t \geq 0$

$$\mathbb{P}(S - \mathbb{E}S \geq t) \leq e^{-\frac{t^2}{16\mathbb{E}Y^2 + 16t}}.$$

□

2.3. Configuration functions and convex functions

So far we have been estimating the entropy of a function of independent random variables $X_1, \ldots, X_n$ by expressions involving their independent copies. In some situations it is useful not to introduce such independent variables but rather to drop some of the variables X_i, that is to approximate the statistic by functions, which do not depend on all of the variables $X_1, \ldots, X_n$. The next theorem, due to S. Boucheron, G. Lugosi and P. Massart ([3]), will constitute a good basis for such a method.

Theorem 7 *Consider independent random variables $X_1, \ldots, X_n$ with values in a measurable space $(\Sigma, \mathcal{F})$. Let $f\colon \Sigma^n \to \mathbb{R}$ and $f_i\colon \Sigma^{n-1} \to \mathbb{R}$ $(i = 1, \ldots, n)$ be measurable functions and denote $S = f(X_1, \ldots, X_n)$, $S_i = f(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$. If $\mathbb{E}Se^S < \infty$ then*

$$\mathrm{Ent}e^S \leq \sum_{i=1}^n \mathbb{E}(\phi(S_i - S))e^S),$$

where $\phi(x) = e^x - x - 1$.

Proof. Obviously we may consider $X_1, \ldots, X_n$ as coordinates on a product space (Ω, μ), $\mu = \mu_1 \otimes \ldots \otimes \mu_n$. Let us notice that if we fix the values of all of the variables $X_1, \ldots, X_n$ except for X_i, then S_i becomes a constant. Moreover by Fubini theorem $E_{\mu_i}Se^S < \infty$ a.e. (with respect to $\mu_1 \otimes \ldots \otimes \mu_{i-1} \otimes \mu_{i+1} \otimes \ldots \otimes \mu_n$). Therefore we may use (1.3) with $u = e^{S_i}$, $X = e^S$ to obtain

$$\mathrm{Ent}_{\mu_i}e^S \leq \mathbb{E}_{\mu_i}(e^S(S - S_i) - (e^S - e^{S_i})) = \mathbb{E}_{\mu_i}e^S(e^{S_i - S} - (S_i - S) - 1) = \mathbb{E}_{\mu_i}e^S\phi(S_i - S).$$

The theorem follows now immediately from the tensorization property of entropy.

□

2.3.1. Configuration functions

We will use Theorem 7 to obtain a result analogous to Theorem 4.3. in [17]. Before we proceed, let us introduce a few definitions.

Definition 2 (The penalized Hamming distance) *For a non-negative vector $\alpha = (\alpha_1, \ldots, \alpha_n)$, define $d\colon \Sigma^n \times \Sigma^n \to \mathbb{R}$ with the formula*

$$d_\alpha(x, y) = \sum_{i=1}^n \alpha_i \mathbf{1}_{x_i \neq y_i}.$$

Definition 3 *Consider a measurable space $(\Sigma, \mathcal{F})$. A measurable function $f\colon \Sigma^n \to \mathbb{R}^+$ will be called a c*-configuration function *if for every $x \in \Sigma^n$ there is a non-negative unit vector $\alpha \in \mathbb{R}^n$, such that*

$$f(y) \geq f(x) - \sqrt{cf(x)}d_\alpha(x,y)$$

for all $y \in \Sigma^n$.

Theorem 8 *Let $X_1,\ldots,X_n$ be independent random variables with values in a Polish space $(\Sigma,\mathcal{F})$, where $\mathcal{F}$ is the Borel σ-field on Σ. Then for every c-configuration function $f\colon \Sigma^n \to \mathbb{R}^+$, the random variable $S = f(X_1,\ldots,X_n)$ satisfies the following deviation inequality*

$$\mathbb{P}(S \geq \mathbb{E}S + t) \leq e^{-\frac{t^2}{2c\mathbb{E}S+2ct}}.$$

Remark Boucheron, Lugosi and Massart in [3] define configuration function in a different way. They consider the so called *hereditary properties*, i.e. properties $\mathcal{P}$, defined for sequences of arbitrary length, such that whenever a sequence $(x_1,\ldots,x_n)$ has the property $\mathcal{P}$, so do its all subsequences. The length of the longest subsequence satisfying a hereditary property $\mathcal{P}$ is then called a configuration function. It is easy to see that such functions satisfy the definition of 1-configuration functions. Indeed, let $\mathcal{P}$ be a hereditary property. Fix a vector $x = (x_1,\ldots,x_n)$ and consider a sequence of indices $i_1 < \ldots < i_m$ such that $(x_{i_1},\ldots,x_{i_m})$ is one of the longest subsequences of x which satisfy the property $\mathcal{P}$. Then $f(x) = m$ and for every vector $y = (y_1,\ldots,y_n)$

$$f(y) \geq \#\{k \in \{1,\ldots,m\}\colon x_{i_k} = y_{i_k}\} = f(x) - \#\{k\colon x_{i_k} \neq y_{i_k}\} = f(x) - \sqrt{f(x)}d_\alpha(x,y),$$

where $\alpha_j = 1/\sqrt{f(x)}$ if $j = i_k$ for some k and 0 otherwise.

Theorem 8 may be thus used to obtain concentration inequalities for instance for the length of a longest increasing subsequence of an i.i.d sample. However the bounds provided by this theorem in the case of hereditary properties may be improved as shown in [3].

To present another application of Theorem 8, arising in many situations in computer science, let us consider independent random variables $X_1,\ldots,X_{n_1},Y_1,\ldots,Y_{n_2}$ and define the random variable L as the length of a *longest common subsequence* of $(X_1,\ldots,X_{n_1})$ and $(Y_1,\ldots,Y_{n_2})$. By an argument analogous to the one presented above for hereditary properties, L is a 2-configuration function, hence by Theorem 8

$$\mathbb{P}(L \geq \mathbb{E}L + t) \leq e^{-\frac{t^2}{4\mathbb{E}L+4t}}.$$

Proof of Theorem 8. Let us fix $x_0 \in \Sigma^n$ and notice that for every $x \in \Sigma^n$ we have for some non-negative unit vector α

$$f(x) - n\sqrt{cf(x)} \leq f(x) - \sqrt{cf(x)}d_\alpha(x,x_0) \leq f(x_0).$$

Thus $\sup_{x\in\Sigma^n}(f(x)^2 - n\sqrt{cf(x)}) < \infty$ and in consequence $||S||_\infty < \infty$ and $\mathbb{E}\lambda Se^{\lambda S} < \infty$ for every λ.
Define for $i = 1,\ldots,n$ the functions $f_i\colon \Sigma^{n-1} \to \mathbb{R}^+$ by

$$f_i(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n) = \inf_{y\in\Sigma} f(x_1,\ldots,x_{n-1},y,x_{n+1},\ldots,x_n)$$

and set $S_i = f_i(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n)$. Theorem 7 gives us

$$\mathrm{Ent} e^{\lambda S} \le \mathbb{E} e^{\lambda S} \sum_{i=1}^{n} \phi(\lambda(S_i - S)). \tag{2.21}$$

But from the Taylor extension of the exponential function we have $\phi(x) = e^{\xi} x^2/2$ for some ξ between 0 and x. Now from the monotonicity of e^x we have

$$\phi(x) \le \frac{1}{2} x^2 \tag{2.22}$$

for $x \le 0$.
On the other hand, by the definition of S_i, we have $S_i \le S$ and thus from (2.21) and (2.22) we obtain for $\lambda \ge 0$

$$\mathrm{Ent} e^{\lambda S} \le \frac{\lambda^2}{2} \mathbb{E} e^{\lambda S} \sum_{i=1}^{n} (S - S_i)^2.$$

Now by the definition of configuration functions

$$f(x_1, \dots, x_n) - f_i(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) \le \sqrt{c f(x)} \alpha_i,$$

for $i = 1, \dots, n$, where $\alpha = (\alpha_1, \dots, \alpha_n)$ is a positive unit vector corresponding to $(x_1, \dots, x_n)$. Therefore

$$(S - S_i)^2 \le c S \alpha_i^2$$

and in consequence

$$\mathrm{Ent} e^{\lambda S} \le c \frac{\lambda^2}{2} \mathbb{E} S e^{\lambda S}.$$

In other words, we have obtained a differential inequality for $F(\lambda) = \mathbb{E} e^{\lambda S}$, namely

$$\lambda F'(\lambda) - F(\lambda) \log F(\lambda) \le c \frac{\lambda^2}{2} F'(\lambda).$$

Let us define $\psi(\lambda) = \log F(\lambda)$ and rewrite the above inequality as

$$\frac{\lambda \psi'(\lambda) - \psi(\lambda)}{\lambda^2} \le \frac{c}{2} \psi'(\lambda)$$

or

$$\left(\frac{\psi(\lambda)}{\lambda} \right)' \le \frac{c}{2} \psi'(\lambda).$$

Since $\lim_{\lambda \to 0+} \psi(\lambda)/\lambda = \mathbb{E} S$ and $\psi(0) = 0$, we can integrate the last inequality and obtain

$$\log F(\lambda) - \lambda \mathbb{E} S \le \lambda \frac{c}{2} \log F(\lambda),$$

that is

$$\log \mathbb{E} e^{\lambda(S - \mathbb{E} S)} \le \frac{\lambda^2 c}{2 - \lambda c} \mathbb{E} S$$

for all $\lambda \in [0, 2/c)$. Now

$$\mathbb{P}(S \ge \mathbb{E} S + t) \le \inf_{\lambda \in [0, 2/c)} e^{\frac{\lambda^2 c}{2 - \lambda c} \mathbb{E} S - \lambda t}.$$

The infimum is obtained for $\lambda = \frac{2}{c}(1 - (t/\mathbb{E} S + 1)^{-1/2})$ and equals $e^{-2 \mathbb{E} S h(t/\mathbb{E} S)/c}$, where $h(u) = (\sqrt{u+1} - 1)^2 \ge u^2/(4u + 4)$, which proves the theorem.

□

Remark The formulation of Theorem 8 involves the notion of a Polish space. Let us stress that the only reason for this is the potential problem with measurability of functions f_i defined in the proof of the theorem. In general the infimum of a family of measurable functions need not be measurable, however in this case measurability (at least with respect to the completed σ-field) is guaranteed by Suslin Theorem. Of course in applications configuration functions appear mainly in discrete mathematics, so measurability is not a real problem and Theorem 8 has been formulated in such an "involved" way just for the sake of accuracy. It is also worth mentioning that a version of this theorem (with slightly weaker constants) may be obtained in a similar way from Theorem 4.

2.3.2. Deviation inequalities for convex functions

Corollary 6 *Consider a convex, L-lipschitz function* $\varphi\colon [0,1]^n \to \mathbb{R}$. *Let* X_i $(i = 1, \ldots, n)$ *be independent random variables with values in* $[0,1]$. *Denote* $S = \varphi(X_1, \ldots, X_n)$.
Then for all $t > 0$ *we have*

$$\mathbb{P}(S - \mathbb{E}S \geq t) \leq e^{-t^2/2L^2}.$$

Proof. Let us define, similarly as in the proof of Theorem 8, $\varphi_i\colon [0,1]^{n-1} \to \mathbb{R}$ with the formula

$$\varphi_i(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) = \inf_{y \in [0,1]} \varphi(x_1, \ldots, x_{i-1}, y, x_{i+1}, \ldots, x_n)$$

and denote $S_i = f(X_1, \ldots, X_n)$. We will show that

$$||\sum_{i=1}^n (S - S_i)^2||_\infty \leq L^2.$$

For fixed $x = (x_1, \ldots, x_n) \in [0,1]^n$, $y = (y_1, \ldots, y_n) \in [0,1]^n$ let x^i be the point obtained from x by replacing the i-th coordinate with $y_i \in [0,1]$. We will first find an upper bound for

$$M = \sum_{i=1}^n (\varphi(x) - \varphi(x^i))_+^2.$$

Since for $(n+1)$-tuples of points $(x, x^1, \ldots, x^n)$ such that $\varphi(x) \leq \varphi(x^i)$ for all i, we have $M = 0$, we can assume that

$$\sum_{i=1}^n (\varphi(x) - \varphi(x^i))_+ > 0.$$

Let

$$z = \frac{\sum_{i=1}^n (\varphi(x) - \varphi(x^i))_+ x^i}{\sum_{i=1}^n (\varphi(x) - \varphi(x^i))_+} = (z_1, \ldots, z_n)$$

From Jensen's inequality it follows that

$$\frac{\sum_{i=1}^n (\varphi(x) - \varphi(x^i))_+ \varphi(x^i)}{\sum_{i=1}^n (\varphi(x) - \varphi(x^i))_+} \geq \varphi(z). \tag{2.23}$$

Moreover

$$\begin{aligned} z_i &= \frac{\sum_{j \neq i} (\varphi(x) - \varphi(x^j))_+ x_i}{\sum_{j=1}^n (\varphi(x) - \varphi(x^j))_+} + \frac{(\varphi(x) - \varphi(x^i))_+}{\sum_{j=1}^n (\varphi(x) - \varphi(x^j))_+} y_i \\ &= x_i + \frac{(\varphi(x) - \varphi(x^i))_+}{\sum_{j=1}^n (\varphi(x) - \varphi(x^j))_+} (y_i - x_i). \end{aligned}$$

