Concentration inequalities using the entropy method

By Stéphane Boucheron [1] Gábor Lugosi [2] Pascal Massart

CNRS-Université Paris-Sud *Pompeu Fabra University* *Université Paris-Sud*

We investigate a new methodology, worked out by Ledoux and Massart, to prove concentration-of-measure inequalities. The method is based on certain modified logarithmic Sobolev inequalities. We provide some very simple and general ready-to-use inequalities. One of these inequalities may be considered as an exponential version of the Efron-Stein inequality. The main purpose of this paper is to point out the simplicity and the generality of the approach. We show how the new method can recover many of Talagrand's revolutionary inequalities and provide new applications in a variety of problems including Rademacher averages, Rademacher chaos, the number of certain small subgraphs in a random graph, and the minimum of the empirical risk in some statistical estimation problems.

1. Introduction Concentration inequalities bound tail probabilities of general functions of independent random variables. Several methods have been known to prove such inequalities, including martingale methods (see the surveys of McDiarmid [26], [27]), information-theoretic methods (see Alhswede, Gács, and Körner [1], Marton [20], [21],[22], Dembo [7], Massart [23] and Rio [28]), Talagrand's induction method [34],[32],[33], and various problem-specific methods, see Janson, Łuczak, and Ruciński [11] for a survey.

A novel way of deriving powerful inequalities, the "entropy method", based on logarithmic Sobolev inequalities, was developed by Ledoux [17],[16], Bobkov and Ledoux [3], Massart [24], Rio [28], and Bousquet [5] for proving sharp concentration bounds for maxima of empirical processes. Recently Boucheron, Lugosi, and Massart [4] pointed out that the methodology may be used effectively outside of the context of empirical process theory as well.

In this paper we follow the line of [4] to develop new easy-to-apply inequalities for general functions of independent random variables. These inequalities may be considered as exponential versions of the well-known Efron-Stein inequality. One of the goals of this paper is to point out that the methodology based on logarithmic Sobolev inequalities provides a transparent and general way of obtaining powerful results in a large variety of applications. The proofs are elementary, and are all based on variations of the same theme. We work out several applications. We show that the new method can recover a version of Talagrand's celebrated "convex-distance" inequality. We give a new and simple proof of a revolutionary and very difficult result of Talagrand [33] for the concentration of the suprema of Rademacher chaos. We also provide new applications concerning the number of certain small subgraphs present in a random graph and the minimum of the empirical risk in nonparametric statistical problems.

AMS 1991 subject classifications. Primary 60E15, 60C05, 28A35; Secondary 05C80

Key words and phrases. Concentration inequalities; Empirical processes; Random graphs

[1]Supported by EU Working Group RAND-APX, binational PICASSO Grant 025 43RM

[2]The work of the second author was supported by DGI grant BMF2000-0807

The paper is organized as follows. In Section 2 the basic logarithmic Sobolev inequality is presented, which is used to derive the main inequalities. The proofs of these inequalities are given in a separate Section 3. In Section 4 we review the relationship of the new results with some of the existing work. In particular, we point out that Talagrand's [32] "convex distance" inequality may be recovered easily from some of the new inequalities. Sections 5, 6, and 7 are devoted to concrete applications. In Section 5 we show how the concentration of Rademacher averages and Rademacher chaos can be derived from the new inequalities in a very simple way. In particular, we offer a new and simple proof for a difficult and important result of Talagrand [33] for the concentration of Rademacher chaos. In Section 6 we study upper tail estimates for the number of occurrences of certain small subgraphs in a random graph and show that the new method often compares favorably with the large variety of other methods surveyed by Janson and Ruciński in [13]. Finally, in Section 7 we derive new sharp concentration inequalities for the minimum of the empirical risk in statistical estimation problems, which may have important consequences in efficient model selection.

2. Main inequalities We begin by introducing some notation that is used throughout the paper. We assume that $X_1, \ldots, X_n$ are independent random variables taking values in a measurable space $\mathcal{X}$. Denote by X_1^n the vector of these n random variables. Let $f : \mathcal{X}^n \to \mathbb{R}$ be some measurable function. We are concerned with concentration of the random variable

$$Z = f(X_1, \ldots, X_n) \ .$$

Throughout, $X_1', \ldots, X_n'$ denote independent copies of $X_1, \ldots, X_n$, and we write

$$Z^{(i)} = f(X_1, \ldots, X_{i-1}, X_i', X_{i+1}, \ldots, X_n) \ .$$

One of the first concentration inequalities was proved by Efron and Stein [8], and further improved by Steele [30]:

Proposition 1. (Efron-Stein inequality.)

$$\mathrm{Var}(Z) \le \frac{1}{2}\mathbb{E}\left[\sum_{i=1}^{n}(Z - Z^{(i)})^2\right] .$$

Note that the inequality becomes an equality if f is the sum of its components. This result provides a simple and often sharp way of bounding the variance of complicated functions of independent random variables. While extremely useful, this result fails to capture the exponential nature of tails, present under many circumstances. Our purpose is to discover simple additional conditions under which the Efron-Stein inequality can be converted into exponential upper bounds for either $\mathbb{P}[Z > \mathbb{E}Z + t]$ or $\mathbb{P}[Z < \mathbb{E}Z - t]$ where $t > 0$.

Define the random variables V_+ and V_- by

$$V_+ = \mathbb{E}\left[\sum_{i=1}^{n}(Z - Z^{(i)})^2 \mathbb{1}_{Z > Z^{(i)}} \Big| X_1^n\right]$$

and

$$V_- = \mathbb{E}\left[\sum_{i=1}^{n}(Z - Z^{(i)})^2 \mathbb{1}_{Z<Z^{(i)}} \middle| X_1^n\right] .$$

Then the Efron-Stein inequality may be re-written as

$$\mathrm{Var}(Z) \le \mathbb{E}[V_+] = \mathbb{E}[V_-] .$$

Our first main result shows that if the random variables V_+ and V_- behave "nicely" in the sense that their moment generating function can be controlled, then, indeed, we may obtain an exponential version of the Efron-Stein inequality:

THEOREM 2. *For all $\theta > 0$ and $\lambda \in (0, 1/\theta)$,*

$$\log \mathbb{E}[\exp(\lambda(Z - \mathbb{E}[Z]))] \le \frac{\lambda\theta}{1-\lambda\theta} \log \mathbb{E}\left[\exp\left(\frac{\lambda V_+}{\theta}\right)\right] . \tag{2.1}$$

On the other hand, we have for all $\theta > 0$ and $\lambda \in (0, 1/\theta)$,

$$\log \mathbb{E}[\exp(-\lambda(Z - \mathbb{E}[Z]))] \le \frac{\lambda\theta}{1-\lambda\theta} \log \mathbb{E}\left[\exp\left(\frac{\lambda V_-}{\theta}\right)\right] . \tag{2.2}$$

The value of θ appearing in the upper bounds is a free parameter that can be optimized. Roughly speaking, if V_+ (or V_-) has small tails, small values of θ may give better results. In the next two corollaries two different choices of θ appear. The simplest, but already quite powerful, corollary is obtained if V_+ (or V_-) is bounded by a constant:

COROLLARY 3. *Assume that there exists a positive constant c such that, almost surely, $V_+ \le c$. Then for all $t > 0$,*

$$\mathbb{P}[Z > \mathbb{E}Z + t] \le e^{-t^2/4c} .$$

Moreover, if $V_- \le c$ almost surely, then for all $t > 0$,

$$\mathbb{P}[Z < \mathbb{E}Z - t] \le e^{-t^2/4c} .$$

Proof. To prove the first inequality, note that (2.1) now implies

$$\log \mathbb{E}[\exp(\lambda(Z - \mathbb{E}[Z]))] \le \frac{\lambda\theta}{1-\lambda\theta}\frac{\lambda c}{\theta} .$$

Thus, letting θ approach zero, we obtain

$$\mathbb{E}[\exp(\lambda(Z - \mathbb{E}[Z]))] \le e^{\lambda^2 c} .$$

Hence, by Markov's inequality, for all $\lambda > 0$ and $t > 0$,

$$\mathbb{P}[Z > \mathbb{E}Z + t] \le e^{\lambda^2 c - \lambda t} .$$

Optimizing the value of λ yields the first inequality. The second inequality is implied by (2.2) similarly. □

In the next corollary of Theorem 2 we describe a situation when even though V_+ (or V_-) cannot be bounded by a constant, it is concentrated around its mean value. We will face such situations frequently below when we describe applications. For brevity we only state the upper-tail version.

COROLLARY 4. *Assume that the random variable V_+ is such that there exists a positive constant a such that for $\lambda \in (0, 1/a)$,*

$$\log \mathbb{E}\left[e^{\lambda(V_+ - \mathbb{E}V_+)}\right] \leq \frac{\lambda^2 a \mathbb{E}[V_+]}{1 - a\lambda} .$$

Then

$$\log \mathbb{E}\left[\exp(\lambda(Z - \mathbb{E}[Z]))\right] \leq \frac{\lambda^2 \mathbb{E}[V_+]}{1 - (a+1)\lambda}$$

and, in particular,

$$\mathbb{P}\left[Z > \mathbb{E}Z + t\right] \leq \exp\left(\frac{-t^2}{4\mathbb{E}[V_+] + 2(a+1)t/3}\right) .$$

Proof. Taking $\theta = 1$ in (2.1) and using the condition on the moment generating function of V_+, we obtain

$$\begin{aligned}\log \mathbb{E}\left[\exp(\lambda(Z - \mathbb{E}[Z]))\right] &\leq \frac{\lambda}{1-\lambda}\left(\lambda \mathbb{E}V_+ + \frac{\lambda^2 a \mathbb{E}V_+}{1 - a\lambda}\right) \\ &= \frac{\lambda^2 \mathbb{E}V_+}{(1-\lambda)(1-a\lambda)} \leq \frac{\lambda^2 \mathbb{E}V_+}{1-(a+1)\lambda} .\end{aligned}$$

The bound on the tail probability is now obtained by straightforward calculations summarized in Lemma 11 in Section 3. □

In applications we often face situations when V_+ or V_- may be bounded as some function of Z itself. In many of these cases Theorem 2 is useless, yet a slight modification of the same proof methodology proves to be useful. The next result considers a frequent situation.

THEOREM 5. *Assume that there exist positive constants a and b such that*

$$V_+ \leq aZ + b . \tag{2.3}$$

Then for $\lambda \in (0, 1/a)$,

$$\log \mathbb{E}[\exp(\lambda(Z - \mathbb{E}[Z]))] \leq \frac{\lambda^2}{1 - a\lambda}(a\mathbb{E}Z + b)$$

and for all $t > 0$,

$$\mathbb{P}\left[Z > \mathbb{E}Z + t\right] \leq \exp\left(\frac{-t^2}{4a\mathbb{E}Z + 4b + 2at}\right) .$$

Bounds for the lower tail $\mathbb{P}\left[Z < \mathbb{E}Z - t\right]$ may be easily derived under much more general conditions on V_- due to a simple association inequality:

THEOREM 6. *Assume that for some nondecreasing function g,*

$$V_- \leq g(Z) \ .$$

Then for all $t > 0$,

$$\mathbb{P}\left[Z < \mathbb{E}Z - t\right] \leq \exp\left(\frac{-t^2}{4\mathbb{E}[g(Z)]}\right) \ .$$

In some situations no suitable bound for V_- is available while V_+ is manageable and furthermore we may have the guarantee that $|Z - Z^{(i)}|$ remains bounded. Without loss of generality, we assume that the bound is 1. In such cases we obtain an exponential lower-tail inequality by controlling V_+ only.

THEOREM 7. *Assume that there exists a nondecreasing function g such that $V_+ \leq g(Z)$ and for any value of X_1^n and $X_i{}'$, $|Z - Z^{(i)}| \leq 1$. Then for all $K > 0$, $\lambda \in [0, 1/K)$*

$$\log \mathbb{E}\big[\exp(-\lambda(Z - \mathbb{E}[Z]))\big] \leq \lambda^2 \frac{\psi(K)}{K^2} \mathbb{E}[g(Z)] \ ,$$

and for all $t > 0$, with $t \leq (e-1)\mathbb{E}[g(Z)]$ we have

$$\mathbb{P}\left[Z < \mathbb{E}Z - t\right] \leq \exp\left(-\frac{t^2}{4(e-1)\mathbb{E}[g(Z)]}\right) \ .$$

Our last general result deals with a situation we have often faced in applications. In these cases V_+ may be bounded by the product of Z and another random variable W with well-behaved moment generating function. The following theorem provides a way to deal with such functionals in an efficient and rather painless way.

THEOREM 8. *Assume that f is nonnegative. Assume that there exists a random variable W, such that:*

$$V_+ \leq WZ \ .$$

Then for all $\theta > 0$ and $\lambda \in (0, 1/\theta)$,

$$\log \mathbb{E}\left[\exp(\lambda(\sqrt{Z} - \mathbb{E}[\sqrt{Z}]))\right] \leq \frac{\lambda\theta}{1 - \lambda\theta} \log \mathbb{E}\left[\exp\left(\frac{\lambda W}{\theta}\right)\right] \ .$$

Bounds for the upper-tail probability of Z may be easily derived using Theorem 8. Indeed, since $\sqrt{\mathbb{E}[Z]} \geq \mathbb{E}[\sqrt{Z}]$, writing $x = \sqrt{\mathbb{E}[Z] + t} - \sqrt{\mathbb{E}[Z]}$, we have, for $\lambda > 0$,

$$\mathbb{P}[Z > \mathbb{E}[Z] + t] \leq \mathbb{P}[\sqrt{Z} > \mathbb{E}[\sqrt{Z}] + x] \leq \mathbb{E}\left[\exp(\lambda(\sqrt{Z} - \mathbb{E}[\sqrt{Z}]))\right] e^{-\lambda x}$$

by Markov's inequality. Some concrete examples are worked out in Section 6.

We close this section by an extension of Theorem 5 which shows that if V_+ is bounded by a sub-quadratic polynomial of Z, then Z is concentrated. Once again, we obtain bounds for the moment generating function of Z^p for some $p < 1$ which may be converted into tail bounds for Z as noted above.

Theorem 9. *Assume that f is nonnegative. Assume that there exist constants $a > 0$ and $\alpha \in (0, 2)$ such that $V_+ \leq aZ^\alpha$. Then for all $\lambda > 0$*

$$\log \mathbb{E}\left[\exp\left(\lambda\left(Z^{(2-\alpha)/2} - \mathbb{E}Z^{(2-\alpha)/2}\right)\right)\right] \leq \lambda^2 a \ .$$

and for all $\lambda \in (0, 1/a)$

$$\log \mathbb{E}\left[\exp\left(\lambda\left(Z^{2-\alpha} - \mathbb{E}Z^{2-\alpha}\right)\right)\right] \leq \frac{\lambda^2 a \mathbb{E}Z^{2-\alpha}}{1 - \alpha\lambda}$$

3. Proofs of the main inequalities In this section we derive the main inequalities of the paper. All of them follow from the next logarithmic Sobolev inequalities, which are straightforward variations of an inequality proposed by Massart [24].

Proposition 10. (Logarithmic Sobolev inequalities.) *For any function $f : \mathcal{X}^n \to \mathbb{R}$, noting $Z = f(X_1 \ldots X_n)$ and for all $\lambda \in \mathbb{R}$,*

$$\lambda \mathbb{E}\left[Ze^{\lambda Z}\right] - \mathbb{E}\left[e^{\lambda Z}\right] \log \mathbb{E}\left[e^{\lambda Z}\right] \leq \sum_{i=1}^n \mathbb{E}\left[e^{\lambda Z}\psi(-\lambda(Z - Z^{(i)}))\mathbb{1}_{Z > Z^{(i)}}\right],$$

$$\lambda \mathbb{E}\left[Ze^{\lambda Z}\right] - \mathbb{E}\left[e^{\lambda Z}\right] \log \mathbb{E}\left[e^{\lambda Z}\right] \leq \sum_{i=1}^n \mathbb{E}\left[e^{\lambda Z}\psi(\lambda(Z^{(i)} - Z))\mathbb{1}_{Z < Z^{(i)}}\right],$$

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

Now we are prepared to derive our main theorems.

Proof of Theorem 2. Before proceeding to the proof, we recall the following decoupling device that is described in [24]. For any W such that $\mathbb{E}[\exp(\lambda W)] < \infty$, and for any $\lambda \in \mathbb{R}$,

$$\frac{\mathbb{E}[\lambda W e^{\lambda Z}]}{\mathbb{E}[e^{\lambda Z}]} \leq \frac{\mathbb{E}[\lambda Z e^{\lambda Z}]}{\mathbb{E}[e^{\lambda Z}]} - \log \mathbb{E}[e^{\lambda Z}] + \log \mathbb{E}[e^{\lambda W}] \ . \tag{3.4}$$

(This inequality may be proved easily by recalling the variational formulation of the Kullback-Leibler divergence between the probability measures Q and P

$$K(Q\|P) = \sup\left[\mathbb{E}_Q[W] - \log \mathbb{E}_P[e^W]\right]$$

where the supremum is taken over all random variables W such that $\mathbb{E}_Q[W] < \infty$ and $\mathbb{E}_P[\exp(W)] < \infty$. Taking $\frac{dQ}{dP} = \frac{e^{\lambda Z}}{\mathbb{E}_P[e^{\lambda Z}]}$ and the fact that $K(Q\|P) \geq 0$ imply (3.4) above.)

We first prove inequality (2.1). Let $\lambda \geq 0$ and introduce $F(\lambda) = \mathbb{E}[\exp(\lambda Z)]$. Observe that for $x > 0$, $\psi(-x) \leq x^2$, and therefore, for any $\lambda > 0$, the first inequality in Proposition 10 implies

$$\begin{aligned}\lambda F'(\lambda) - F(\lambda)\log F(\lambda) &\leq \sum_{i=1}^n \lambda^2 \mathbb{E}[e^{\lambda Z}(Z - Z^{(i)})^2 \mathbb{1}_{Z > Z^{(i)}}] \\ &= \lambda^2 \mathbb{E}[V_+ e^{\lambda Z}]\end{aligned}$$

Define $G(\lambda) = \log \mathbb{E}[\exp(\lambda V_+)]$. Note that $G(0) = 0$ and that G is convex. Moreover, G is differentiable in some neighborhood of 0. Applying (3.4) with $W = V_+/\theta$, and noting that $F'(\lambda) = \mathbb{E}[Z \exp(\lambda Z)]$, we obtain

$$\lambda F'(\lambda) - F(\lambda) \log F(\lambda) \leq \lambda^2 \theta \left(F'(\lambda) + \frac{1}{\lambda} F(\lambda) G(\lambda/\theta) - \frac{1}{\lambda} F(\lambda) \log F(\lambda) \right) .$$

Dividing both sides by $\lambda^2 F(\lambda)$ this may be rearranged as

$$\frac{1}{\lambda} \frac{F'(\lambda)}{F(\lambda)} - \frac{1}{\lambda^2} \log F(\lambda) \leq \frac{\theta G(\lambda/\theta)}{\lambda(1-\lambda\theta)} .$$

Here we observe that the left-hand side is just the derivative of $H(\lambda) = (1/\lambda) \log F(\lambda)$. Thus, the obtained differential inequality may be integrated. Recalling that $H(\lambda) \to \mathbb{E}[Z]$ as $\lambda \to 0$, we obtain

$$H(\lambda) \leq \mathbb{E}Z + \int_0^\lambda \frac{\theta G(s/\theta)}{s(1-s\theta)} ds .$$

In order to simplify the obtained expression, we note that the convexity of G implies that $G(s/\theta)/s(1-s\theta)$ is a nondecreasing function of s and therefore

$$\log F(\lambda) \leq \lambda \mathbb{E}Z + \frac{\lambda\theta G(\lambda/\theta)}{(1-\lambda\theta)} ,$$

proving (2.1).