Thus

$$||z-x||^2 = \frac{\sum_{i=1}^n (\varphi(x)-\varphi(x^i))_+^2 (y_i - x_i)^2}{(\sum_{i=1}^n (\varphi(x)-\varphi(x^i))_+)^2} \leq \frac{M}{(\sum_{i=1}^n (\varphi(x)-\varphi(x^i))_+)^2}.$$

Now from the Lipschitz property of φ we get

$$\varphi(z) \geq \varphi(x) - \frac{\sqrt{M}L}{\sum_{i=1}^n (\varphi(x)-\varphi(x^i))_+} \tag{2.24}$$

Putting together (2.23) and (2.24) we conclude that

$$\sum_{i=1}^n (\varphi(x)-\varphi(x^i))_+ \varphi(x^i) \geq \sum_{i=1}^n (\varphi(x)-\varphi(x^i))_+ \varphi(x) - \sqrt{M}L$$

or equivalently

$$M = \sum_{i=1}^n (\varphi(x)-\varphi(x^i))_+ (\varphi(x)-\varphi(x^i)) \leq \sqrt{M}L,$$

that is

$$M \leq L^2.$$

Let now $\tilde{x}^i \in [0,1]^{n-1}$ denote the vector, obtained from x by skipping its i-th coordinate. The function φ is continuous, hence there exist numbers $y_i \in [0,1]$ $(i = 1, \ldots, n)$, such that

$$\varphi_i(\tilde{x}^i) = \varphi(x_1, \ldots, x_{i-1}, y_i, x_{i+1}, \ldots, x_n),$$

then

$$\sum_{i=1}^n (\varphi(x) - \varphi_i(\tilde{x}^i))^2 = \sum_{i=1}^n (\varphi(x) - \varphi(x_1, \ldots, x_{i-1}, y_i, x_{i+1}, \ldots, x_n))_+^2 \leq L^2.$$

Now just like in the proof of Theorem 8 we have

$$\mathrm{Ent} e^{\lambda S} \leq \frac{\lambda^2}{2} \mathbb{E} e^{\lambda S} \sum_{i=1}^n (S - S_i)^2,$$

for $\lambda \geq 0$, so

$$\mathrm{Ent} e^{\lambda S} \leq \frac{\lambda^2}{2} L^2 \mathbb{E} e^{\lambda S},$$

which by *Herbst argument* (Lemma 2) implies the Corollary.

□

Remark Above we have derived a bound only for the upper-tail of $S = \varphi(X_1, \ldots, X_n)$. The same estimation for the lower-tail of a convex function (or equivalently upper-tail of a concave function) has been obtained in [21] from logarithmic Sobolev inequalities derived from transportation of measure approach. Moreover with use of some less general, tailored to the situation logarithmic Sobolev inequalities, Corollary 6 may be generalized to separately convex functions, i.e. functions which are convex with respect to every variable (compare [13]).

Concentration inequalities for convex functions of independent bounded random variables appeared first in M. Talagrand's paper [22] in the case of Rademacher random variables and were generalized by the same author to arbitrary random variables in [24]. Talagrand considers concentration around median which is however equivalent to concentration around the mean in a sense that if a random variable concentrates around its median with the tail Ke^{-ct^2} for some K, c, then it concentrates around its mean with the tail of the form $K'e^{-c't^2}$ where K', c' depend only on K,c and vice versa. It can be easily seen, since in the case of both types of concentration we have

$$|\mathbb{E}X - M| \leq L,$$

with the constant L depending only on K,c (K',c'). Indeed, if X concentrates around median we have by Jensen's inequality

$$|\mathbb{E}X - M| \leq \mathbb{E}|X - M| = \int_0^\infty \mathbb{P}(|X - M| \geq t)dt \leq \int_0^\infty Ke^{-ct^2}dt.$$

On the other hand we have

$$|M - \mathbb{E}X| < t$$

for any t such that $\mathbb{P}(|X - \mathbb{E}X| \geq t) < 1/2$ (and in the case of Gaussian concentration around the mean, such t can be defined by K' and c' alone).

Now, following [22], we can prove that the convexity assumption in Corollary 6 is important. From the above remark it follows that it is enough to consider concentration around median. Consider the discrete cube $I^n = \{-1, 1\}^n$ with the uniform probability measure and define

$$A_n = \{x = (x_1, \ldots, x_n) \in I^n \colon \sum_{i=1}^n x_i \leq 0\}.$$

Now set

$$f_n(y) = \inf\{||x - y|| \colon x \in A_n\}.$$

The functions f_n are of course 1-Lipschitz continuous and since uniform measure on I^n is the product of n symmetric measures on $\{+1, -1\}$ we can regard f_n as functions of n independent Rademacher variables. Obviously $\mathbb{P}(f_n \leq 0) = \mathbb{P}(f_n = 0) \geq 1/2$ and $\mathbb{P}(f_n \geq 0) = 1$, so 0 is a median of f_n. However

$$f_n(y) = 2(\lceil \frac{(\sum_{i=1}^n y_i)_+}{2} \rceil)^{1/2},$$

so by the Central Limit Theorem we get that $\mathbb{P}(f_n \geq cn^{1/4}) > 1/4$ for some constant c and every n, which shows that there cannot be a universal Gaussian bound on tail probabilities for all 1-Lipschitz functions.

2.3.3. Random matrices

Concentration inequalities for convex functions may be used for instance to analyse deviation from the mean for eigenvalues of random matrices. Namely, let X_{ij} for $1 \leq i \leq j \leq n$ be independent random variables such that $|X_{ij}| \leq 1$ a.e. Denote $X_{ji} = X_{ij}$ for $i < j$ and consider a random symmetric matrix $A = (X_{ij})_{i,j=1}^n$. The spectral theorem asserts that all eigenvalues of A are real, so we can consider a random variable λ_i ($i = 1, \ldots, n$) defined as the i-th largest eigenvalue of A (counting with multiplicities). We are interested in concentration around mean for λ_i.

The first obvious observation we have to make is that all symmetric matrices constitute a linear space of dimension $\frac{n(n+1)}{2}$, which can be identified with $\mathbb{R}^{\frac{n(n+1)}{2}}$. To be able to use concentration inequalities for convex functions we need the following lemma

Lemma 6 *For every $k = 1, \ldots, n$ the function $\varphi\colon \mathbb{R}^{\frac{n(n+1)}{2}} \to \mathbb{R}$ given by*

$$\varphi(A) = \lambda_1(A) + \ldots + \lambda_k(A)$$

is convex.

Proof. To show this lemma it is enough to prove the so-called Ky-Fan theorem, which claims that

$$\lambda_1(A)+\ldots+\lambda_k(A) = \sup\left\{\sum_{i=1}^{k} x_i^T A x_i\colon x_1, \ldots, x_k - \text{an orthonormal system in } \mathbb{R}^n\right\}. \quad (2.25)$$

Indeed, the expression $\sum_{i=1}^{k} x_i^T A x_i$ defines a linear function of A, hence having (2.25), we can claim $\varphi(A)$ to be convex as a pointwise supremum of linear functions. In the case of $k = 1$, equality (2.25) is a basic theorem of linear algebra or functional analysis, which we will assume to be known. On the other hand, the case of $k = n$ reduces (2.25) to the theorem about preserving the trace of a matrix under a change of basis transformation.

We will now prove (2.25). It is quite obvious that the right-hand side of (2.25) is greater than the left-hand side, since we can diagonalize A with a unitary isomorphism of $\mathbb{R}^n$ and pick up k orthonormal eigenvectors $x_1, \ldots, x_k$ corresponding to k greatest eigenvalues. Then

$$\sum_{i=1}^{k} x_i^T A x_i = \lambda_1(A) + \ldots + \lambda_k(A).$$

Since a unitary change of basis preserves orthogonality, we can assume that A is diagonal. Now we will proceed in several steps. First of all let us introduce matrices I_l, defined as diagonal matrices $(a_{ij})_{i,j=1}^{n}$ with $a_{ii} = 1$ for $i \leq l$ and $a_{ii} = 0$ for $i > l$. It is quite obvious that I_l satisfies (2.25). To every orthonormal system $x_1, \ldots, x_k$ in $\mathbb{R}^n$ we can add some vectors $x_{k+1}, \ldots, x_n$ in such a way that $x_1, \ldots, x_n$ form an orthonormal basis of $\mathbb{R}^n$. As I_l is positively definite, we get

$$\sum_{i=1}^{k} x_i^T I_l x_i \leq \sum_{i=1}^{n} x_i^T I_l x_i = \mathrm{tr} I_l = l. \quad (2.26)$$

Moreover it is clear that $x_i^T I_l x_i \leq 1$, so

$$\sum_{i=1}^{k} x_i^T I_l x_i \leq k. \quad (2.27)$$

Inequalities (2.26) and (2.27) give us

$$\sum_{i=1}^{k} x_i^T I_l x_i \leq \min(k, l) = \lambda_1(I_l) + \ldots + \lambda_k(I_l),$$

which proves (2.25) for $A = I_l$.

A diagonal matrix A can be written as

$$A = (\lambda_1 - \lambda_2)I_1 + (\lambda_2 - \lambda_3)I_2 + \ldots + (\lambda_{n-1} - \lambda_n)I_{n-1} + \lambda_n I_n.$$

Now

$$\sum_{i=1}^{k} x_i^T A x_i \leq \sum_{l=1}^{n-1} (\min(l,k)(\lambda_l - \lambda_{l+1})) + \min(n,k)\lambda_n = \lambda_1 + \ldots + \lambda_k.$$

Equality (2.25) and therefore also Lemma 6 have thus been proved.

□

We are now in position to derive deviation inequalities for λ_i.

Theorem 9 *For all $t > 0$*

$$\mathbb{P}(|\lambda_1 - \mathbb{E}\lambda_1| \geq t) \leq 2e^{-\frac{t^2}{4}},$$

whereas for $k = 2, \ldots, n$ and $t > 0$ we have

$$\mathbb{P}(|\lambda_k - \mathbb{E}\lambda_k| \geq t) \leq 4e^{-\frac{t^2}{4(\sqrt{k}+\sqrt{k-1})^2}} \leq 4e^{-\frac{t^2}{16k}}.$$

Proof.
Denote $\varphi_k(A) = \sum_{i=1}^{k} \lambda_i(A)$. We have already proved that φ_k are convex. It remains to show that they are Lipschitz functions of A with respect to the Hilbert-Schmidt norm, defined as

$$||A||_{HS} = \sqrt{\sum_{i,j=1}^{n} a_{ij}^2} = \sqrt{\sum_{i=1}^{n} \lambda_i(A)^2}.$$

Consider two symmetric matrices A and B. Let $x_1,\ldots,x_k$ be an orthonormal system of vectors, such that $Ax_i = \lambda_i x_i$. We have

$$\begin{aligned} \varphi_k(A) &= \sum_{i=1}^{k} x_i^T A x_i \\ \varphi_k(B) &\geq \sum_{i=1}^{k} x_i^T B x_i, \end{aligned}$$

hence

$$\varphi_k(A) - \varphi_k(B) \leq \sum_{i=1}^{k} x_i^T (A - B) x_i \leq \varphi_k(A - B) \leq k\sqrt{\frac{\sum_{i=1}^{k} \lambda_i(A-B)^2}{k}} \leq \sqrt{k}||A - B||_{HS}.$$

By analogy

$$\varphi_k(B) - \varphi_k(A) \leq \sqrt{k} \cdot ||A - B||_{HS},$$

so $\varphi_k(A)$ are indeed Lipschitz continuous with respect to $|| \cdot ||_{HS}$, with Lipschitz constant equal to $\sqrt{k}$.

We identify the space of all symmetric matrices with $\mathbb{R}^{\frac{n(n+1)}{2}}$. Although the Hilbert-Schmidt norm of a symmetric matrix is not exactly the same as the euclidean norm ($||A|| = \sum_{i\leq j} a_{ij}^2$) of its image under the natural isomorphism between the two spaces, they satisfy

$$||A||_{HS} \leq \sqrt{2} \cdot ||A||$$

so φ_k is $\sqrt{2k}$-Lipschitz continuous with respect to the euclidean norm.
Now we can take use of concentration inequalities for convex Lipschitz continuous functions and write for any $t > 0$

$$\mathbb{P}(|\varphi_k(A) - \mathbb{E}\varphi_k(A)| \geq t) \leq 2e^{-\frac{t^2}{4k}}.$$

For $k = 1$ the above inequality gives us concentration of λ_1, but for other values of k we still have to do some computations. Namely, since $\lambda_k = \varphi_k - \varphi_{k-1}$, we have for any $\theta \in [0, 1]$

$$\mathbb{P}(|\lambda_k - \mathbb{E}\lambda_k| \geq t) \leq \mathbb{P}(|\varphi_k - \mathbb{E}\varphi_k| \geq \theta t) + \mathbb{P}(|\varphi_{k-1} - \mathbb{E}\varphi_{k-1}| \geq (1-\theta t)) \leq 2e^{-\frac{\theta^2 t^2}{4k}} + 2e^{-\frac{(1-\theta)^2 t^2}{4(k-1)}}.$$

To finish the proof it is now enough to substitute $\theta = \sqrt{k}/(\sqrt{k} + \sqrt{k-1})$.