(2.2) follows similarly by replacing Z by $-Z$ and noting that a bound on V_- for Z is equivalent to a bound on V_+ for $-Z$. $\square$

Before proceeding to the proof of Theorem (5), we state a useful technical lemma which summarizes some of the straightforward computations carried out in [24]. Introduce

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

One can check that

$$h(x) \geq \frac{x^2}{2 + 2x/3} \quad \text{if } x \geq 0 \qquad \text{and} \qquad h(x) \geq \frac{x^2}{2} \quad \text{if } x \leq 0.$$

Then we have the following.

Lemma 11. *Let C and a denote two positive real numbers. Then*

$$\sup_{\lambda \in [0,1/a)} \left(\lambda t - \frac{C\lambda^2}{1-a\lambda} \right) = \frac{2C}{a^2} h\left(\frac{at}{2C}\right) \geq \frac{t^2}{2\left(2C + \frac{at}{3}\right)}$$

and the supremum is attained at

$$\lambda = \frac{1}{a}\left(1 - \left(1 + \frac{at}{C}\right)^{-1/2} \right) .$$

Also,

$$\sup_{\lambda\in[0,\infty)}\left(\lambda t-\frac{C\lambda^2}{1+a\lambda}\right)=\frac{2C}{a^2}h\left(\frac{-at}{2C}\right)\geq\frac{t^2}{4C}$$

if $t < C/a$ and the supremum is attained at

$$\lambda=\frac{1}{a}\left(\left(1-\frac{at}{C}\right)^{-1/2}-1\right) .$$

Proof of Theorem 5. Let $\lambda > 0$. The same way as in the first steps of the proof of Theorem 2, we have

$$\begin{aligned}\lambda\mathbb{E}\left[Ze^{\lambda Z}\right]-\mathbb{E}\left[e^{\lambda Z}\right]\log\mathbb{E}\left[e^{\lambda Z}\right]&\leq\mathbb{E}\left[e^{\lambda Z}V_+\right]\\&\leq\lambda^2\left(a\mathbb{E}\left[Ze^{\lambda Z}\right]+b\mathbb{E}\left[e^{\lambda Z}\right]\right) ,\end{aligned}$$

where at the last step we used assumption (2.3).

Denoting, once again, $F(\lambda)=\mathbb{E}\left[e^{\lambda Z}\right]$, the above inequality becomes

$$\lambda F'(\lambda)-F(\lambda)\log F(\lambda)\leq a\lambda^2F'(\lambda)+b\lambda^2F(\lambda) .$$

After dividing both sides by $\lambda^2F(\lambda)$, once again we see that the left-hand side is just the derivative of $H(\lambda)=\lambda^{-1}\log F(\lambda)$, so we obtain

$$H'(\lambda)\leq a(\log F(\lambda))'+b .$$

Using the fact that $H(0)=F'(0)/F(0)=\mathbb{E}Z$ and $\log F(0)=0$, and integrating the inequality, we obtain

$$H(\lambda)\leq\mathbb{E}Z+a\log F(\lambda)+b\lambda ,$$

or, if $\lambda<1/a$,

$$\log\mathbb{E}[\lambda(Z-\mathbb{E}[Z])]\leq\frac{\lambda^2}{1-a\lambda}(a\mathbb{E}Z+b) ,$$

proving the first inequality. The inequality for the upper tail now follows by Markov's inequality and Lemma 11. □

Proof of Theorem 6. To prove lower-tail inequalities we obtain upper bounds for $F(\lambda)=\mathbb{E}[\exp(\lambda Z)]$ with $\lambda<0$. By the second inequality of Proposition 10,

$$\begin{aligned}&\lambda\mathbb{E}\left[Ze^{\lambda Z}\right]-\mathbb{E}\left[e^{\lambda Z}\right]\log\mathbb{E}\left[e^{\lambda Z}\right]\\&\leq\sum_{i=1}^n\mathbb{E}\left[e^{\lambda Z}\psi(\lambda(Z^{(i)}-Z))\mathbb{1}_{Z<Z^{(i)}}\right]\\&\leq\sum_{i=1}^n\mathbb{E}\left[e^{\lambda Z}\lambda^2(Z^{(i)}-Z)^2\mathbb{1}_{Z<Z^{(i)}}\right]\\&\quad\text{(using }\lambda<0\text{ and that }\psi(-x)\leq x^2\text{ for }x>0)\\&=\lambda^2\mathbb{E}\left[e^{\lambda Z}V_-\right]\\&\leq\lambda^2\mathbb{E}\left[e^{\lambda Z}g(Z)\right] .\end{aligned}$$

Since $g(Z)$ is a nondecreasing and $e^{\lambda Z}$ is a decreasing function of Z, Chebyshev's association inequality (see, e.g., [9]) implies that

$$\mathbb{E}\left[e^{\lambda Z} g(Z)\right] \le \mathbb{E}\left[e^{\lambda Z}\right] \mathbb{E}[g(Z)] \ .$$

Thus, dividing both sides of the obtained inequality by $\lambda^2 F(\lambda)$ and writing $H(\lambda) = (1/\lambda)\log F(\lambda)$, we obtain

$$H'(\lambda) \le \mathbb{E}[g(Z)] \ .$$

integrating the inequality in the interval $[\lambda, 0)$ we obtain

$$F(\lambda) \le \exp(\lambda^2 \mathbb{E}[g(Z)] + \lambda \mathbb{E}[Z]) \ .$$

Markov's inequality and optimizing in λ now implies the theorem. □

Proof of Theorem 7. The key observation is that the function $\psi(x)/x^2$ is increasing if $x > 0$. Choose $K > 0$. Thus, for $\lambda \in (-1/K, 0)$, the first inequality of Proposition 10 implies that

$$\begin{aligned}
\lambda \mathbb{E}\left[Ze^{\lambda Z}\right] - \mathbb{E}\left[e^{\lambda Z}\right] \log \mathbb{E}\left[e^{\lambda Z}\right] &\le \sum_{i=1}^n \mathbb{E}\left[e^{\lambda Z} \psi(-\lambda(Z - Z^{(i)})) \mathbb{1}_{Z > Z^{(i)}}\right] \\
&\le \lambda^2 \frac{\psi(K)}{K^2} \mathbb{E}\left[e^{\lambda Z} \sum_{i=1}^n (Z - Z^{(i)})^2 \mathbb{1}_{Z > Z^{(i)}}\right] \\
&\le \lambda^2 \frac{\psi(K)}{K^2} \mathbb{E}\left[g(Z) e^{\lambda Z}\right] ,
\end{aligned}$$

where at the last step we used the condition on V_+.

Just like as in the proof of Theorem 6, we bound $\mathbb{E}\left[g(Z)e^{\lambda Z}\right]$ by $\mathbb{E}[g(Z)]\mathbb{E}\left[e^{\lambda Z}\right]$. The rest of the proof is identical to that of Theorem 6. □

Proof of Theorem 8. Introduce $Y = \sqrt{Z}$ and $Y^{(i)}$ as $\sqrt{Z^{(i)}}$. Then

$$\begin{aligned}
\mathbb{E}\left[\sum_i (Y - Y^{(i)})^2 \mathbb{1}_{Y > Y^{(i)}} \mid X_1^n\right] &= \mathbb{E}\left[\sum_i (\sqrt{Z} - \sqrt{Z^{(i)}})^2 \mathbb{1}_{Z > Z^{(i)}} \mid X_1^n\right] \\
&\le \mathbb{E}\left[\sum_i \left(\frac{Z - Z^{(i)}}{\sqrt{Z}}\right)^2 \mathbb{1}_{Z > Z^{(i)}} \mid X_1^n\right] \\
&\le \frac{1}{Z} \mathbb{E}\left[\sum_i \left(Z - Z^{(i)}\right)^2 \mathbb{1}_{Z > Z^{(i)}} \mid X_1^n\right] \\
&\le W \ .
\end{aligned}$$

Thus, applying Theorem 2 for Y proves the statement. □

Proof of Theorem 9. For any $p > 0$,

$$\begin{aligned}\mathbb{E}\left[\sum_i \left(Z^p - \left(Z^{(i)}\right)^p\right)^2 \mathbb{1}_{Z > Z^{(i)}} \mid X_1^n\right] &= \mathbb{E}\left[\sum_i \left(\frac{Z}{Z^{1-p}} - \frac{Z^{(i)}}{{Z^{(i)}}^{1-p}}\right)^2 \mathbb{1}_{Z > Z^{(i)}} \mid X_1^n\right] \\ &\leq \frac{1}{Z^{2-2p}} \mathbb{E}\left[\sum_i \left(Z - Z^{(i)}\right)^2 \mathbb{1}_{Z > Z^{(i)}} \mid X_1^n\right] \\ &\leq a Z^{\alpha+2p-2}\end{aligned}$$

where at the last step we used the condition of the theorem. The first inequality is obtained by choosing $p = (2-\alpha)/2$ and invoking Corollary 3. The choice $p = 2-\alpha$ and Corollary 4 give the second inequality. □

4. Corollaries, relation to previous results This section is devoted to surveying some of the existing general concentration inequalities and to pointing out their relationship to the inequalities presented in Section 2.

We first mention a classical inequality whose simplicity and the transparency of its conditions have made it one the most useful concentration result. However, the inequality is too rigid in many situations, as it does not take variance information into account. It was proven explicitly first by McDiarmid [26].

Proposition 12. (Bounded difference inequality.) *Assume that*

$$\sup_{\substack{x_1,\ldots,x_n,\\ x_i' \in \mathcal{X}}} |f(x_1,\ldots,x_n) - f(x_1,\ldots,x_{i-1},x_i',x_{i+1},\ldots,x_n)| \leq c_i \ , \ 1 \leq i \leq n \ .$$

Then for all $t > 0$

$$\begin{aligned} &\mathbb{P}\left[Z \geq \mathbb{E}Z + t\right] \leq e^{-2t^2 / \sum_{i=1}^n c_i^2} \ , \\ \textit{and} \qquad &\mathbb{P}\left[Z \leq \mathbb{E}Z - t\right] \leq e^{-2t^2 / \sum_{i=1}^n c_i^2} \ .\end{aligned}$$

It is immediate to see that Corollary 3 implies (up to constant factors in the exponent) the bounded difference inequality.

In a remarkable series of papers (see [34],[32],[33]), Talagrand developed an induction method to prove powerful concentration results in many cases when the bounded difference inequality fails. Perhaps the most widely used of these is the so-called "convex-distance inequality" which we recall here:

Proposition 13. (Convex distance inequality.) *For any subset $A \subseteq \mathcal{X}^n$ with $\mathbb{P}[X_1^n \in A] \geq 1/2$ and $t > 0$,*

$$\mathbb{P}\left[d_T(X_1^n, A) \geq t\right] \leq 2e^{-t^2/4} \ ,$$

where for any $x_1^n = (x_1,\ldots,x_n) \in \mathcal{X}^n$,

$$d_T(x_1^n, A) = \sup_{\alpha \in [0,\infty)^n : \|\alpha\| = 1} d_\alpha(x_1^n, A)$$

denotes the convex distance of x_1^n from the set A. Here

$$d_\alpha(x_1^n, A) = \inf_{y_1^n \in A} d_\alpha(x_1^n, y_1^n) = \inf_{y_1^n \in A} \sum_{i: x_i \neq y_i} |\alpha_i| \ .$$

Even though at the first sight it is not obvious how Talagrand's result can be used to prove concentration for general functions f of X_1^n, apparently with relatively little work, the theorem may be converted into very useful inequalities. Talagrand [32] and Steele [31] survey a large variety of applications. Here we show that Corollary 3 and Theorem 7 imply the convex distance inequality, though with a worse constant in the exponent.

Define the random variable $Z = d_T(X_1^n, A)$. It is known [32] that $d_T(\cdot,\cdot)$ can be represented as a saddle point. Let $\mathcal{M}(A)$ denote the set of probabilities on A. Then

$$\begin{aligned} d_T(X_1^n, A) &= \inf_{\nu \in \mathcal{M}(A)} \sup_{\alpha: \|\alpha\|_2 \le 1} \sum_j \alpha_j \mathbb{E}_\nu [\mathbb{1}_{X_j \neq Y_j}] \\ &= \sup_{\alpha: \|\alpha\|_2 \le 1} \inf_{\nu \in \mathcal{M}(A)} \sum_j \alpha_j \mathbb{E}_\nu [\mathbb{1}_{X_j \neq Y_j}] \, , \end{aligned}$$

where the saddle point is achieved. This follows from Sion's minmax Theorem [29]: let $f(x,y)$ denote a function from $\mathcal{X} \times \mathcal{Y}$ to $\mathbb{R}$ that is convex and lower-semi-continuous with respect to x, concave and upper-semi-continuous with respect to y, if $\mathcal{X}$ is convex and compact, then

$$\inf_x \sup_y f(x,y) = \sup_y \inf_x f(x,y) = \min_x \sup_y f(x,y) \, .$$

The summation is indeed linear with respect to its two arguments. Moreover, let us fix X_1^n for a moment. Rather than minimizing in the large space $\mathcal{M}(A)$, we may as well perform minimization on a convex closed set of probabilities on $\{0,1\}^n$ by mapping $y_1^n \in A$ on $(\mathbb{1}_{y_j \neq X_j})_{1 \le j \le n}$. Let us denote this mapping by χ (note that the mapping depends on X_1^n, we omit this dependence to alleviate notations). The set $\mathcal{M}(A) \circ \chi^{-1}$ of image probability measures on $\{0,1\}^n$ coincides with $\mathcal{M}(\chi(A))$. It is convex and compact. We may rewrite $d_T(X_1^n, A)$ as:

$$\sup_{\alpha: \|\alpha\|_2 \le 1} \inf_{\mu \in \mathcal{M}(A) \circ \chi^{-1}} \sum_j \alpha_j \mathbb{E}_\mu [\omega_j] = \inf_{\mu \in \mathcal{M}(A) \circ \chi^{-1}} \sup_{\alpha: \|\alpha\|_2 \le 1} \sum_j \alpha_j \mathbb{E}_\mu [\omega_j]$$

where ω denotes a generic element of $\{0,1\}^n$. Then the summation is continuous with respect to its arguments, and the extrema are taken in compact spaces. If $(\widehat{\mu}, \widehat{\alpha})$ is a saddle point in $\mathcal{M}(A) \circ \chi^{-1} \times \mathbb{R}^n$, then any $\widehat{\nu} \in \mathcal{M}(A)$ with $\widehat{\nu} \circ \chi^{-1} = \widehat{\mu}$ is such that $(\widehat{\nu}, \widehat{\alpha})$ is a saddle point.

Let $(\widehat{\nu}, \widehat{\alpha})$ be a saddle point for X_1^n. We have

$$Z^{(i)} = \inf_{\nu \in \mathcal{M}(A)} \sup_\alpha \sum_j \alpha_j \mathbb{E}_\nu [\mathbb{1}_{X_j^{(i)} \neq Y_j}] \ge \inf_{\nu \mathcal{M}(A)} \sum_j \widehat{\alpha}_j \mathbb{E}_\nu [\mathbb{1}_{X_j^{(i)} \neq Y_j}] \, .$$

Let $\tilde{\nu}$ denote the distribution on A that achieves the infimum in the latter expression. Now we have

$$Z = \inf_\nu \sum_j \widehat{\alpha}_j \mathbb{E}_\nu [\mathbb{1}_{X_j \neq Y_j}] \le \sum_j \widehat{\alpha}_j \mathbb{E}_{\tilde{\nu}} [\mathbb{1}_{X_j \neq Y_j}] \, .$$

Hence we get

$$Z - Z^{(i)} \leq \sum_j \widehat{\alpha}_j \mathbb{E}_{\tilde{\nu}}[\mathbb{1}_{X_j \neq Y_j} - \mathbb{1}_{X_j^{(i)} \neq Y_j}] = \widehat{\alpha}_i \mathbb{E}_{\tilde{\nu}}[\mathbb{1}_{X_i \neq Y_i} - \mathbb{1}_{X_i^{(i)} \neq Y_i}] \leq \widehat{\alpha}_i \ .$$

Therefore $V_+ \leq \sum_i \widehat{\alpha}_i^2 = 1$. Therefore, by Corollary 3, for any $t > 0$,

$$\mathbb{P}\big[d_T(X_1^n, A) - \mathbb{E}d_T(X_1^n, A) \geq t\big] \leq e^{-t^2/4}.$$

Note that by Efron-Stein inequality $\mathrm{Var}[d_T(X_1^n, A)] \leq \mathbb{E}[V_+] \leq 1$. Writing $P(A) = \mathbb{P}[X_1^n \in A]$, as by Chebyshev inequality:

$$\mathbb{P}[d_T(X_1^n, A) - \mathbb{E}d_T(X_1^n, A) \leq -t] \leq \frac{\mathrm{Var}[d_T(X_1^n, A)]}{t^2} \leq \frac{1}{t^2} \ ,$$

we get

$$\mathbb{E}d_T(X_1^n, A) \leq \frac{1}{\sqrt{P(A)}} \ ,$$

which is less than $\sqrt{2}$ if $P(A) \geq 1/2$. Let $a = \sqrt{2}$, plugging this into the upper-tail inequality for $d_T(X_1^n, A)$ above, we get

$$\mathbb{P}[d_T(X_1^n, A) \geq t + a] \leq e^{-t^2/4}$$

but, for $t > a$, $(t-a)^2 \geq t^2/2 - 4\log 2$ and thus for such t's,

$$\mathbb{P}[d_T(X_1^n, A) \geq t] \leq 2e^{-t^2/8} \ ,$$

which, for $\mathbb{P}[X_1^n \in A] \geq 1/2$, is Talagrand's convex distance inequality except that the constant 4 in the exponent is now replaced by the worse value 8. □

Next we recall one of the corollaries of Proposition 13 which is at the basis of many successful applications of the convex distance inequality.