□

Remark For comparison purposes let us mention the Gaussian counterpart of the above theorem, which asserts that if H is a random symmetric matrix with Gaussian entries X_{ij}, such that $\mathrm{Var}X_{ij} \leq 1$ then

$$\mathbb{P}(|\lambda_k - \mathbb{E}\lambda_k| \geq t) \leq 2e^{-t^2/4}.$$

This statement follows easily from the obvious fact that λ_k is 1-Lipschitz continuous with respect to the Hilbert-Schmidt norm and Gaussian concentration inequality for Lipschitz functions (Corollary 2.15).

2.3.4. Rademacher averages

Another application of tail estimates for convex functions may be found in the area of probability in Banach spaces.

Corollary 7 *Let $(x_i)_{i=1}^n$ be a sequence of vectors from a Banach space E. Define*

$$\sigma^2 = \sup\{\sum_{i=1}^n x^*(x_i)^2 \colon x^* \in E^*,\ ||x^*|| \leq 1\}.$$

Let S be a random variable defined by

$$S = ||\sum_{i=1}^n \varepsilon_i x_i||,$$

where $(\varepsilon_i)_{i=1}^n$ is a sequence of independent Rademacher variables. Then

$$\mathbb{P}(S - \mathbb{E}S \geq t) \leq e^{-t^2/8\sigma^2}$$

for all $t \geq 0$.

Proof. Obviously, the function $\varphi\colon \mathbb{R}^n \to \mathbb{R}$ defined with the formula

$$\varphi(t_1,\ldots,t_n) = ||\sum_{i=1}^n t_i x_i||$$

is convex. Thus to prove the Corollary it is enough to find its Lipschitz constant. We have

$$|\varphi(t_1,\ldots,t_n) - \varphi(s_1,\ldots,s_n)| \le ||\sum_{i=1}^n (t_i - s_i)x_i||.$$

By the Hahn-Banach Theorem $||\sum_{i=1}^n (t_i - s_i)x_i|| = x^*(\sum_{i=1}^n (t_i - s_i)x_i)$ for some $x^* \in E^*$, $||x^*|| = 1$. Hence, by the Cauchy-Schwarz inequality

$$|\varphi(t_1,\ldots,t_n) - \varphi(s_1,\ldots,s_n)| \le \sqrt{\sum_{i=1}^n (t_i - s_i)^2}\sqrt{\sum_{i=1}^n x^*(x_i)^2} \le \sigma \cdot \sqrt{\sum_{i=1}^n (t_i - s_i)^2},$$

and thus φ is σ-Lipschitz.

□

It is worth mentioning that Corollary 7 implies the following Khintchin-Kahane type inequality

Corollary 8 *There exists a universal constant K such that for any Banach space E, $x_1,\ldots,x_n \in E$ and all $p \ge 1$*

$$||\sum_{i=1}^n \varepsilon_i x_i||_p \le ||\sum_{i=1}^n \varepsilon_i x_i||_1 + K\sigma p^{1/2},$$

with $\sigma = (\sup\{\sum_{i=1}^n x^*(x_i)^2 \colon x^* \in E^*,\ ||x^*|| \le 1\})^{1/2} \le ||\sum_{i=1}^n \varepsilon_i x_i||_2$.

Proof. Using the notation from Corollary 7, we have

$$0 \le S \le \mathbb{E}S + (S - \mathbb{E}S)_+,$$

hence

$$\begin{aligned}
||S||_p &\le ||\mathbb{E}S + (S - \mathbb{E}S)_+||_p \\
&\le \mathbb{E}S + ||(S - \mathbb{E}S)_+||_p \\
&= \mathbb{E}S + (\int_0^\infty pt^{p-1}\mathbb{P}(S - \mathbb{E}S > t)dt)^{1/p} \\
&\le \mathbb{E}S + (\int_0^\infty pt^{p-1}e^{-t^2/8\sigma^2}dt)^{1/p} \\
&\le \mathbb{E}S + K\sigma p^{1/2}.
\end{aligned}$$

Let us notice that to prove the Corollary it was enough to use the bound on the upper tail of S.

□

Chapter 3

Moments estimates

In this chapter we present another method of deriving tail inequalities for random variables. Roughly speaking, it relies on estimates of all the (integer) moments of a random variable, which in some cases together with the Chebyshev inequality can yield exponential concentration. In the first section we explain the method on relatively easy examples of sub-Gaussian random variables, in the second we present a powerful general moment inequality, discovered recently by S. Boucheron, O.Bosquet, G. Lugosi and P. Massart (comp. [2]) and apply it to some special random variables. What is especially interesting from our point of view, is the fact that at the core of the proof of the aforementioned general inequality there are tensorization properties of some entropy functionals.

3.1. Random variables with sub-Gaussian tails

Theorem 10 *Let $X_1, \dots, X_n$ be independent mean zero random variables, such that for all i*

$$\mathbb{P}(|X_i| \geq t) \leq K e^{-t^2/L_i^2}$$

for all $t \geq 0$. Then the random variable $S = \sum_{i=1}^n X_i$ satisfies

$$\mathbb{P}(|S - \mathbb{E}S| \geq t) \leq e^2 e^{-\frac{t^2}{C_K^2(\sum_{i=1}^n L_i^2)}}$$

for all $t \geq 0$, with $C_K = 2e(D + \sqrt{\log K + \log\sqrt{2}})$, where D is a universal constant.

Before we proceed with the proof of Theorem 10, we need three easy lemmas.

Lemma 7 *Let $X_1, \dots, X_n$ be independent mean zero random variables and $\varepsilon_1, \dots, \varepsilon_n$ a sequence of independent Rademacher variables, independent of $X_1, \dots, X_n$. Then for every $p \geq 0$ we have*

$$\mathbb{E}|\sum_{i=1}^n X_i|^p \leq 2^p \mathbb{E}|\sum_{i=1}^n \varepsilon_i X_i|^p.$$

Proof. Let the random vector $(\tilde{X}_i, \dots, \tilde{X}_n)$ be an independent copy of $(X_1, \dots, X_n)$. Then

$$\mathbb{E}|\sum_{i=1}^n X_i|^p = \mathbb{E}|\sum_{i=1}^n (X_i - \mathbb{E}X_i)|^p = \mathbb{E}|\sum_{i=1}^n (X_i - \mathbb{E}\tilde{X}_i)|^p \leq \mathbb{E}|\sum_{i=1}^n (X_i - \tilde{X}_i)|^p,$$

where the last inequality follows from the Jensen inequality applied to the function $t \mapsto |t|^p$ and the expectation with respect to $(\tilde{X}_i)_{i=1}^n$. Notice now that for every fixed sequence $\varepsilon_1, \ldots, \varepsilon_n$, the random variable $\sum_{i=1}^n \varepsilon_i (X_i - \tilde{X}_i)$ has the same distribution. Hence

$$\begin{aligned}
\mathbb{E}|\sum_{i=1}^n X_i|^p &\leq \mathbb{E}|\sum_{i=1}^n \varepsilon_i (X_i - \tilde{X}_i)|^p \leq \mathbb{E}\left|\frac{2\sum_{i=1}^n \varepsilon_i X_i - 2\sum_{i=1}^n \varepsilon_i \tilde{X}_i}{2}\right|^p \\
&\leq \mathbb{E}\frac{|2\sum_{i=1}^n \varepsilon_i X_i|^p + |2\sum_{i=1}^n \varepsilon_i \tilde{X}_i|^p}{2} = 2^p \mathbb{E}|\sum_{i=1}^n \varepsilon_i X_i|^p.
\end{aligned}$$

□

Lemma 8 *Let $\varphi\colon \mathbb{R} \to \mathbb{R}$ be a convex function, $\varepsilon_1, \ldots, \varepsilon_n$ a sequence of independent Rademacher variables and $a_1, \ldots, a_n$, $b_1, \ldots, b_n$ two sequences of nonnegative real numbers, such that for every i $a_i \leq b_i$. Then*

$$\mathbb{E}\varphi(\sum_{i=1}^n a_i \varepsilon_i) \leq \mathbb{E}\varphi(\sum_{i=1}^n \varepsilon_i b_i).$$

Proof. It is enough to prove the monotonicity of function $f(t) = \mathbb{E}\varphi(a + t\varepsilon_1)$, for every choice of the parameter a. By the convexity assumption we have for $0 < s < t$

$$\frac{\varphi(a+t) - \varphi(a+s)}{t-s} \geq \frac{\varphi(a-s) - \varphi(a-t)}{t-s},$$

or equivalently

$$f(s) = \frac{1}{2}(\varphi(a+s) + \varphi(a-s)) \leq \frac{1}{2}(\varphi(a+t) + \varphi(a-t)) = f(t).$$

□

Lemma 9 *Let $X_1, \ldots, X_n$, $Y_1, \ldots, Y_n$ be independent, symmetric random variables, such that for all $i = 1, \ldots, n$ and $t \geq 0$, we have $\mathbb{P}(|X_i| \geq t) \leq \mathbb{P}(|Y_i| \geq t)$. Then for all $p \geq 1$*

$$\mathbb{E}|\sum_{i=1}^n X_i|^p \leq \mathbb{E}|\sum_{i=1}^n Y_i|^p.$$

Proof. Let $\varepsilon_1, \ldots, \varepsilon_n$ be a sequence of independent Rademacher variables, independent of $(X_i)_{i=1}^n$ and $(Y_i)_{i=1}^n$. Let us notice that by symmetry X_i (Y_i) has the same distribution as $\varepsilon_i |X_i|$ $(\varepsilon_i |Y_i|)$. Since we may consider $|X_i|$ and $|Y_i|$ as defined on $\Omega_i = (0,1)$ as the 'inverse' of their distribution functions, without loss of generality, we can assume that for $= 1, \ldots, n$ $|X_i| \leq |Y_i|$ a.e. Thus

$$\mathbb{E}|\sum_{i=1}^n X_i|^p = \mathbb{E}_{X,Y}\mathbb{E}_\varepsilon|\sum_{i=1}^n \varepsilon_i |X_i||^p \leq \mathbb{E}_{X,Y}\mathbb{E}_\varepsilon|\sum_{i=1}^n \varepsilon_i |Y_i||^p = \mathbb{E}|\sum_{i=1}^n Y_i|^p,$$

where the inequality follows from Lemma 8.

□

Lemma 10 *Let γ be a Gaussian random variable with the density $g(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$. Then for every $t \geq 0$, we have*

$$\frac{1}{\sqrt{2}} e^{-t^2} \leq \mathbb{P}(|\gamma| \geq t). \tag{3.1}$$

Proof. For every $s, t \in \mathbb{R}$ we have $(s-t)^2 \geq \frac{s^2}{2} - t^2$. Therefore

$$\frac{\sqrt{\pi}}{2} = \int_t^\infty e^{-(s-t)^2} ds \leq \int_t^\infty e^{t^2} e^{-s^2/2} ds,$$

which is equivalent to (3.1).

□

Proof of Theorem 10. Let γ be a standard Gaussian random variable. Define $C = \sqrt{\log K + \log \sqrt{2}}$. Then for $t \geq 0$, $\mathbb{P}(|X_i| - L_i C \geq t) \leq Ke^{-\frac{(t+L_iC)^2}{L_i^2}} \leq Ke^{-C^2} e^{-\frac{t^2}{L_i^2}} \leq \mathbb{P}(|L_i \gamma| \geq t)$, where the last inequality follows from Lemma 10. Thus for $t \geq 0$

$$\mathbb{P}((|X_i| - L_i C)_+ \geq t) \leq \mathbb{P}(|L_i\gamma| \geq t).$$

Now we have for $p \geq 1$

$$\begin{aligned} || \sum_{i=1}^n X_i ||_p &\leq 2|| \sum_{i=1}^n \varepsilon_i X_i ||_p = 2|| \sum_{i=1}^n \varepsilon_i |X_i| \, ||_p \qquad (3.2) \\ &= 2|| \sum_{i=1}^n \varepsilon_i (|X_i| - L_i C)_+ + \sum_{i=1}^n \varepsilon_i |X_i| \mathbf{1}_{\{|X_i| \leq L_i C\}} + \sum_{i=1}^n \varepsilon_i C L_i \mathbf{1}_{\{|X_i| > L_i C\}} \, ||_p \\ &\leq 2(|| \sum_{i=1}^n \varepsilon_i (|X_i| - C)_+ ||_p + || \sum_{i=1}^n \varepsilon_i (|X_i| \mathbf{1}_{\{|X_i| \leq L_i C\}} + L_i C \mathbf{1}_{\{|X_i| > L_i C\}}) \, ||_p). \end{aligned}$$

Let now $\gamma_1, \ldots, \gamma_n$ be i.i.d. random variables, distributed identically as γ. We have $|| \sum_{i=1}^n L_i \gamma_i ||_p \leq D\sqrt{p}\sqrt{\sum_{i=1}^n L_i^2}$ for some universal constant D. We may thus use Lemma 9 to bound the first summand and the Khintchine inequality (conditionally to $(X_i)_{i=1}^n$) to bound the other terms at the right-hand side of (3.2). In consequence we obtain

$$|| \sum_{i=1}^n X_i ||_p \leq \frac{C_K}{e} \sqrt{p} \sqrt{\sum_{i=1}^n L_i^2} \qquad (3.3)$$

for all $p \geq 2$. Let now t be an arbitrary nonnegative number. Define $p = \frac{t^2}{C_K^2 \sum_{i=1}^n L_i^2}$. If $p \geq 2$, the Chebyshev inequality yields

$$\mathbb{P}(|\sum_{i=1}^n X_i| \geq t) \leq \frac{\mathbb{E}|\sum_{i=1}^n X_i|^p}{t^p} \leq \frac{C_K^p p^{p/2} (\sum_{i=1}^n L_i^2)^{p/2}}{e^p t^p} = e^{-p}.$$

On the other hand, if $p < 2$, then $\mathbb{P}(|\sum_{i=1}^n X_i| \geq t) \leq e^2 e^{-p}$, which proves the Theorem.

□

Let us now introduce another lemma, which, together with Theorem 10 will allow us to derive a more general theorem, which may be considered a 'sub-Gaussian' version of the bounded differences inequality.

Lemma 11 *Let $\varphi\colon \mathbb{R} \to \mathbb{R}$ be a convex function and $S = f(X_1, \ldots, X_n)$, where $X_1, \ldots, X_n$ are independent random variables. Denote as usual $S_i = f(X_1, \ldots, X_{i-1}, \tilde{X}_i, X_{i+1}, \ldots, X_n)$, where $(\tilde{X}_1, \ldots, \tilde{X}_n)$ is an independent copy of $(X_1, \ldots, X_n)$ and assume that*

$$|S - S_i| \le F_i(X_i, \tilde{X}_i), \quad i = 1,\ldots,n.$$

Then

$$\mathbb{E}\varphi(S - \mathbb{E}S) \le \mathbb{E}\varphi(\sum_{i=1}^{n} \varepsilon_i F_i(X_i, \tilde{X}_i)), \tag{3.4}$$

where $\varepsilon_1, \ldots, \varepsilon_n$ is a sequence of independent Rademacher variables, independent of $(X_i)_{i=1}^n$ and $(\tilde{X}_i)_{i=1}^n$.

Proof. We will use induction with respect to n. For $n = 0$ the statement is obvious, since both the left-hand and the right-hand side of (3.4) equal $\varphi(0)$. Let us therefore assume that the Theorem is true for $n - 1$. Then

$$\begin{aligned}
\mathbb{E}\varphi(S - \mathbb{E}S) &= \mathbb{E}\varphi(S - \mathbb{E}_{\tilde{X}_n} S_n + \mathbb{E}_{X_n} S - \mathbb{E}S) \\
&\le \mathbb{E}\varphi(S - S_n + \mathbb{E}_{X_n} S - \mathbb{E}S) = \mathbb{E}\varphi(S_n - S + \mathbb{E}_{X_n} S - \mathbb{E}S) \\
&= \mathbb{E}\varphi(\varepsilon_n |S - S_n| + \mathbb{E}_{X_n} S - \mathbb{E}S) \\
&\le \mathbb{E}\varphi(\varepsilon_n F_n(X_n, \tilde{X}_n) + \mathbb{E}_{X_n} S - \mathbb{E}S),
\end{aligned}$$

with the last inequality following from Lemma 8. Now, denoting $Z = \mathbb{E}_{X_n} S$, $Z_i = \mathbb{E}_{X_n} S_i$, we have for $i = 1, \ldots, n-1$

$$|Z - Z_i| = |\mathbb{E}_{X_n} S - \mathbb{E}_{X_n} S_i| \le \mathbb{E}_{X_n} |S - S_i| \le F_i(X_i, \tilde{X}_i),$$

and thus for fixed X_n,$\tilde{X}_n$ and ε_n, we can apply the induction assumption to the function $t \mapsto \varphi(\varepsilon_n F(X_n, \tilde{X}_n) + t)$ instead of φ and $\mathbb{E}_{X_n}$ in the place of S, to obtain

$$\mathbb{E}\varphi(S - \mathbb{E}S) \le \mathbb{E}\varphi\left(\sum_{i=1}^{n} F_i(X_i, \tilde{X}_i)\varepsilon_i\right).$$

□

Remark Let us notice that we can now provide an alternate proof of the bounded differences inequality. Indeed if $|S - S_i| \le c_i$ for $i = 1, \ldots, n$, then, using the above lemma for $\varphi(t) = |t|^p$ we get for $p \ge 2$

$$\mathbb{E}|S - \mathbb{E}S|^p \le \mathbb{E}|\sum_{i=1}^{n} c_i \varepsilon_i|^p \le p^{p/2}\sqrt{\sum_{i=1}^{n} c_i^2}.$$

Thus, similarly as in the proof of Theorem 10, we obtain

$$\mathbb{P}(|S - \mathbb{E}S| \ge t) \le e^2 e^{-\frac{2t^2}{e^2 \sum_{i=1}^n c_i^2}},$$

which is (up to constants) the bounded difference inequality.

Theorem 11 *In the setting of Lemma 11, assume that for $i = 1, \ldots, n$ and all $t \ge 0$ we have*

$$\mathbb{P}(F_i(X_i, \tilde{X}_i) \ge t) \le K e^{-t^2/L_i^2}.$$

Then for all $t \ge 0$

$$\mathbb{P}(|S - \mathbb{E}S| \ge t) \le e^2 e^{-\frac{4t^2}{C_K^2 \sum_{i=1}^n L_i^2}}.$$

Proof. By Lemma 11 we have for $p \geq 2$

$$\mathbb{E}|S - \mathbb{E}S|^p \leq \mathbb{E}|\sum_{i=1}^{n} \varepsilon_i F(X_i, \tilde{X}_i)|^p.$$

But $F(X_i, \tilde{X}_i)$ are independent random variables and exactly as in the proof of Theorem 10 (inequalities (3.2) and (3.3)), we conclude that

$$\mathbb{E}|\sum_{i=1}^{n} \varepsilon_i F(X_i, \tilde{X}_i)|^p \leq \frac{C_K^p}{(2e)^p} \left(p \sum_{i=1}^{n} L_i^2 \right)^{p/2}.$$

Consider now $t \geq 0$ and define $p = \frac{4t^2}{C_K^2 \sum_{i=1}^n L_i^2}$. If $p \geq 2$, then

$$\mathbb{P}(|S - \mathbb{E}S| \geq t) \leq \frac{\mathbb{E}|S - \mathbb{E}S|^p}{t^p} \leq \frac{C_K^p p^{p/2} (\sum_{i=1}^n L_i^2)^{p/2}}{2^p e^p t^p} = e^{-p},$$

whereas if $p < 2$, we have $\mathbb{P}(|S - \mathbb{E}S| \geq t) \leq e^2 e^{-p}$.

□

Actually the following version of the above theorem is more useful in the applications

Theorem 12 *In the setting of Lemma 11, assume that for $i = 1, \ldots, n$*

$$\begin{aligned} F_i(X_i, \tilde{X}_i) &\leq G_i(X_i) + G_i(\tilde{X}_i) \\ \mathbb{P}(G(X_i) \geq t) &\leq K e^{-t^2/L_i^2}. \end{aligned}$$

for all $t \geq 0$. Then for all $t \geq 0$, we have

$$\mathbb{P}(|S - \mathbb{E}S| \geq t) \leq e^2 e^{-\frac{t^2}{2C_K^2 \sum_{i=1}^n L_i^2}}.$$

Proof. By Lemma 11 and Lemma 8, we have

$$||S - \mathbb{E}S||_p \leq ||\sum_{i=1}^{n} \varepsilon_i F(X_i, \tilde{X}_i)||_p \leq ||\sum_{i=1}^{n} \varepsilon_i (G(X_i) + G(\tilde{X}_i))||_p \leq 2||\sum_{i=1}^{n} \varepsilon_i G(X_i)||_p.$$

Thus for $p \geq 2$

$$||S - \mathbb{E}S||_p \leq \frac{C_K}{e} \sqrt{p} \sqrt{\sum_{i=1}^{n} L_i^2},$$

which implies the Theorem.

□

The following Corollary generalizes Theorem 10.

Corollary 9 *Let $X_1, \ldots, X_n$ be a sequence of independent random variables with values in a measurable space $(\Sigma, \mathcal{F})$ and $\mathcal{T}$ be a countable family of real measurable functions on Σ. Assume that for all $f \in \mathcal{T}$*

$$|f| \leq F$$

for some $F\colon \Sigma \to \mathbb{R}$, *satisfying*

$$\mathbb{P}(F(X_i) \geq t) \leq Ke^{-t^2/L_i^2}$$

for $i = 1, \ldots, n$. *Let now* S *be the random variable defined with the formula*

$$S = \sup_{f\in\mathcal{T}} \sum_{i=1}^{n} f(X_i).$$

Then, for all $t \geq 0$

$$\mathbb{P}(|S - \mathbb{E}S| \geq t) \leq e^2 e^{-\frac{t^2}{C_K^2 \sum_{i=1}^n L_i^2}}.$$

Proof. It is enough to check the assumption of Theorem 12. We have

$$S - S_i \leq \sup_{f\in\mathcal{T}} (f(X_i) - f(\tilde{X}_i)) \leq F(X_i) + F(\tilde{X}_i),$$

which by symmetry yields

$$|S - S_i| \leq F(X_i) + F(\tilde{X}_i).$$

□

3.2. General moment inequalities

Now we are going to show how the moment method can be linked with the entropy method. We will first state and prove a general moment inequality from ([2]), and then apply it to obtain some tail and moment estimates for U-statistics in Banach spaces.

Theorem 13 *Let* $X_1, \ldots, X_n$ *be independent random variables taking values in a measurable space* $(\Sigma, \mathcal{F})$ *and* $f\colon \Sigma^n \to \mathbb{R}$ *a measurable function (with respect to the product σ-field). Denote* $S = f(X_1, \ldots, X_n)$, $S_i = f(X_1, \ldots, X_{i-1}, \tilde{X}_i, X_{i+1}, \ldots, X_n)$, *where* $(X_1, \ldots, X_n)$ *and* $(\tilde{X}_1, \ldots, \tilde{X}_n)$ *are independent random vectors, equal in distribution. Define* $V = \sum_{i=1}^{N} \mathbb{E}_{\tilde{X}_i}(S - S_i)_+^2$. *Then for all* $p \in \mathbb{N}$, $p \geq 2$

$$\mathbb{E}(S - \mathbb{E}S)_+^p \leq 2^{p/2}\kappa_p^{p/2}\left(1 - \frac{1}{p}\right)^{p/2} p^{p/2}\mathbb{E}V^{p/2} \leq 2^{p/2}\kappa^{p/2}\mathbb{E}V^{p/2},$$

where

$$\kappa_p = \frac{1}{2}\left(1 - \left(1 - \frac{1}{p}\right)^{p/2}\right)^{-1}$$

and

$$\kappa = \lim_{p\to\infty} \kappa_p = \frac{\sqrt{e}}{2(\sqrt{e} - 1)}.$$

To prove the above Theorem we shall follow the arguments from [2]. First we need to examine some properties of the functional $E_\alpha(X) = \mathbb{E}X^\alpha - (\mathbb{E}X)^\alpha$, for $\alpha \in (1, 2]$. Recall from Chapter 1, that E satisfies the convexity condition (1.1).