Corollary 14. (Configuration function bound.) *Assume that* $f : \mathcal{X} \to \mathbb{R}^+$ *is a* configuration function, *that is, for all* $\mathbf{x} = (x_1, \ldots, x_n) \in \mathcal{X}^n$*, there exists a set* $I \in \{1, \ldots, n\}$ *of indices such that* $f(\mathbf{x}) = |I|$ *and for all* $\mathbf{y} = (y_1, \ldots, y_n) \in \mathcal{X}^n$, $f(\mathbf{y}) \geq \sum_{i \in I} \mathbb{1}_{x_i = y_i}$. *Then for all* $t > 0$, *if* $\mathbb{M}Z$ *denotes the median of* Z,

$$\mathbb{P}\left[Z \geq \mathbb{M}Z + t\right] \leq 2\exp\left[-\frac{t^2}{4\mathbb{M}Z + 4t}\right] \ ,$$

$$\text{and} \qquad \mathbb{P}\left[Z \leq \mathbb{M}Z - t\right] \leq 2\exp\left[-\frac{t^2}{4\mathbb{M}Z}\right] \ .$$

The configuration function bound has proved useful in obtaining sharp concentration results for functions such as the length of the longest increasing subsequence in a random permutation, the length of the longest common subsequence of two random strings, the number of occupied bins in "bins-and-balls" occupancy problems, etc.

An improved version of the configuration function bound follows from the following inequality which was proved in [4] using the logarithmic-Sobolev-inequality approach.

PROPOSITION 15. *Assume that f is nonnegative and that there exists a function $g : \mathcal{X}^{n-1} \to R^+$ such that for all $x_1, \ldots, x_n \in \mathcal{X}$*

(1) $0 \le f(x_1, \ldots, x_n) - g(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) \le 1$ *for all* $i = 1, \ldots, n$*;*

(2) $\sum_{i=1}^n [f(x_1, \ldots, x_n) - g(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)] \le f(x_1, \ldots, x_n)$.

Then for any $t > 0$,

$$\mathbb{P}\left[Z \ge \mathbb{E}Z + t\right] \le \exp\left[-\frac{t^2}{2\mathbb{E}Z + 2t/3}\right],$$

$$\textit{and} \qquad \mathbb{P}\left[Z \le \mathbb{E}Z - t\right] \le \exp\left[-\frac{t^2}{2\mathbb{E}Z}\right],$$

It is easy to see that configuration functions satisfy the conditions of Proposition 15, and so the proposition is a generalization (and improvement) of the configuration function bound. It is also shown in [4] that Proposition 15 may be used to prove concentration bounds for other types of functions such as certain combinatorial entropies.

It is also clear that a function f satisfying the conditions of Proposition 15 satisfies the condition of Theorem 5 as well with $b = 0$ and $a = 1$. Thus, Theorem 5 may be considered as a generalization of Proposition 15, though the constants in the bound of Proposition 15 are somewhat better.

In the next sections we present some new applications in which none of the cited methods give satisfactory answers.

5. Rademacher averages and chaos The first applications we present concern Rademacher averages and Rademacher chaos, quantities which play an important role in empirical process theory and in the theory of probability in Banach spaces, see, for example, Ledoux and Talagrand [18] and van der Vaart and Wellner [35]. We start with investigating Rademacher averages of independent Banach-space-valued random variables. We obtain a sharp concentration inequality as a simple application of Proposition 15. Second, we investigate the supremum of Rademacher chaos and provide a simple and transparent proof of Talagrand's concentration inequality published in [33].

5.1. *Rademacher averages* Let B denote a separable Banach space and let $X_1, \ldots, X_n$ be independent and identically distributed bounded B-valued random variables. Without loss of generality we assume that $\|X_1\| \le 1$ almost surely. The quantity of interest is the conditional Rademacher average

$$Z = \mathbb{E}\left[\left\|\sum_{i=1}^n \epsilon_i X_i\right\| \mid X_1^n\right]$$

where the ϵ_i are independent centered $\{1, -1\}$-valued random variables. We offer the following concentration inequalities for Z:

Theorem 16. *For any $t > 0$,*

$$\mathbb{P}\left[Z \geq \mathbb{E}Z + t\right] \leq \exp\left[-\frac{t^2}{2\mathbb{E}Z + 2t/3}\right] ,$$

and
$$\mathbb{P}\left[Z \leq \mathbb{E}Z - t\right] \leq \exp\left[-\frac{t^2}{2\mathbb{E}Z}\right] ,$$

Proof. It suffices to prove that the function $f : B^n \to \mathbb{R}^+$ defined by $f(x_1^n) = \mathbb{E}[\|\sum_{i=1}^n \epsilon_i x_i\|]$ satisfies the conditions of Proposition 15. Introduce

$$Z_i = g(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) = \mathbb{E}\left[\left\|\sum_{\substack{j=1\\ j\neq i}}^n \epsilon_j X_j\right\| \mid X_1^n\right] .$$

Clearly, by monotonicity, $Z \geq Z_i$, and since the X_j are bounded, $Z - Z_i \leq 1$.

Let D denote a dense countable collection in the unit ball of the dual B^* of B. Then, for each choice of the ϵ_j, the Hahn-Banach theorem implies that there exists some element $v_\epsilon \in D$ such that

$$\left\|\sum_{i=1}^n \epsilon_i X_i\right\| = \langle v_\epsilon, \sum_{j=1}^n \epsilon_j X_j\rangle .$$

Then for the same realization of the Rademacher variables

$$\left\|\sum_{\substack{j=1\\ j\neq i}}^n \epsilon_j X_j\right\| \geq \langle v_\epsilon, \sum_{\substack{j=1\\ j\neq i}}^n \epsilon_j X_j\rangle .$$

Hence, for any given realization of the Rademacher variables.

$$\begin{aligned}\sum_{i=1}^n (Z - Z_i) &= \sum_{i=1}^n \left[\left\|\sum_{i=1}^n \epsilon_i X_i\right\| - \left\|\sum_{\substack{j=1\\ j\neq i}}^n \epsilon_j X_j\right\|\right] \\ &\leq \sum_{i=1}^n \langle v_\epsilon, \epsilon_i X_i\rangle \\ &= \left\|\sum_{i=1}^n \epsilon_i X_i\right\| = Z ,\end{aligned}$$

concluding the proof. □

Deriving (upper and lower) tail inequalities for conditional Rademacher averages using Proposition 15 was previously carried out using the "control by q-points" concentration inequality derived by Talagrand [18, 35] or the bounded difference inequality. The first method did not seem to provide bounds for lower tails and the second method provided a conservative inequality and did not seem to be able to capture the fact that the conditional Rademacher averages may be much smaller than $\sqrt{n}$. This point may be important in statistical applications.

5.2. *Rademacher chaos* In this section $\mathcal{F}$ denotes a collection of $n \times n$ symmetric matrices M, and $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. Rademacher variables. We assume that if $M \in \mathcal{F}$, then $-M \in \mathcal{F}$. To avoid problems with measurability we assume that $\mathcal{F}$ is a finite set. For convenience assume that the matrices M have zero diagonal, that is, $M(i,i) = 0$ for all $M \in \mathcal{F}$ and $i = 1, \ldots, n$. We investigate concentration of the random variable

$$Z = \sup_{M \in \mathcal{F}} \sum_{i,j \leq n} \epsilon_i \epsilon_j M(i,j) .$$

Suppose the supremum of the L_2 operator norm of matrices $(M)_{M \in \mathcal{F}}$ is finite, and without loss of generality we assume that this supremum equals one, that is,

$$\sup_{M \in \mathcal{F}} \sup_{\alpha : \sum_{i=1}^n \alpha_i^2 \leq 1} \alpha^\dagger M \alpha = 1$$

where $\alpha^\dagger$ denotes the transpose of the vector $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n$.

The next result is now obtained easily as a consequence of Corollary 4 and Theorem 5. A similar inequality appears in the revolutionary paper of Talagrand [33]. We believe that the proof presented here is more transparent than Talagrand's. Note also that Ledoux [17] already used the logarithmic-Sobolev-inequality approach to prove a version of Talagrand's theorem, but first Ledoux's constants are significantly worse than the ones obtained here, and second the tail bound does not have exactly the same shape.

Theorem 17. *For all $t > 0$,*

$$\mathbb{P}\{Z \geq \mathbb{E}[Z] + t\} \leq \exp\left(-\frac{t^2}{32\mathbb{E}[Y^2] + 65t/3}\right)$$

where the random variable Y is defined as

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

Proof. In order to apply Corollary 4, we need to obtain a suitable upper bound for V_+. This will be achieved by appealing to Theorem 5. If M^* denotes an element of $\mathcal{F}$ such that $Z = \sum_{i,j: i \neq j} \epsilon_i \epsilon_j M^*(i,j)$ then by a straightforward argument we see that for each $k = 1, \ldots, n$,

$$\mathbb{E}\left[(Z - Z^{(k)})^2 \mathbb{1}_{Z > Z^{(k)}} \middle| \epsilon_1^n \right] \leq 8 \left(\sum_{i=1}^n \epsilon_i M^*(i,k) \right)^2$$

and therefore

$$\begin{aligned} V_+ &= \mathbb{E}\left[\sum_{k=1}^n (Z - Z^{(k)})^2 \mathbb{1}_{Z > Z^{(k)}} \middle| \epsilon_1^n \right] \\ &\leq 8 \sup_{M \in \mathcal{F}} \sum_{k=1}^n \left(\sum_{i=1}^n \epsilon_i M(i,k) \right)^2 = 8Y^2. \end{aligned}$$

We may now apply Corollary 4 if we can obtain a suitable bound for the moment generating function of $8Y^2$. We do this by applying the same method to the variable Y^2. First, using the Cauchy-Schwarz inequality, note that

$$Y^2 = \sup_{M \in \mathcal{F}} \sup_{\alpha \in \mathbb{R}^n : \|\alpha\|_2 \le 1} \left(\sum_{i=1}^n \epsilon_i \sum_{k=1}^n \alpha_k M(i,k) \right)^2$$

Introducing the notation $b_i = b_i(M, \alpha) = \sum_{k=1}^n \alpha_k M(i,k)$ we may rewrite the expression of Y^2 as

$$Y^2 = \sup_b \left(\sum_{i=1}^n \epsilon_i b_i \right)^2$$

where the supremum is taken over all vectors $b = b(M,\alpha) = (b_1(M,\alpha), \ldots, b_n(M,\alpha))$ with $M \in \mathcal{F}$ and $\alpha : \|\alpha\|_2 \le 1$. Let $Y^{(i)}$ be defined as

$$\sup_b \left(\sum_{j=1}^n \epsilon_j^{(i)} b_j \right)^2 .$$

Clearly, for all $i \le n$,

$$(Y^2 - {Y^{(i)}}^2) \mathbb{1}_{Y^2 > {Y^{(i)}}^2} \le 2Y(Y - Y^{(i)}) .$$

Hence, denoting by b^* the vector that achieves $Y^2 = \left(\sum_{i=1}^n \epsilon_i b_i^* \right)^2$,

$$\begin{aligned} \mathbb{E}\big[(Y^2 - {Y^{(i)}}^2)^2 \mathbb{1}_{Y^2 > {Y^{(i)}}^2} \mid \epsilon_1^n \big] &\le 4Y^2 \mathbb{E}\big[(Y - Y^{(i)})^2 \mathbb{1}_{Y > Y^{(i)}} \mid \epsilon_1^n \big] \\ &\le 4Y^2 (b_i^*)^2 \mathbb{E}\big[(\epsilon_i - {\epsilon_i}')^2 \mid \epsilon_1^n \big] \\ &\le 8Y^2 (b_i^*)^2 . \end{aligned}$$

From this it follows that

$$\mathbb{E}\left[\sum_{i=1}^n \left(Y^2 - {Y^{(i)}}^2 \right)^2 \mathbb{1}_{Y^2 > {Y^{(i)}}^2} | \epsilon_1^n \right] \le 8Y^2 \sup_b \sum_{i=1}^n b_i^2 .$$

But

$$\sup_b \sum_{i=1}^n b_i^2 = \sup_{M \in \mathcal{F}} \sup_{\alpha \in \mathbb{R}^n : \|\alpha\|_2 \le 1} \sum_{i=1}^n \left(\sum_{k=1}^n \alpha_k M(i,k) \right)^2$$

is just the maximal norm of the matrices M which we assumed to be equal to one. Therefore,

$$\mathbb{E}\left[\sum_{i=1}^n \left(Y^2 - {Y^{(i)}}^2 \right)^2 \mathbb{1}_{Y^2 > {Y^{(i)}}^2} | \epsilon_1^n \right] \le 8Y^2$$

hence Y^2 satisfies the conditions of Theorem 5 with $a = 8$ and $b = 0$. Thus, we obtain

$$\log \mathbb{E} \left[\exp \left(\lambda (Y^2 - \mathbb{E}[Y^2]) \right) \right] \le \frac{\lambda^2}{1 - 8\lambda} 8\mathbb{E}[Y^2] .$$

This bound is exactly of the form required by Corollary 4 which, in turn, implies the stated inequality. □

Next we study lower-tail inequalities for Z. In this case we have been unable to recover Talagrand's bound which states a bound of the same form as that for the upper tail. Here we summarize an alternative bound which is an easy consequence of Theorem 7.

Let U' denote

$$U' = \sup_{M\in\mathcal{F}} \max_i \sum_j |M(i,j)| = \sup_{M\in\mathcal{F}} \sup_{\epsilon_1^n} \max_i \sum_j \epsilon_j M(i,j) \ .$$

Thus, U' is the supremum of the norms of the operators $\ell_\infty^n \to \ell_\infty^n$ defined by the matrices M. Note that as the supremum of the L_2 operator norm of matrices $(M)_{M\in\mathcal{F}}$ is at most U', $U' \geq 1$.

In order to apply Theorem 7, we need first an upper bound on $Z - Z^{(i)}$ when $Z > Z^{(i)}$. Let again M^* denote an element of $\mathcal{F}$ such that $\sum_{i,j}\epsilon_i\epsilon_j M^*(i,j) = \sup_{M\in\mathcal{F}}\sum_{i,j}\epsilon_i\epsilon_j M(i,j)$. Then

$$\begin{aligned}
\left(Z - Z^{(i)}\right)\mathbb{1}_{Z>Z^{(i)}} &\leq 4\left|\sum_j \epsilon_j M^*(i,j)\right| \\
&\leq 4 \sup_{M\in\mathcal{F}} \sup_{\epsilon_1^n} \max_i \sum_j \epsilon_j M(i,j) \\
&\leq 4U' \ .
\end{aligned}$$

Thus for $\lambda \in [0, 1/(4U'))$, using the same notation as above for Y,

$$\begin{aligned}
\sum_i \mathbb{E}\left[e^{-\lambda Z}\psi\left(\lambda(Z - Z^{(i)})\right)\mathbb{1}_{Z>Z^{(i)}}\right] &\leq \sum_i \lambda^2\psi(1)\mathbb{E}\left[e^{-\lambda Z}\left(Z - Z^{(i)}\right)^2\mathbb{1}_{Z>Z^{(i)}}\right] \\
&\leq \lambda^2\psi(1)8\mathbb{E}\left[e^{-\lambda Z}Y^2\right] \ .
\end{aligned}$$

Hence for $\lambda \in [0, 1/(4U'))$,

$$\log \mathbb{E}\left[e^{-\lambda(Z-\mathbb{E}[Z])}\right] \leq \frac{\lambda^2 8\psi(1)\mathbb{E}[Y^2]}{1 - (8\psi(1))\lambda} \ .$$

Using Lemma 11, this leads to the following lower tail bounds. If either $U' \leq 2\psi(1)$ or $U' \geq 2\psi(1)$ and $t \leq \mathbb{E}[Y^2]\left((1 - 2\psi(1)/U')^{-2} - 1\right)$, then

$$\mathbb{P}\left\{Z \leq \mathbb{E}[Z] - t\right\} \leq \exp\left(\frac{-t^2}{16\psi(1)\left(2\mathbb{E}[Y^2] + t/3\right)}\right).$$

On the other hand, if $U' \geq 2\psi(1)$ and $t \geq \mathbb{E}[Y^2]\left((1 - 2\psi(1)/U')^{-2} - 1\right)$, then

$$\mathbb{P}\left\{Z \leq \mathbb{E}[Z] - t\right\} \leq \exp\left(\frac{-t}{8U'}\frac{U' - 2\psi(1)}{U' - \psi(1)}\right) \ .$$

□

These bounds should be compared with Theorem 1.2 in [33]. According to this theorem, there exists a universal constant K such that

$$\mathbb{P}\left\{|Z - \mathbb{M}[Z]| > t\right\} \leq 2\exp\left(-\frac{1}{K}\min\left(\frac{t^2}{\mathbb{E}^2[Y]}, t\right)\right) .$$

As Y is concentrated around its mean, $\mathbb{E}[Y^2]$ and $\mathbb{E}^2[Y]$ scale in the same way, and deviations from above in Theorem 17 are controlled essentially in the same way as in [33]. When U' is of the same order of magnitude as U this is also true for deviations from below.

Theorem 17 should also be compared with Theorem 3.1 in Ledoux [17]. Ledoux's tail bound has the form

$$\mathbb{P}\left\{Z + \mathbb{E}[Z] + t\right\} \leq 2\exp\left(-\frac{1}{K}\min\left(t, \frac{t^2}{\mathbb{E}[Z] + \mathbb{E}^2[Y]}\right)\right) .$$

The term $\mathbb{E}[Z]$ in the denominator of the exponent appears because of the way $\mathbb{E}[Y^2\exp(\lambda Z)]$ is dealt with. Instead of using the decoupling device (3.4), Ledoux uses a truncation argument and integration by parts.

For more information on related results on Rademacher chaos we refer the reader to Ledoux and Talagrand [18] and Giné and de la Peña [6].

6. Counting small subgraphs in random graphs Recently an important body of work has been dedicated to the concentration of the number of occurrences of certain small subgraphs in a random graph, see Kim and Vu [14], Vu [36],[37], Janson and Ruciński [12], [13] and the references therein. The purpose of this section is to point out that the inequalities presented in this paper provide a convenient and often sharp tool for deriving such results. In particular, we study two simple examples. First, we consider the number of triangles in a random graph, and show how the best known inequalities can be recovered by the new methodology. Second, new concentration inequalities are derived for the number of cycles of length four which improve the best known results in a certain range of the parameters.

Throughout the section, we consider the Erdős-Rényi $G(n,p)$ model of a random graph. Such a graph has n vertices and for each pair (u,v) of vertices an edge is inserted between u and v with probability p, independently. We write $m = \binom{n}{2}$, and denote the indicator variables of the m edges by $X_1, \ldots, X_m$.