Lemma 12 *Let* X *be a nonnegative, integrable random variable and* Y *an independent copy of* X. *Then*

$$E_\alpha(X) \leq \mathbb{E}(X - Y)_+(X^{\alpha-1} - Y^{\alpha-1}).$$

Proof. From the concavity of the function $x \mapsto x^{\alpha-1}$, we have

$$\begin{aligned}
E_\alpha(X) &= \mathbb{E}X^\alpha - (\mathbb{E}X)^\alpha \\
&= \mathbb{E}X^\alpha - (\mathbb{E}X)(\mathbb{E}Y)^{\alpha-1} \\
&\le \mathbb{E}X^\alpha - (\mathbb{E}X)\mathbb{E}Y^{\alpha-1} \\
&= \mathbb{E}X(X^{\alpha-1} - Y^{\alpha-1}) \\
&= \frac{1}{2}\mathbb{E}(X-Y)(X^{\alpha-1} - Y^{\alpha-1}) \\
&= \mathbb{E}(X-Y)_+(X^{\alpha-1} - Y^{\alpha-1}).
\end{aligned}$$

□

Lemma 13 *In the setting of Theorem 13, let $p \ge 2$ and let α satisfy $p/2 \le \alpha \le p-1$. Let us assume that $\mathbb{E}(S - \mathbb{E}S)_+^p < \infty$. Then*

$$\mathbb{E}(S - \mathbb{E}S)_+^p \le (\mathbb{E}(S - \mathbb{E}S)_+^\alpha)^{p/\alpha} + \alpha(p-\alpha)\mathbb{E}V(S - \mathbb{E}S)_+^{p-2}.$$

Proof. The statement of the lemma can be expressed in terms of $E_{p/\alpha}$ $(p/\alpha \in (1,2])$ as

$$E_{p/\alpha}((S - \mathbb{E}S)_+^\alpha) \le \alpha(p-\alpha)\mathbb{E}V(S - \mathbb{E}S)_+^{p-2}. \tag{3.5}$$

Thus, to prove the lemma, it is enough to show, that for every number $m \in \mathbb{R}$, such that $\mathbb{E}(S-m)_+^p < \infty$, we have

$$E_{p/\alpha}(F(S)) \le \alpha(p-\alpha)\mathbb{E}V(S-m)_+^{p-2}, \tag{3.6}$$

where $F(s) = (s-m)_+^\alpha$ (since (3.5) follows from (3.6) by substituting $m = \mathbb{E}S$). Now, by the tensorization property of E (Theorem 3) we can restrict our attention to the case $n = 1$. We have thus $V = \mathbb{E}_Y(S-Y)_+^2$ where Y is an independent copy of S. Since F is non-decreasing, by Lemma 12, we have

$$\begin{aligned}
E_{p/\alpha}(F(S)) &\le \mathbb{E}(F(S) - F(Y))\mathbf{1}_{\{S \ge Y\}}(F(S)^{p/\alpha-1} - F(Y)^{p/\alpha-1}) \\
&= \mathbb{E}(F(S) - F(Y))\mathbf{1}_{\{S \ge Y\}}((S-m)_+^{p-\alpha} - (Y-m)_+^{p-\alpha}). \qquad (3.7)
\end{aligned}$$

But both F and the function $x \mapsto (x-m)_+^{p-\alpha}$ are convex and non-decreasing, and thus for $x \ge y$ we have

$$\begin{aligned}
0 &\le F(x) - F(y) \le (x-y)\alpha(x-m)_+^{\alpha-1} \\
0 &\le (x-m)_+^{p-\alpha} - (y-m)_+^{p-\alpha} \le (x-y)(p-\alpha)(x-m)_+^{p-\alpha-1},
\end{aligned}$$

hence

$$(F(S) - F(Y))\mathbf{1}_{\{S \ge Y\}}((S-m)_+^{p-\alpha} - (Y-m)_+^{p-\alpha}) \le \alpha(p-\alpha)(S-Y)_+^2(S-m)_+^{p-2},$$

which together with (3.7) proves the 1-dimensional version of (3.6).

□

Proof of Theorem 13. The proof will consist of two parts. The first part will constitute the Theorem for random variables S, such that $\mathbb{E}(S-\mathbb{E}S)_+^p < \infty$. We will use the induction with respect to p. We have $\kappa_2 = 1$, so for $p = 2$ the statement of the Theorem is

$$\mathbb{E}(S-\mathbb{E}S)_+^2 \leq 2\mathbb{E}\sum_{i=1}^{n}(S-S_i)_+^2.$$

But $\mathbb{E}(S-\mathbb{E}S)_+^2 \leq \mathbb{E}S^2 - (\mathbb{E}S)^2$, so it is enough to prove

$$\mathrm{Var}S \leq 2\mathbb{E}\sum_{i=1}^{n}(S-S_i)_+^2$$

and due to the tensorization property of the variance, we can restrict to $n = 1$. Let thus X, Y be i.i.d. random variables. Then

$$\mathrm{Var}X = \mathbb{E}(X-\mathbb{E}Y)^2 \leq \mathbb{E}(X-Y)^2 = 2\mathbb{E}(X-Y)_+^2.$$

Let us now proceed with the induction step. By Hölder's inequality, for non-negative random variables Y, we have

$$\mathbb{E}Y(S-\mathbb{E}S)_+^{p-2} \leq ||Y||_{p/2}||(S-\mathbb{E}S)_+||_p^{p-2}.$$

Now, by Lemma 13, applied with $\alpha = p-1$, we obtain

$$\mathbb{E}(S-\mathbb{E}S)_+^p \leq \left(\mathbb{E}(S-\mathbb{E}S)_+^{p-1}\right)^{\frac{p}{p-1}} + (p-1)||V||_{p/2}||(S-\mathbb{E}S)_+||_p^{p-2}.$$

If we denote $c_p = 2||V||_{p/2}(1-1/p)$ and $x_p = (\mathbb{E}(S-\mathbb{E}S)_+^p)(p\kappa_p c_p)^{-p/2}$, the above inequality translates as

$$x_p p^{p/2} c_p^{p/2}\kappa_p^{p/2} \leq x_{p-1}^{p/(p-1)}(p-1)^{p/2}c_{p-1}^{p/2}\kappa_{p-1}^{p/2} + \frac{1}{2}x_p^{1-2/p}p^{p/2}c_p^{p/2}\kappa_p^{p/2-1}.$$

But $\kappa_{p-1} \leq \kappa_p$, $c_{p-1} \leq c_p$ and by the induction assumption $x_{p-1} \leq 1$, so this inequality yields

$$x_p \leq \left(1-\frac{1}{p}\right)^{p/2} + \frac{1}{2\kappa_p}x_p^{1-2/p}.$$

Consider now the function f_p, defined on $\mathbb{R}_+$ as

$$f_p(x) = \left(1-\frac{1}{p}\right)^{p/2} + \frac{1}{2\kappa_p}x^{1-2/p} - x.$$

Since f_p' is decreasing, f_p is strictly concave. Moreover, $f_p(0) > 0$ and $f_p(1) = 0$, so for $x > 1$ we have $f_p(x) < 0$. Thus $f_p(x_p) \geq 0$ implies $x_p \leq 1$.

What still remains to be done is to prove the Theorem for S, such that $\mathbb{E}(S-\mathbb{E}S)_+^p = \infty$. We want to show that then also $\mathbb{E}V^{p/2} = \infty$. To prove it we will once again use the induction, this time with respect to the number of coordinates n. Let $n = 1$ and Y be an independent copy of S. By Jensen's inequality we have

$$\mathbb{E}(S-\mathbb{E}S)_+^p = \mathbb{E}(S-\mathbb{E}Y)_+^p \leq \mathbb{E}(\mathbb{E}_Y(S-Y)_+^2)^{p/2} = \mathbb{E}V^{p/2},$$

which proves the Theorem in the case $n = 1$. For $n > 1$, let us notice that

$$||(S-\mathbb{E}S)_+||_p \leq ||(S-\mathbb{E}_{X_n}S)_+||_p + ||(\mathbb{E}_{X_n}S-\mathbb{E}S)_+||_p,$$

since for $x, y \in \mathbb{R}$, we have $(x+y)_+ \leq x_+ + y_+$. Thus if $\mathbb{E}(S - \mathbb{E}S)_+ = \infty$, then $\mathbb{E}(S - \mathbb{E}_{X_n} S)_+^p = \infty$ or $\mathbb{E}(\mathbb{E}_{X_n} S - \mathbb{E}S)_+^p = \infty$. But we have

$$\mathbb{E}(S - \mathbb{E}_{X_n} S)_+^p = \mathbb{E}(S - \mathbb{E}_{\tilde{X}_n} S_n)_+^p \leq \mathbb{E}(\mathbb{E}_{\tilde{X}_n}(S - S_n)_+^2)^{p/2} \leq \mathbb{E}V^{p/2},$$

so in the first case the Theorem is satisfied. On the other hand

$$\mathbb{E}(\sum_{i=1}^{n-1} \mathbb{E}_{\tilde{X}_i}(\mathbb{E}_{X_n} S - \mathbb{E}_{X_n} S_i)_+^2)^{p/2} \leq \mathbb{E}(\sum_{i=1}^{n-1} \mathbb{E}_{X_n} \mathbb{E}_{\tilde{X}_i}(S - S_i)_+^2)^{p/2} \leq \mathbb{E}(\sum_{i=1}^{n-1} \mathbb{E}_{\tilde{X}_i}(S - S_i)_+^2)^{p/2} \leq \mathbb{E}V^{p/2}.$$

If $\mathbb{E}(\mathbb{E}_{X_n} S - \mathbb{E}S)_+^p = \infty$, then (by the induction assumption) the left hand side of the above inequality is also infinite and so is $\mathbb{E}V^{p/2}$, which proves the Theorem.

□

3.2.1. Application to U-statistics

Let now B be a separable Banach space, such that B^* is separable. Let $X_1, \ldots, X_N$, $Y_1, \ldots, Y_N$ be independent random variables, with values in a Polish space Σ and $h \colon \Sigma \times \Sigma \to B$ be a measurable function. Assume that $\mathbb{E}_{X_i} h(X_i, Y_j) = 0$ and $\mathbb{E}_{Y_j} h(X_i, Y_j) = 0$ a.e. and define

$$Z = || \sum_{i,j=1}^{N} h(X_i, Y_j) ||.$$

We will need some additional facts, that will be stated without proofs.

Fact 1 (Theorem 11 in [2]) *Let $X_1, \ldots, X_n$ be a sequence of independent random variables with values in a measurable space $(\Sigma, \mathcal{F})$ and $\mathcal{T}$ be a countable family of nonnegative measurable functions on Σ. Let $S = \sup_{f \in \mathcal{T}} \sum_{i=1}^{n} f(X_i)$. Then there exists a universal constant K, such that for $p = 2, 3, \ldots$ we have*

$$\mathbb{E}S^p \leq K^p((\mathbb{E}S)^p + p^p \mathbb{E} \max_{i \leq 1 \leq n} \sup_{f \in \mathcal{T}} f(X_i)^p).$$

Fact 2 (Proposition 3.1. in [9]) *Let $X_1, \ldots, X_n$ be a sequence of independent random variables with values in a measurable space $(\Sigma, \mathcal{F})$ and $\mathcal{T}$ be a countable family of measurable functions on Σ. Assume furthermore that for each $f \in \mathcal{T}$ we have $\mathbb{E}f(X_i) = 0$ for all i. Consider the random variable*

$$S = \sup_{f \in \mathcal{T}} | \sum_{i=1}^{n} f(X_i) |.$$

Define now

$$\sigma^2 = \sup_{f \in \mathcal{T}} \sum_{i=1}^{n} \mathbb{E}f(X_i)^2.$$

Then there exists a universal constant K such that

$$\mathbb{E}S^p \leq K^p((\mathbb{E}S)^p + p^{p/2}\sigma^p + p^p \mathbb{E} \max_{1 \leq i \leq n} \sup_{f \in \mathcal{T}} |f(X_i)|^p).$$

for $p = 2, 3, \ldots$

Let us stress here that both of the above facts can be proved using Theorem 13. We refer to [2] for details. The latter Fact was first proved in [9] for all $p \geq 2$, non necessarily natural, from the upper tail bound for the random variable S.

Corollary 10 *Let $X_1, \ldots, X_n$ be independent centered random variables with values in a Banach space B, such that B^* is separable. Then there exists a universal constant K, such that for all $p \geq 2$ we have the following estimate*

$$\mathbb{E}||\sum_{i=1}^n X_i||^p \leq K^p((\mathbb{E}||\sum_{i=1}^n X_i||)^p + p^{p/2}(\sum_{i=1}^n \mathbb{E}||X_i||^2)^{p/2} + p^p\mathbb{E}\sum_{j=1}^n ||X_i||^p).$$

Proof. The proof involves just expressing the norm $||\cdot||$ as $\sup_v\langle v, \cdot\ \rangle$ over a countable set of elements $v \in B^*$ and applying Fact 2.

□

The next theorem is an improvement of classical Rosenthal inequalities, due to R. Latała ([11],[9])

Fact 3 (Inequality R1 in [9]) *Let $X_1, \ldots, X_n$ be independent, nonnegative random variables. Then for all $p \geq 1$*

$$\mathbb{E}(\sum_{i=1}^n X_i)^p \leq (2e)^p \max\left(\frac{e}{p}p^p\sum_{i=1}^n \mathbb{E}X_i^p, e^p(\sum_{i=1}^n \mathbb{E}X_i)^p\right).$$

Fact 4 (Inequality (2.6) in [9]) *Let $X_1, \ldots, X_n$ be independent nonnegative random variables. Then for all $p > 1$ and $\alpha \geq 0$*

$$p^{\alpha p}\sum_{i=1}^n \mathbb{E}X_i^p \leq 2(1+p^\alpha)\max\left(p^{\alpha p}\mathbb{E}\max_{1\leq i\leq n} X_i^p, (\sum_{i=1}^n \mathbb{E}X_i)^p\right).$$

We will also use the following technical lemma

Lemma 14 *Let B be a Banach space such that B^* is separable. Let Σ be a Polish space, equipped with a Borel probability measure. Then there exists a countable set $\mathcal{T}$ of functions $g\colon \Sigma \to B^*$ with $\mathbb{E}||g||^2 \leq 1$, such that*

$$(\mathbb{E}||f||^2)^{1/2} = \sup_{g\in\mathcal{T}} \mathbb{E}\langle g, f\rangle$$

for every measurable function $f\colon \Sigma \to B$, such that $\mathbb{E}||f||^2 = 1$.

Proof. For every vector $v \in B$ let $\Gamma(v) = \{w \in B^*\colon \langle w, v\rangle = ||v||, ||w|| = 1\}$. By the Hahn-Banach Theorem $\Gamma(v) \neq \emptyset$ for every v. Moreover $\Gamma(v)$ is closed in B^* and hence complete in the metric induced from B^*. We would like to choose a measurable function $\tilde{g}\colon B \to B^*$ such that for every $v \in B$, $\tilde{g}(v) \in \Gamma(v)$ (i.e. $\tilde{g}$ is a measurable selection of Γ). For this purpose we will use the following theorem, which can be found in ([5]), p. 65.

Fact 5 *Let X be separable metric space, $(T, \mathcal{F})$ a measurable space, Γ a multifunction from T to the collection of complete, nonempty subsets of X. If for each open set $U \subseteq X$, $\Gamma^{-1}(U) = \{t \colon \Gamma(t) \cap U \neq \emptyset\} \in \mathcal{F}$, then Γ admits a measurable selection.*

For an open set $U \subset B^*$, let us consider the set $\Gamma^{-1}(U) = \{v \in B \colon \Gamma(v) \cap U \neq \emptyset\}$. Since $\Gamma^{-1}(\bigcup U_i) = \bigcup \Gamma^{-1}(U_i)$ and every open subset of B^* is a countable union of open balls, to check the assumption of the above fact, it is enough to prove the Borel measurability of $\Gamma^{-1}(U)$ in the case when U is an open ball. Let thus w,r denote respectively the centre and the radius of U.

Let $\mathcal{A} = \{w_1, w_2, \ldots\}$ be a countable set, dense in the unit sphere of B^*. If $v \in \Gamma^{-1}(U)$, then there exists $w_\infty \in U$, with $||w_\infty|| = 1$, $\langle w, v\rangle = ||v||$. Thus for some $\varepsilon > 0$, there exists a sequence $w_n \in \mathcal{A}$, $||w_n - w|| < r - \varepsilon$, such that $\lim_{n\to\infty} \langle w_n, v\rangle = ||v||$.

On the other hand, if there exists such a sequence, then there exists a subsequence w_{n_k}, converging to some w_∞ in the $*$-weak topology. Then $\langle w_\infty, v\rangle = ||v||$ and $||w - w_\infty|| \leq r - \varepsilon$, $||w|| = 1$, so $w_\infty \in U$ and $v \in \Gamma^{-1}(U)$. Thus

$$\Gamma^{-1}(U) = \bigcup_{\varepsilon\in\mathbb{Q}^+} \bigcap_{\rho\in\mathbb{Q}^+} \bigcup_{u\in\mathcal{A}, ||u-w||<r-\varepsilon} \{v \colon |\langle u, v\rangle - \|v\|| < \rho\}.$$

Since $\{v \colon |\langle u, v\rangle - ||v||| < \rho\}$ is closed in B, we conclude that $\Gamma^{-1}(U)$ is Borel measurable.

We have thus proved that there exists a measurable function $\tilde{g} \colon B \to B^*$, such that $||v|| = \langle \tilde{g}(x), v\rangle$ and $||\tilde{g}(v)|| = 1$ for all $v \in B$. Thus, for every $f \in L^2(\Sigma, B)$ there exists $g \in L^2(\Sigma, B^*)$, such that $||f(x)|| = \langle g(x), f(x)\rangle$ and $||g(x)|| = 1$ for all $x \in \Sigma$.

Now we are ready to construct the set $\mathcal{T}$. Let $\mathcal{B} = \{w_1, w_2, \ldots\}$ be a countable set, dense in B^*. Every function from $L^2(\Sigma, B^*)$ can be approximated in this space by bounded functions and such functions can be approximated by $\mathcal{B}$-valued *step functions* i.e. functions of the form

$$h(x) = \sum_{i=1}^{n} w_i \mathbf{1}_{A_i}(x),$$

where A_i are Borel subsets of Σ. Now, since every Borel measure on a Polish space is regular, we can approximate such step functions by $\mathcal{B}$-valued step functions such that every set A_i is a finite sum of open sets from a countable basis. All such functions constitute a countable set, which we will denote by $\mathcal{S}$.

Recall, that for fixed $f \in L^2(\Sigma, B)$, we denote by g a function from $L^2(\Sigma, B^*)$, such that $||f(x)|| = \langle g(x), f(x)\rangle$ and $||g(x)|| = 1$ for all $x \in \Sigma$. Define $h = g||f||/(\mathbb{E}||f||^2)^{1/2}$. We have

$$\begin{aligned} \mathbb{E}||h||^2 &= 1, \\ \mathbb{E}\langle h, f\rangle &= (\mathbb{E}||f||^2)^{1/2}. \end{aligned}$$

Consider a sequence $\tilde{g}_n \in \mathcal{S}$, such that $\tilde{g}_n \to h$ in $L^2(\Sigma, B^*)$. Then

$$|\mathbb{E}\langle \tilde{g}_n, f\rangle - (\mathbb{E}||f||^2)^{1/2}| \leq \mathbb{E}|\langle \tilde{g}_n - h, f\rangle| \leq \mathbb{E}(||\tilde{g}_n - h|| \cdot ||f||) \leq (\mathbb{E}||\tilde{g}_n - h||^2)^{1/2}(\mathbb{E}||f||^2)^{1/2}.$$

The expression at the right-hand side converges to 0 as $n \to \infty$, so we get

$$\lim_{n\to\infty} \langle \tilde{g}_n, f\rangle = (\mathbb{E}||f||^2)^{1/2}.$$

Moreover $(\mathbb{E}||h||^2)^{1/2} - (\mathbb{E}||h - \tilde{g}_n||^2)^{1/2} \leq (\mathbb{E}||\tilde{g}_n||^2)^{1/2} \leq (\mathbb{E}||h||^2)^{1/2} + (\mathbb{E}||h - \tilde{g}_n||^2)^{1/2}$, so $\lim_{n\to\infty} \mathbb{E}||\tilde{g}_n||^2 = 1$. Define $g_n = \tilde{g}_n/(\mathbb{E}||\tilde{g}_n||^2)^{1/2}$. We have

$$\lim_{n\to\infty} \mathbb{E}\langle g_n, f\rangle \quad = \quad \lim_{n\to\infty} \frac{\mathbb{E}\langle \tilde{g}_n, f\rangle}{(\mathbb{E}||\tilde{g}_n||^2)^{1/2}} = \frac{(\mathbb{E}||f||^2)^{1/2}}{1} = (\mathbb{E}||f||^2)^{1/2},$$

On the other hand, for every $g \in L^2(\Sigma, B^*)$, with $\mathbb{E}||g||^2 \leq 1$, we have $\mathbb{E}\langle g, f\rangle \leq \mathbb{E}||f||||g|| \leq (\mathbb{E}||f||^2)^{1/2}$, so

$$(\mathbb{E}||f||^2)^{1/2} = \sup\{\mathbb{E}\langle \frac{g}{(\mathbb{E}||g||^2)^{1/2}}, f\rangle \colon g \in \mathcal{S}\}$$

Since the set $\mathcal{T} = \{g/(E||g||^2)^{1/2} \colon g \in \mathcal{S}\}$ is countable and for all $h \in \mathcal{T}$ we have $\mathbb{E}||h||^2 = 1$, the lemma has been proved.

□

We will now use the moment method to find a bound for the upper tail of Z, following the idea from [11]. For convenience and consistency with the previous part we will use sometimes the notation $X_{N+i} = Y_i$. Let $\mathcal{T}$ be a countable set dense in the unit ball of B^*. By the Hahn-Banach theorem we have

$$Z = \sup_{v\in\mathcal{T}} \langle v, \sum_{i,j=1}^{N} h(X_i, Y_j)\rangle$$

Let us now fix a sample $(X_i)_{i=1}^{2N}$ and consider a sequence $v_n \in \mathcal{T}$, such that

$$\lim_{n\to\infty} \langle v_n, \sum_{i,j=1}^{n} h(X_i, Y_j)\rangle = Z.$$

Pointwise, we have

$$\sum_{k=1}^{N} (Z - Z_k)_+^2 = \sum_{k=1}^{N} \lim_{n\to\infty} (\langle v_n, \sum_{i,j=1}^{N} h(X_i, Y_j)\rangle - Z_k)_+^2$$

and by the Lebesgue dominated convergence theorem

$$\mathbb{E}_{\tilde{X}} \sum_{k=1}^{N} (Z - Z_k)_+^2 = \lim_{n\to\infty} \mathbb{E}_{\tilde{X}} \sum_{k=1}^{N} (\langle v_n, \sum_{i,j=1}^{N} h(X_i, Y_j)\rangle - Z_k)_+^2.$$

But for each n

$$\begin{aligned}
\sum_{k=1}^{N} \mathbb{E}_{\tilde{X}} (\langle v_n, \sum_{i,j=1}^{n} h(X_i, Y_j)\rangle - Z_k)_+^2 &\leq \sum_{k=1}^{N} \mathbb{E}_{\tilde{X}} (\langle v_n, \sum_{j=1}^{N} h(X_k, Y_j) - h(\tilde{X}_k, Y_j)\rangle)^2 \\
&= \sum_{k=1}^{N} \langle v_n, \sum_{j=1}^{N} h(X_k, Y_j)\rangle^2 + \sum_{k=1}^{N} \mathbb{E}_X \langle v_n, \sum_{j=1}^{N} h(X_k, Y_j)\rangle^2 \\
&\leq \sup_{v\in\mathcal{T}} \sum_{k=1}^{N} \langle v, \sum_{j=1}^{N} h(X_k, Y_j)\rangle^2 + \sup_{v\in\mathcal{T}} \sum_{k=1}^{N} \mathbb{E}_X \langle v, \sum_{j=1}^{N} h(X_k, Y_j)\rangle^2
\end{aligned}$$

with the equality following from the assumption $\mathbb{E}_X h = 0$ a.e. After handling the case of $k > N$ in an analogous way we finally obtain

$$\begin{aligned}
\sum_{k=1}^{2N} \mathbb{E}_{\tilde{X}} (Z - Z_k)_+^2 &\leq \sup_{v\in\mathcal{T}} \sum_{i=1}^{N} \langle v, \sum_{j=1}^{N} h(X_i, Y_j)\rangle^2 + \sup_{v\in\mathcal{T}} \sum_{i=1}^{N} \mathbb{E}_X \langle v, \sum_{j=1}^{N} h(X_i, Y_j)\rangle^2 \qquad (3.8) \\
&+ \sup_{v\in\mathcal{T}} \sum_{j=1}^{N} \langle v, \sum_{i=1}^{N} h(X_i, Y_j)\rangle^2 + \sup_{v\in\mathcal{T}} \sum_{j=1}^{N} \mathbb{E}_Y \langle v, \sum_{i=1}^{N} h(X_i, Y_j)\rangle^2
\end{aligned}$$

Thus by Theorem 13 we get

$$\begin{aligned}
\mathbb{E}(Z-\mathbb{E}Z)^p &\leq K^p p^{p/2}\Bigg(\mathbb{E}(\sup_{v\in\mathcal{T}}\sum_{i=1}^N\langle v,\sum_{j=1}^N h(X_i,Y_j)\rangle^2)^{p/2}+\mathbb{E}_Y(\sup_{v\in\mathcal{T}}\sum_{i=1}^N\mathbb{E}_X\langle v,\sum_{j=1}^N h(X_i,Y_j)\rangle^2)^{p/2} \\
&+ \mathbb{E}(\sup_{v\in\mathcal{T}}\sum_{j=1}^N\langle v,\sum_{i=1}^N h(X_i,Y_j)\rangle^2)^{p/2}+\mathbb{E}_X(\sup_{v\in\mathcal{T}}\sum_{j=1}^N\mathbb{E}_Y\langle v,\sum_{i=1}^N h(X_i,Y_j)\rangle^2)^{p/2}\Bigg) \qquad (3.9)\\
&\leq K^p p^{p/2}(A+B+C+D).
\end{aligned}$$

Let us notice that two latter terms are analogous to the former, so in what follows we will not put attention to them in any of partial computations, but just include their influence at the final steps. Let us thus handle the first term at the right hand side of (3.9), denoting for the time being

$$S=\sup_{v\in\mathcal{T}}\sum_{i=1}^N\langle v,\sum_{j=1}^N h(X_i,Y_j)\rangle^2.$$

Fact 1, applied conditionally to Y gives

$$\begin{aligned}
\mathbb{E}S^{p/2} &\leq K^p\left(\mathbb{E}_Y(\mathbb{E}_X\sup_{v\in\mathcal{T}}\sum_{i=1}^N\langle v,\sum_{j=1}^N h(X_i,Y_j)\rangle^2)^{p/2}+p^{p/2}\mathbb{E}\max_{1\leq i\leq N}\sup_{v\in\mathcal{T}}\langle v,\sum_{j=1}^N h(X_i,Y_j)\rangle^p\right) \\
&= K^p\left(\mathbb{E}_Y(\mathbb{E}_X\sup_{v\in\mathcal{T}}\sum_{i=1}^N\langle v,\sum_{j=1}^N h(X_i,Y_j)\rangle^2)^{p/2}+p^{p/2}\mathbb{E}\max_{1\leq i\leq N}||\sum_{j=1}^N h(X_i,Y_j)||^p\right).
\end{aligned}$$