Remark. Janson's inequality. For problems of the type studied in this section, Janson's inequality [10] provides a sharp tool to obtain upper bounds for the lower tail. More precisely, let $\mathcal{I}$ denote a family of subsets of the edges and define the number of occurrences of elements of $\mathcal{I}$ in the random graph by

$$Z = \sum_{\alpha\in\mathcal{I}}\prod_{i\in\alpha} X_i .$$

One typical example is when $\mathcal{I}$ contains all triples of edges which form a triangle. Write

$$I_\alpha \triangleq \prod_{i\in\alpha} X_i.$$

Then expectation and variance of Z can be computed easily:

$$\mathbb{E}[Z] = \sum_{\alpha\in\mathcal{I}} \mathbb{E}[I_\alpha]$$

$$\mathrm{Var}[Z] = \sum_{\alpha} (\mathbb{E}[I_\alpha] - \mathbb{E}[I_\alpha]^2) + \sum_{\alpha,\beta:\alpha\cap\beta\neq\emptyset} [\mathbb{E}[I_\alpha I_\beta] - \mathbb{E}[I_\alpha]\mathbb{E}[I_\beta]] \,.$$

Define

$$\delta = \frac{\sum_{\alpha,\beta:\alpha\cap\beta\neq\emptyset} \mathbb{E}[I_\alpha I_\beta]}{\mathbb{E}[Z]}.$$

Note that

$$\mathrm{Var}[Z] \le \sum_{\alpha} \mathbb{E}[I_\alpha] + \sum_{\alpha,\beta:\alpha\cap\beta\neq\emptyset} \mathbb{E}[I_\alpha I_\beta] = (1+\delta)\mathbb{E}[Z] \ .$$

On the other hand, let k be the maximum cardinality of the sets in $\mathcal{I}$. Then the Efron-Stein estimate of the variance becomes

$$\begin{aligned}
\mathbb{E}[V_-] &= \sum_i \mathbb{E}\left[\left(Z - Z^{(i)}\right)^2 \mathbb{1}_{Z<Z^{(i)}}\right] \\
&\le \sum_i \mathbb{E}\left[\sum_{\alpha:i\in\alpha} (I_\alpha)^2\right] \\
&= \sum_i \mathbb{E}\left[\sum_{\alpha:\in i\in\alpha} I_\alpha + \sum_{\alpha\neq\beta:i\in\alpha\cap\beta} I_\alpha I_\beta\right] \\
&= \sum_{\alpha} \sum_{i\in\alpha} \mathbb{E}[I_\alpha] + \sum_{\alpha\neq\beta} \sum_{i\in\alpha\cap\beta} \mathbb{E}[I_\alpha I_\beta] \\
&\le k(1+\delta)\mathbb{E}[Z]
\end{aligned}$$

which is k times larger than the direct bound on the variance.

Janson's inequality states that for all $t \in [0, \mathbb{E}Z]$,

$$\mathbb{P}[Z < \mathbb{E}Z - t] \le \exp\left(\frac{-t^2}{2(1+\delta)\mathbb{E}[Z]}\right) \ .$$

This nice and elegant result is basically unimprovable. It is quite straightforward to derive a similar inequality using the methods of this paper. More precisely (the details are omitted), one obtains, using Theorem 6,

$$\mathbb{P}[Z < \mathbb{E}Z - t] \le \exp\left(\frac{-t^2}{4k(1+\delta)\mathbb{E}[Z]}\right) \ .$$

In typical cases k is a small constant (for example, $k = 3$ in the case when $\mathcal{I}$ is the class of all triangles) and we come within a constant factor of Janson's inequality, though we do not quite achieve it. The reason is that, as we have seen it above, the Efron-Stein estimate is at least k times larger than the true value of the variance. This explains the appearance of the extra factor of k in the exponential version of the Efron-Stein inequality. $\square$

In contrast to lower-tail bounds, obtaining sharp bounds for $\mathbb{P}[Z > \mathbb{E}Z + t]$ remains a challenge. Janson and Ruciński [12] give an excellent overview of the different attempts that have been made to derive bounds for the upper tail. In the next two subsections our aim is to demonstrate that the method of this paper may be a serious contender to get sharp results.

6.1. *Counting triangles* In this section we consider the number of triangles in a random graph. A triangle is a set of three edges defined by vertices u, v, w such that the edges are of the form $\{u, v\}, \{v, w\}$ and $\{w, u\}$. Let Z denote the number of triangles in a random graph. Note that

$$\mathbb{E}[Z] = \frac{n(n-1)(n-2)}{6} p^3 \approx \frac{n^3 p^3}{6}$$

and

$$\mathrm{Var}(Z) = \binom{n}{3}(p^3 - p^6) + \binom{n}{4}\binom{4}{2}(p^5 - p^6) .$$

We offer the following exponential inequality for the upper tail probabilities of Z, which matches the best results announced in [13].

Theorem 18. *Let $K > 1$. Then for all $0 \le t \le (K^2 - 1)\mathbb{E}Z$,*

$$\begin{aligned} &\mathbb{P}\{Z > \mathbb{E}[Z] + t\} \\ &\le \exp\left(- \frac{t^2}{(K+1)^2 \mathbb{E}[Z] \left(24np^2 + 24 \log n + \frac{14t}{(K+1)\sqrt{\mathbb{E}[Z]}} \right)} \vee \frac{t^2}{12n\mathbb{E}[Z] + 6nt} \right) . \end{aligned}$$

Proof. We apply Theorem 8. First we derive a suitable bound for V_+. If v and u denote the extremities of edge i $(1 \le i \le m)$, then we denote by B_i the number of vertices w such that edges (u, w) and (v, w) exist in the random graph. Then

$$\begin{aligned} V_+ &= \sum_{i=1}^m \mathbb{E}[(Z - Z^{(i)})^2 \mathbb{1}_{Z > Z^{(i)}} \mid X_1^m] \\ &= \sum_{i=1}^m X_i (1-p) B_i^2 \\ &\le \sum_{i=1}^m X_i B_i^2 . \end{aligned}$$

Since $\sum_{i=1}^m X_i B_i = 3Z$ we have

$$V_+ \le \sum_{i=1}^m X_i \left(\max_{j=1,\ldots,m} B_j \right) B_i \le \left(\max_{j=1,\ldots,m} B_j \right) \sum_{i=1}^m X_i B_i \le 3 \left(\max_{j=1,\ldots,m} B_j \right) Z .$$

Using the trivial upper bound $\max_j B_j \le n$ and Theorem 5, one gets the second upper bound in the theorem.

To obtain the first bound, we use Theorem 8. In order to do so, we need to bound the moment generating function of $W \triangleq 3\max_{j=1,\ldots,m} B_j$. Observe that

$$\sum_i \left(W - W^{(i)}\right)^2 \mathbb{1}_{W>W^{(i)}} \leq 18W \ .$$

Hence, by Theorem 5,

$$\log \mathbb{E}[e^{\lambda(W-\mathbb{E}[W])}] \leq \frac{18\lambda^2\mathbb{E}[W]}{1-18\lambda} \ .$$

Denote $Y = \sqrt{Z}$. Theorem 8 leads to

$$\log \mathbb{E}[\exp\left(\lambda(Y - \mathbb{E}[Y])\right)] \leq \frac{\lambda}{1-\lambda}\left(\frac{18\lambda^2\mathbb{E}[W]}{1-18\lambda} + \lambda\mathbb{E}[W]\right) \leq \frac{\lambda^2\mathbb{E}[W]}{1-19\lambda} \ .$$

This, by Lemma 11, implies

$$\mathbb{P}\left\{Y > \mathbb{E}[Y] + t\right\} \leq \exp\left(-\frac{t^2}{4\mathbb{E}[W] + 14t}\right) \ .$$

$W/3$ is the maximum of $m = \binom{n}{2}$ binomial random variables with parameters (n, p^2). In order to upper bound $\mathbb{E}[W/3]$, it is convenient to use the following device which we learned from G. Pisier. Let S_i with $i \leq m$ denote a sequence of binomially distributed random variables with parameters n and p^2. By Jensen's inequality,

$$\begin{aligned}
\mathbb{E}[W/3] &\leq \log\left(\mathbb{E}\left[\max_{i\leq m} e^{S_i}\right]\right) \\
&\leq \log\left(\mathbb{E}\left[me^{S_1}\right]\right) \\
&= \log m + \log\left(\mathbb{E}\left[e^{S_1}\right]\right) \\
&\leq \log m + (e-1)np^2 \\
&\leq 2\log n + 2np^2 \ .
\end{aligned}$$

Hence, we obtain the following bound for the tail of Y.

$$\mathbb{P}\left\{Y \geq \mathbb{E}[Y] + t\right\} \leq \exp\left(-\frac{t^2}{24(np^2 + \log n) + 14t}\right) \ .$$

Now it is straightforward to get tail bounds for the number Z of triangles. Let $K > 1$ be arbitrary, and assume that $t \le (K^2 - 1)\mathbb{E}[Z]$. Then

$$
\begin{aligned}
&\mathbb{P}\{Z > \mathbb{E}[Z] + t\} \\
&\le \mathbb{P}\left\{Y \ge \sqrt{\mathbb{E}[Y^2]}\sqrt{1 + t/\mathbb{E}[Y^2]}\right\} \\
&\le \mathbb{P}\left\{Y \ge \sqrt{\mathbb{E}[Y^2]}(1 + t/((K+1)\mathbb{E}[Y^2]))\right\} \\
&\quad (\text{since } t \le (K^2 - 1)\mathbb{E}[Z]) \\
&\le \mathbb{P}\left\{Y \ge \mathbb{E}[Y] + t/((K+1)\sqrt{\mathbb{E}[Z]})\right\} \\
&\le \exp\left(-\frac{t^2}{(K+1)^2\mathbb{E}[Z]\left(24np^2 + 24\log n + \frac{14t}{(K+1)\sqrt{\mathbb{E}[Z]}}\right)}\right)
\end{aligned}
$$

as desired. $\square$

To understand the inequality, we summarize some of its consequences for different choices of t and for different ranges of the parameter p. For different ranges of t, we obtain the following bounds:

(a) $0 \le t \le \sqrt{\mathbb{E}[Z]}(np^2 + \log n)$

$$\mathbb{P}\{Z \ge \mathbb{E}[Z] + t\} \le \exp\left(-\frac{t^2}{256\mathbb{E}[Z](np^2 + \log n)}\right) .$$

This is the "Gaussian" range. Note that, up to the $\log n$ term, the denominator coincides with the variance.