Since the first term at the right-hand side of the last inequality is greater then the second term at the right-hand side of (3.9), after taking into account the analogous contributions from C and D we get

$$\begin{aligned}
\mathbb{E}(Z-\mathbb{E}Z)_+^p &\leq K^p\Bigg(p^{p/2}\mathbb{E}_Y(\mathbb{E}_X\sup_{v\in\mathcal{T}}\sum_{i=1}^N\langle v,\sum_{j=1}^N h(X_i,Y_j)\rangle^2)^{p/2}+p^p\mathbb{E}\max_{1\leq i\leq N}||\sum_{j=1}^N h(X_i,Y_j)||^p \\
&+ p^{p/2}\mathbb{E}_X(\mathbb{E}_Y\sup_{v\in\mathcal{T}}\sum_{j=1}^N\langle v,\sum_{i=1}^N h(X_i,Y_j)\rangle^2)^{p/2}+p^p\mathbb{E}\max_{1\leq j\leq N}||\sum_{i=1}^N h(X_i,Y_j)||^p\Bigg).
\end{aligned} \qquad (3.10)$$

Obviously

$$\begin{aligned}
(\mathbb{E}_X\sup_{v\in\mathcal{T}}\sum_{i=1}^N\langle v,\sum_{j=1}^N h(X_i,Y_j)\rangle^2)^{1/2} &\leq (\mathbb{E}_X\sum_{i=1}^N\sup_{v\in\mathcal{T}}\langle v,\sum_{j=1}^N h(X_i,Y_j)\rangle^2)^{1/2} \\
&= (\mathbb{E}_X\sum_{i=1}^N||\sum_{j=1}^N h(X_i,Y_j)||^2)^{1/2} =: S.
\end{aligned}$$

This estimate is quite crude, however it will allow us to replace the 'troublesome' random variable by one that can be handled with the use of Fact 2. Indeed, by Lemma 14, there exists a countable set $\mathcal{V}$, consisting of elements $f = (f_1, \ldots, f_n)$, such that for each i, $f_i \colon \Sigma \to B^*$ and $\sum_{i=1}^{n} \mathbb{E}||f_i(X_i)||^2 \leq 1$ and

$$S = \sup_{f\in\mathcal{V}} |\sum_{i=1}^{N}\sum_{j=1}^{N} \mathbb{E}_X \langle f_i(X_i), h(X_i, Y_j)\rangle|.$$

Hence, identifying $f \in \mathcal{V}$ with the function $y \mapsto \sum_i \mathbb{E}_X \langle f_i(X_i), h(X_i, Y_j)\rangle$, we have

$$S = \sup_{f\in\mathcal{V}} |\sum_{j=1}^{N} f(Y_j)|,$$

so we can apply to S the inequality from Fact 2. In this case we get

$$\begin{aligned}\sigma^2 &= \sup_{f\in\mathcal{V}} \mathbb{E}_Y \sum_{j=1}^{N} \left(\sum_{i=1}^{N} \mathbb{E}_X \langle f_i(X_i), h(X_i, Y_j)\rangle \right)^2 \\ &\leq \sup \left\{ \mathbb{E} \sum_{i,j=1}^{N} \langle f_i(X_i), h(X_i, Y_j)\rangle g_j(Y_j) \colon\, f_i \colon \Sigma \to B^*,\ g_j \colon \Sigma \to \mathbb{R}, \right. \\ &\qquad \left. \sum_{i=1}^{N} \mathbb{E}||f_i(X_i)||^2 \leq 1,\ \sum_{j=1}^{N} \mathbb{E} g_j(Y_j)^2 \leq 1 \right\}^2\end{aligned}$$

For simplicity reasons and analogy with the real-valued case, let us denote the square root of the right-hand side by $||h||^{(1)}_{L^2 \to L^2}$. Similarly, we define

$$\begin{aligned}||h||^{(2)}_{L_2 \to L_2} &= \sup \left\{ \mathbb{E} \sum_{i,j=1}^{N} \langle f_j(Y_j), h(X_i, Y_j)\rangle g_i(X_i) \colon\, f_j \colon \Sigma \to B^*,\ g_i \colon \Sigma \to \mathbb{R}, \right. \\ &\qquad \left. \sum_{j=1}^{N} \mathbb{E}||f_j(Y_j)||^2 \leq 1,\ \sum_{i=1}^{N} \mathbb{E} g_i(X_i)^2 \leq 1 \right\}.\end{aligned}$$

Now $|\mathbb{E}S|^2 \leq \mathbb{E}S^2 = \mathbb{E}\sum_{i=1}^{N} || \sum_{j=1}^{N} h(X_i, Y_j)||^2$ and finally

$$\begin{aligned}\mathbb{E}_Y \max_{1\leq j\leq N} \sup_{f\in\mathcal{V}} |f(Y_j)|^p &= \mathbb{E}_Y \max_{1\leq j\leq N} \sup_{f\in\mathcal{V}} |\sum_{i=1}^{N} \mathbb{E}_X \langle f_i(X_i), h(X_i, Y_j)\rangle|^p \\ &\leq \mathbb{E}_Y \max_{1\leq j\leq N} \sup_{f\in\mathcal{V}} (\sum_{i=1}^{N} \mathbb{E}_X ||f_i(X_i)|| \cdot ||h(X_i, Y_j)||)^p \\ &\leq \mathbb{E}_Y \max_{1\leq j\leq N} (\sum_{i=1}^{N} ||h(X_i, Y_j)||^2)^{p/2}.\end{aligned}$$

After collecting the above estimations, using Fact 2 and plugging the result into (3.10), we obtain

$$\begin{aligned}
\mathbb{E}(Z-\mathbb{E}Z)_+^p \quad &\leq \quad K^p \left(p^{p/2}(\mathbb{E}\sum_{i=1}^N ||\sum_{j=1}^N h(X_i,Y_j)||^2)^{p/2} + p^p(||h||^{(1)}_{L_2\to L_2})^p \right. && (3.11)\\
&+ \quad p^{3p/2}\mathbb{E}_Y \max_{1\leq j\leq N}(\sum_{i=1}^N ||h(X_i,Y_j)||^2)^{p/2} + p^p\mathbb{E}\sum_{i=1}^N ||\sum_{j=1}^N h(X_i,Y_j)||^p \\
&+ \quad p^{p/2}(\mathbb{E}\sum_{j=1}^N ||\sum_{i=1}^N h(X_i,Y_j)||^2)^{p/2} + p^p(|h||^{(2)}_{L_2\to L_2})^p \\
&+ \quad \left. p^{3p/2}\mathbb{E}_X \max_{1\leq i\leq N}(\sum_{j=1}^N ||h(X_i,Y_j)||^2)^{p/2} + p^p\mathbb{E}\sum_{j=1}^N ||\sum_{i=1}^N h(X_i,Y_j)||^p \right).
\end{aligned}$$

Let us note that the fourth and the eight terms at the right-hand side have been obtained by changing maximum into sum in the appropriate term from (3.10). We will now handle the fourth term by applying conditionally to X Corollary 10 (we stick to the introduced convention to derive only one of two analogous terms, derived from X and Y part of (3.9) respectively).

What we get is

$$\begin{aligned}
p^p\mathbb{E}\sum_{i=1}^N ||\sum_{j=1}^N h(X_i,Y_j)||^p \quad &\leq \quad K^p\left(p^p\mathbb{E}_X\sum_{i=1}^N(\mathbb{E}_Y||\sum_{j=1}^N h(X_i,Y_j)||)^p \right. && (3.12)\\
&+ \quad \left. p^{3p/2}\mathbb{E}_X\sum_{i=1}^N(\sum_{j=1}^N \mathbb{E}_Y||h(X_i,Y_j)||^2)^{p/2} + p^{2p}\sum_{i,j=1}^N \mathbb{E}||h(X_i,Y_j)||^p \right).
\end{aligned}$$

We would like to turn the outer sums in i into the maximum over i and the sum in i,j into the maximum over i,j. To achieve this we will use Facts 3 and 4. Before we continue let us note that since for any fixed α we have $1+p^\alpha \leq K^p$, in the sequel we will ommit the multiplicative constant in front of the right-hand side of the inequality in Fact 4 and write just K^p instead.

Let us start with the first term. Applying Fact 4 with $\alpha=1$ and $p/2$ instead of p yields

$$\begin{aligned}
p^p\sum_{i=1}^N \mathbb{E}_X(\mathbb{E}_Y||\sum_{j=1}^N h(X_i,Y_j)||)^p \quad &\leq \quad K^p p^{p/2}\left(p^{p/2}\mathbb{E}_X \max_{1\leq i\leq N}(\mathbb{E}_Y||\sum_{j=1}^N h(X_i,Y_j)||)^p \right.\\
&+ \quad \left. (\sum_{i=1}^N \mathbb{E}||\sum_{j=1}^N h(X_i,Y_j)||^2)^{p/2} \right), && (3.13)
\end{aligned}$$

where to get the last term at the right-hand side we have used the Jensen inequality. This term coincides with the first term at the right-hand side of (3.11).

Now we are going to proceed with the second term at the right-hand side of 3.12. We apply

Fact 4 again, this time with $p/2$ and $\alpha = 3$ to obtain

$$\begin{aligned} p^{3p/2}\mathbb{E}_X \sum_{i=1}^N (\sum_{j=1}^N \mathbb{E}_Y ||h(X_i,Y_j)||^2)^{p/2} &\leq K^p \left(p^{3p/2}\mathbb{E}_X \max_{1\leq i\leq N} (\sum_{j=1}^N \mathbb{E}_Y ||h(X_i,Y_j)||^2)^{p/2} \right. \\ &+ \left. (\sum_{i,j=1}^N \mathbb{E}||h(X_i,Y_j)||^2)^{p/2} \right). \qquad (3.14) \end{aligned}$$

We can see that the first term at the right-hand side has already appeared with the same order of the multiplicative constant in (3.11).

What remains is the last term at the right-hand side of (3.12). We use Fact 4 with $\alpha = 2$ and $p/2$, conditionally to X and obtain

$$\begin{aligned} p^{2p} \sum_{i,j=1}^N \mathbb{E}||h(X_i,Y_j)||^p &\leq K^p \left(p^{2p} \sum_{i=1}^N \mathbb{E} \max_{1\leq j\leq N} ||h(X_i,Y_j)||^p \right. \qquad (3.15) \\ &+ \left. \mathbb{E}_X \sum_{i=1}^N (\sum_{j=1}^N \mathbb{E}_Y ||h(X_i,Y_j)||^2)^{p/2} \right). \end{aligned}$$

To get rid of the second term we use Fact 4 again, this time with $p/2$ and $\alpha = 0$ to get

$$\mathbb{E}_X \sum_{i=1}^N (\sum_{j=1}^N |h(X_i,Y_j)||^2)^{p/2} \leq K^p \left(\mathbb{E}_X \max_{1\leq i\leq N} (\sum_{j=1}^N \mathbb{E}_Y h(X_i,Y_j)||^2)^{p/2} + (\sum_{i,j=1}^N ||h(X_i,Y_j)||^2)^p \right).$$

Since both terms, that we have obtained, have already appeared with greater order of the multiplicative constants in front, we can see that the second term at the right-hand side of (3.15) is negligible.

Thus the last thing that remains is the first term at the right-hand side of (3.15). To bound it, we apply to $\mathbb{E}_X$ Fact 4 with $p/2$ and $\alpha = 4$. We obtain

$$p^{2p} \sum_{i=1}^N \mathbb{E} \max_{1\leq j\leq N} ||h(X_i,Y_j)||^p \leq K^p (p^{2p}\mathbb{E} \max_{1\leq i,j\leq N} ||h(X_i,Y_j)||^p + \mathbb{E}_X (\mathbb{E}_Y \sum_{i=1}^N \max_{1\leq j\leq N} ||h(X_i,Y_j)||^2)^{p/2}).$$

The second term may be bounded from above by $\mathbb{E}_X(\sum_{i=1}^N(\mathbb{E}_Y \sum_{j=1}^N ||h(X_i,Y_j)||^2))^{p/2}$. Thus applying Fact 3 to E_X we can see that it is dominated by

$$K^p \left(p^{p/2}\mathbb{E}_X \sum_{i=1}^N (\mathbb{E}_Y \sum_{j=1}^N ||h(X_i,Y_j)||^2)^{p/2} + (\sum_{i,j=1}^N \mathbb{E}||h(X_i,Y_j)||^2)^{p/2} \right).$$

The first term has already appeared above at the right-hand side of (3.12) and has been bounded in (3.14). Thus we can collect all the terms and using (3.12) and (3.11) obtain

Theorem 14 *There exists a universal constant K, such that for all $p \in \mathbb{N}$, $p > 2$, we have*

$$\begin{aligned}
\mathbb{E}(Z - \mathbb{E}Z)_+^p \quad &\leq \quad K^p \Bigg((\sum_{i,j=1}^{N} \mathbb{E}||h(X_i, Y_j)||^2)^{p/2} \\
&+ \quad p^{p/2}(\mathbb{E}\sum_{i=1}^{N} ||\sum_{j=1}^{N} h(X_i, Y_j)||^2)^{p/2} + p^{p/2}(\mathbb{E}\sum_{j=1}^{N} ||\sum_{i=1}^{N} h(X_i, Y_j)||^2)^{p/2} \\
&+ \quad p^p(||h||_{L_2 \to L_2}^{(1)})^p + p^p(||h||_{L_2 \to L_2}^{(2)})^p \\
&+ \quad p^p \mathbb{E}_X \max_{1 \leq i \leq N} (\mathbb{E}_Y ||\sum_{j=1}^{N} h(X_i, Y_j)||)^p + p^p \mathbb{E}_Y \max_{1 \leq j \leq N} (\mathbb{E}_X ||\sum_{i=1}^{N} h(X_i, Y_j)||)^p \\
&+ \quad p^{3p/2} \mathbb{E}_X \max_{1 \leq i \leq N} (\mathbb{E}_Y \sum_{j=1}^{N} ||h(X_i, Y_j)||^2)^{p/2} + p^{3p/2} \mathbb{E}_Y \max_{1 \leq j \leq N} (\mathbb{E}_X \sum_{i=1}^{N} ||h(X_i, Y_j)||^2)^{p/2} \\
&+ \quad p^{2p} \mathbb{E} \max_{1 \leq i,j \leq N} ||h(X_i, Y_j)||^p \Bigg) .
\end{aligned} \tag{3.16}$$

We are interested in turning the above moment inequality into a bound on the upper tail of Z. We can do it for bounded kernels. Let us define

$$\begin{aligned}
A^2 &= \sum_{i,j=1}^{N} \mathbb{E}||h(X_i, Y_j)||^2 \\
B^2 &= \mathbb{E}\sum_{i=1}^{N} ||\sum_{j=1}^{N} h(X_i, Y_j)||^2 \\
C &= ||h||_{L_2 \to L_2}^{(1)} + ||h||_{L_2 \to L_2}^{(2)} \\
D^2 &= \max \left\{ \left|\left| \mathbb{E}_Y \sum_{j=1}^{N} ||h(\cdot, Y_j)||^2 \right|\right|_\infty , \left|\left| \mathbb{E}_X \sum_{i=1}^{N} ||h(X_i, \cdot)||^2 \right|\right|_\infty \right\} \\
E &= \max_{1 \leq i,j \leq N} ||h(X_i, Y_j)||_\infty \\
F &= \max \left\{ \left|\left| \mathbb{E}_Y ||\sum_{j=1}^{N} h(\cdot, Y_j)|| \right|\right|_\infty , \left|\left| \mathbb{E}_X ||\sum_{i=1}^{N} h(X_i, \cdot)|| \right|\right|_\infty \right\} .
\end{aligned}$$

Then Theorem 14 implies

$$\mathbb{E}(Z - \mathbb{E}Z)_+^p \leq K^p(p^{p/2}(A+B)^p + p^p(C+F)^p + p^{3p/2}D^p + p^{2p}F^p).$$

This implies the following

Theorem 15 *There exists a universal constant K, such that if h is bounded, then for all $t \geq 0$*

$$\mathbb{P}(S - \mathbb{E}S \geq t) \leq K \exp\left(-\frac{1}{K} \min\left(\frac{t^2}{A^2 + B^2}, \frac{t}{C+F}, \frac{t^{2/3}}{D^{2/3}}, \frac{t^{1/2}}{E^{1/2}} \right) \right) .$$

Let us now comment on the special case, when X_i, Y_j are i.i.d random variables. The main interest in inequalities as the one above is their usefulness in proving limit theorems. For example, the real line version of the above theorem has been used to prove the law of iterated logarithm for U-statistics (see [8]). Therefore we are interested in the order of growth of the coefficients $A, \ldots, F$ with the size of the sample (N). Let us therefore take a closer look at the behaviour of those coefficients. We have

$$\begin{aligned}
A^2 &= N^2\mathbb{E}||h(X,Y)||^2, \\
D^2 &\leq N\max(||\mathbb{E}_X||h(X,\cdot)||^2||_\infty, ||\mathbb{E}_Y||h(\cdot,Y)||^2||_\infty), \\
F &\leq N\max(||\mathbb{E}_X||h(X,\cdot)||||_\infty, ||\mathbb{E}_Y||h(\cdot,Y)||||_\infty), \\
C &\leq N\big(\sup\{\mathbb{E}\langle f(X), h(X,Y)\rangle g(Y) \colon f\colon \Sigma \to B^*, g\colon \Sigma \to \mathbb{R},\ \mathbb{E}||f(X)||^2 \leq 1, \mathbb{E}g(Y)^2 \leq 1\} \\
&\quad + \sup\{\mathbb{E}\langle f(Y), h(X,Y)\rangle g(X) \colon f\colon \Sigma \to B^*, g\colon \Sigma \to \mathbb{R},\ \mathbb{E}||f(Y)||^2 \leq 1, \mathbb{E}g(X)^2 \leq 1\}\big),
\end{aligned}$$

where the last line is an easy consequence of the Cauchy-Schwarz inequality. The coefficient E does not depend on the size of the sample. We have however still to deal with B. Let us notice that whenever B^2 is of order N^2, Theorem 15 shows that the upper deviation of Z from its mean is of order N (i.e. $\mathbb{P}(Z - \mathbb{E}Z \geq tN)$ may be bounded by a function depending only on t and vanishing at infinity).

We would like to emphasize that there exists a class of Banach spaces, for which B^2 is indeed of order N^2 for every h and even more, both B^2 and $\mathbb{E}S^2$ can be bounded from above by KA^2, where K is a constant, depending only on the space B. What we have in mind here is the class of Banach spaces of type 2. Below, we define this class and explain how the tail and moment inequalities for U-statistics can be improved in that case.

Definition 4 *A Banach space B is of type p, if there exists a constant T, such that for every $n \in \mathbb{N}$ and every $x_1, \ldots, x_n \in B$, we have*

$$\left(\mathbb{E}||\sum_{i=1}^n \varepsilon_i x_i||^2\right)^{1/2} \leq T\left(\sum_{i=1}^n ||x_i||^p\right)^{1/p}.$$

Remark It is easy to see that every Hilbert space has type 2. Also the spaces L^q for $q \geq 2$ have type 2. The spaces L^p for $1 \leq p \leq 2$ have type p. The proof can be found for example in ([20]).

Let us now notice that for every Banach space valued independent centered random variables $X_1, \ldots, X_n$, we have

$$\mathbb{E}||\sum_{i=1}^n X_i||^p \leq 2^p\mathbb{E}||\sum_{i=1}^n \varepsilon_i X_i||^p,$$

where $\varepsilon_1, \ldots, \varepsilon_n$ is a sequence of independent Rademacher random variables, independent of $X_1, \ldots, X_n$. The proof is analogous to the real case. (comp. Lemma 7). Thus for spaces of type 2, we get

$$\mathbb{E}||\sum_{i=1}^n X_i||^2 \leq 4||\sum_{i=1}^n \varepsilon_i X_i||^2 \leq 4T^2\sum_{i=1}^n \mathbb{E}||X_i||^2$$

and applying it to random variables $h(X_i, Y_j)$, we get

$$\mathbb{E}S^2 = \mathbb{E}||\sum_{i,j=1}^N h(X_i,Y_j)||^2 \leq 4T^2\sum_{i=1}^N \mathbb{E}||\sum_{j=1}^N h(X_i,Y_j)||^2 \leq 16T^4\sum_{i,j=1}^N \mathbb{E}||h(X_i,Y_j)||^2.$$

Thus indeed, the both $\mathbb{E}S^2$ and B^2 ca be bounded by A^2. Let us also take a look at the coefficient F. We have

$$(\mathbb{E}_Y||\sum_{j=1}^{N} h(X_i,Y_j)||)^p \leq (\mathbb{E}_Y||\sum_{j=1}^{N} h(X_i,Y_j)||^2)^{p/2} \leq (4T^2 \sum_{j=1}^{N} \mathbb{E}_Y||h(X_i,Y_j)||^2)^{p/2}$$

Since this quantity appears at the right hand side of moment inequality (3.16), with a greater order of the multiplicative constant in front, we can skip the term corresponding to F at the right-hand side.

The above remarks, together with the inequality

$$\mathbb{E}Z^p \leq \mathbb{E}((Z-\mathbb{E}Z)_+ + \mathbb{E}Z)^p \leq K^p(\mathbb{E}(Z-\mathbb{E}Z)_+^p + (\mathbb{E}Z)^p)$$

give us the following

Theorem 16 *For every Banach space of type 2, there exist constants K, L, depending only on the constant in the definition of type, such that for all $p \in \mathbb{N}$, $p > 2$, we have*

$$\begin{aligned}
\mathbb{E}Z^p \quad &\leq \quad K^p \left(p^{p/2} (\sum_{i,j=1}^{N} \mathbb{E}||h(X_i,Y_j)||^2)^{p/2} \right. \\
&+ \quad p^p(||h||_{L_2 \to L_2}^{(1)})^p + p^p(||h||_{L_2 \to L_2}^{(2)})^p \\
&+ \quad p^{3p/2}\mathbb{E}_X \max_{1\leq i\leq N} (\mathbb{E}_Y \sum_{j=1}^{N} ||h(X_i,Y_j)||^2)^{p/2} + p^{3p/2}\mathbb{E}_Y \max_{1\leq j\leq N} (\mathbb{E}_X \sum_{i=1}^{N} ||h(X_i,Y_j)||^2)^{p/2} \\
&+ \quad \left. p^{2p}\mathbb{E} \max_{1\leq i,j\leq N} ||h(X_i,Y_j)||^p \right). \qquad (3.17)
\end{aligned}$$

and

$$\mathbb{P}(S \geq t) \leq L \exp\left(-\frac{1}{L} \min\left(\frac{t^2}{A^2}, \frac{t}{C}, \frac{t^{2/3}}{D^{2/3}}, \frac{t^{1/2}}{E^{1/2}} \right) \right). \qquad (3.18)$$

for all $t \geq 0$.

Bibliography

[1] N. Alon, M. Krivelevich, V.H Vu. On the concentration of eigenvalues of random symmetric matrices. To appear in Israel J. Math.

[2] S. Boucheron, O. Bosquet, G. Lugosi, P. Massart. Moment inequalities for functions of independent random variables. Preprint.

[3] S. Boucheron, G. Lugosi, P. Massart. A sharp concentration inequality with applications in random combinatorics and learning. *Random Structures and Algorithms*, 16(2000), 277-292.

[4] S. Boucheron, G. Lugosi, P. Massart. Concentration inequalities using the entropy method. *The Annals of Probability*, to appear.

[5] C. Castaing, M. Valadier, *Convex Analysis and Measurable Multifunctions.* Springer Verlag, Berlin 1977.

[6] D. Chafai. Convexity, entropies and functional inequalities. Preprint.

[7] L. Gross. Logarithmic Sobolev inequalities. *Amer. J. Math.* 97 (1975), 1061-1083.

[8] E. Gine, S. Kwapień, R. Latała, J. Zinn. The LIL for canonical U-statistics of order 2. *The Annals of Probability* 29 (2001), 520-557.

[9] E. Gine, R. Latała, J. Zinn. Exponential and moment inequalities for U-statistics. High Dimensional Probability II, Progress in Probability 47, Birkhauser, Boston 2000, 13-38.

[10] D.L. Hanson, F.T. Wright. A bound on tail probabilities for quadratic forms of independent random variables. *Annals of Mathematical Statistics* 42 (1971), 52-61.

[11] R. Latała. Estimation of moments of sums of independent random variables. *Annals of Probability.* 25 (1997), 1502-1513.

[12] R. Latała. K. Oleszkiewicz. Between Sobolev and Poincare. *Geometric Aspects of Functional Analysis*, Lecture Notes in Mathematics 1745, Springer Verlag, Berlin 2000, 147-168.

[13] M. Ledoux. On Talagrand's deviation inequalities for product measures. *ESAIM: Probability and Statistics*, 1(1996), 63-87, `http://www.emath.fr/ps/`

[14] M. Ledoux. *The concentration of measure phenomenon.* Mathematical Surveys and Monographs 89, American Mathematical Society 2001.

[15] M. Ledoux, M. Talagrand. *Probability in Banach spaces.* Springer-Verlag, New York, 1991.

[16] P. Massart. About the constants in Talagrand's deviation inequalities for empirical processes. *Annals of Probability*, 28(2000), 863-884.

[17] C. McDiarmid. On the method of bounded differences. *Surveys in Combinatorics 1989*, 148-188. Cambridge University Press, Cambridge 1989.

[18] C. McDiarmid. Concentration. In M. Habib, C. McDiarmid, J. Ramirez-Alfonsin and B. Reed, editors, *Probabilistic Methods for Algorithmic Discrete Mathematics*, 195-248, Springer Verlag, New York, 1998.

[19] M. W. Meckes. Concentration of norms and eigenvalues of random matrices. Preprint.

[20] V. Milman, G. Schechtman, *Asymptotic theory of finite-dimensional normed spaces*, Lecture Notes in Mathematics 1200, Springer Verlag, Berlin - New York, 1986.

[21] P.M. Samson. Concentration of measure inequalities for Markov chains and Φ-mixing processes. *Annals of Probability*, 28(2000), 416-461.

[22] M. Talagrand, An isoperimetric theorem on the cube and the Khinchine-Kahane inequalities. *Proc. Amer. Math. Soc.*, 104(1988), 905-909.

[23] M. Talagrand, New concentration inequalities in product spaces, *Inventionnes Math* 126(1996), 505-563.

[24] M. Talagrand, A New Look at Independence, *The Annals of Probability*, 24(1996), 1-34.

Contents