(b) $\sqrt{\mathbb{E}[Z]}(np^2 + \log n) \le t \le \mathbb{E}[Z] \wedge n\sqrt{\mathbb{E}[Z]}$

$$\mathbb{P}\{Z \ge \mathbb{E}[Z] + t\} \le \exp\left(-\frac{t}{256\sqrt{\mathbb{E}[Z]}}\right) .$$

(c) $\mathbb{E}[Z] \wedge n\sqrt{\mathbb{E}[Z]} \le t \le n^2 \vee \mathbb{E}[Z]$, then, if $n \ge \sqrt{\mathbb{E}[Z]}$ and $np^2 + \log n < \sqrt{\mathbb{E}[Z]}$:

$$\mathbb{P}\{Z \ge \mathbb{E}[Z] + t\} \le \exp\left(-\frac{\sqrt{t}}{256}\right) ,$$

otherwise

$$\mathbb{P}\{Z \ge \mathbb{E}[Z] + t\} \le \exp\left(-\frac{t^2}{18n\mathbb{E}[Z]}\right) .$$

(d) $n^2 \vee \mathbb{E}[Z] < t$

$$\mathbb{P}\{Z \ge \mathbb{E}[Z] + t\} \le \exp\left(-\frac{t}{18n}\right) .$$

Remark. In [13, Table 2], Janson and Ruciński compare tail bounds obtained by different methods for $t = \mathbb{E}[Z]$. They actually relate the exponents in tail bounds with n and p. Their results are best appreciated by letting n go to infinity while keeping np^α constant for some fixed α. The best relation they get is obtained using some tailored version of the "deletion method". They show that

$$-\max\left(\frac{n}{\mathbb{E}[Z]}, \frac{1}{\sqrt{\mathbb{E}[Z]}}\right) \log \mathbb{P}\{Z \geq 2\mathbb{E}[Z]\}$$

remains bounded away from 0. This also follows from the bounds described above.

6.2. *Counting occurrences of C_4 in random graphs* A cycle C_4 of length 4 is a list of 4 vertices u, v, w, s such that edges $(u,v), (v,w), (w,s)$ and (s,u) exists. Let Z denote now the number of occurrences of C_4 in a random graph and let again $m = \binom{n}{2}$ and $X_i = 1$ if edge i exists in the random graph. If i denotes the pair of vertices (u,v), let

$$B_i \triangleq \sum_{(w,s)} X_{(u,w)} X_{(w,s)} X_{(s,v)},$$

where the sum is over pairs of vertices. This is not a sum of independent random variables as it was the case when dealing with triangles. Note that

$$\mathbb{E}[Z] = \frac{n!}{8(n-4)!} p^4 \approx \frac{n^4 p^4}{8} \ ,$$

and that

$$\sum_i \mathbb{E}[(Z - Z^{(i)})^2] \leq \sum_i 2p(1-p)\mathbb{E}[B_i^2]$$

We first apply Theorem 8 and observe that for some range of p, it provides meaningful results.

In order to be able to apply Theorem 8, note that

$$\begin{aligned} \mathbb{E}\left[\sum_i \left(Z - Z^{(i)}\right)^2 \mathbb{1}_{Z > Z^{(i)}} \mid X_1^m\right] &\leq \sum_i X_i B_i^2 \\ &\leq 4(\max_i B_i) Z \ . \end{aligned}$$

Let now $W = 4 \max_i B_i$.

Observe that for each i, B_i is dominated by a binomial with parameters n^2 and p, and thus, by the union bound,

$$\begin{aligned} \log \mathbb{E}[e^{\lambda W}] &\leq \log \mathbb{E}\left[\sum_i e^{4\lambda B_i}\right] \\ &\leq 2\log n + \log \mathbb{E}[e^{\lambda B_1}] \\ &\leq 4 \log n + n^2 p\big(e^{4\lambda} - 1\big) \ . \end{aligned}$$

Here again, we define $Y = \sqrt{Z}$. Using Theorem 8 and taking $\theta = \lambda$, we get the following upper bound:

$$\log \mathbb{E}\left[e^{\lambda(Y-\mathbb{E}[Y])}\right] \leq \frac{\lambda^2}{1-\lambda} \log \mathbb{E}[e^W] \ .$$

Therefore,

$$\mathbb{P}\left\{Y \geq \mathbb{E}[Y] + t\right\} \leq \exp\left(-\frac{t^2}{2(2\log \mathbb{E}[\exp(W)] + t/3)}\right).$$

This leads, for example, to

$$\mathbb{P}\left\{Z \geq 2\mathbb{E}[Z]\right\} \leq \exp\left(-\frac{n^4p^4}{2(64e^4n^2p + 2n^2p^2)}\right) \leq \exp\left(-\frac{n^2p^3}{128e^8}\right).$$

As of this writing, this is the best available upper bound for $p > n^{-1/3}$. In [13, Table 2], when $p > n^{-1/3}$ while n tends to infinity, the best bounds obtained using the deletion method are of the form $\exp(-Cn^{4/3}p)$, and this bound is weaker than the one presented here since $n^{4/3}p < n^2p^3$ is equivalent to $p > n^{-1/3}$.

7. Minimum of the Empirical Risk Concentration inequalities have been used as a key tool in recent developments of model selection methods in nonparametric classification. In this section we describe an application of logarithmic Sobolev inequalities in this framework, which reveals a new phenomenon providing deeper insight into the model selection problem. In this section we simply present the main result, the reader can consult the background in the recent work of Koltchinskii Panchenko [15], Massart [25], and Bartlett, Boucheron, and Lugosi [2].

Let $\mathcal{F}$ denote a class of $\{0,1\}$-valued functions on some space $\mathcal{X}$. For simplicity of the exposition we assume that $\mathcal{F}$ is finite. The results remain true for general classes as long as the measurability issues are taken care of. Given an i.i.d. sample $D_n = (\langle X_i, Y_i\rangle)_{i\leq n}$ of n pairs of random variables $\langle X_i, Y_i\rangle$ taking values in $\mathcal{X} \times \{0,1\}$, for each $f \in \mathcal{F}$ we define the empirical risk

$$L_n(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(X_i), Y_i)$$

where the loss function ℓ is defined on $\{0,1\}^2$ by

$$\ell(y, y') = |y - y'| \ .$$

In nonparametric classification and learning theory it is common to select an element of $\mathcal{F}$ by minimizing the empirical risk. The quantity of interest in this section is the minimal empirical risk

$$\widehat{L} = \inf_{f\in\mathcal{F}} L_n(f).$$

The bounded difference inequality (Proposition 12) immediately implies that $|\widehat{L} - \mathbb{E}[\widehat{L}]|$ is with overwhelming probability at most of the order of $1/\sqrt{n}$. In this section, we show that $\widehat{L}$ may be much more concentrated than predicted by Proposition 12 if $\mathbb{E}\widehat{L}$ is small and the class $\mathcal{F}$ is not too large. Getting tight results for the fluctuations of $\widehat{L}$ provides better insight into the calibration of penalties in certain model selection methods.

In the sequel W denotes the supremum of the empirical process indexed by $\mathcal{F}$:

$$W = n \sup_{f\in\mathcal{F}} |L_n(f) - \mathbb{E}L_n(f)| \ .$$

Also, introduce

$$v = \mathbb{E}\left[\sup_{f\in\mathcal{F}} \sum_{i=1}^{n} \left(\ell(f(X_i),Y_i) - \ell(f(X_i'),Y_i')\right)^2\right] \ .$$

v plays a key role in Theorem 19 below. It is shown in [24] that

$$v \le 2n\left(\sup_{f\in\mathcal{F}} L(f)(1-L(f))\right) + 16\mathbb{E}[W] \le \frac{n}{2} + 16\mathbb{E}[W] \ .$$

The main result of this section is the following.

THEOREM 19. *Let* $a = \frac{33}{8} + \frac{(e-1)v}{8n\mathbb{E}[\widehat{L}]+4\mathbb{E}[W]}$. *Then for all* $t > 0$

$$\mathbb{P}[\widehat{L} > \mathbb{E}[\widehat{L}] + t] \le \exp\left(-\frac{nt^2}{25\mathbb{E}[\widehat{L}]/2 + 25\mathbb{E}[W]/(4n) + 2at/3}\right)$$

and for $t \in [0, \mathbb{E}[\widehat{L}])$,

$$\mathbb{P}[\widehat{L} < \mathbb{E}[\widehat{L}] - t] \le \exp\left(-\frac{nt^2}{4\mathbb{E}[\widehat{L}] + 4\mathbb{E}[W]/n + \frac{4tv(e-1)}{n\mathbb{E}[\widehat{L}]+\mathbb{E}[W]}}\right) \ .$$

It is well known that if $\mathcal{F}$ is, for example, a Vapnik-Chervonenkis class then $\mathbb{E}[W]/n$ is bounded by $c\sqrt{d/n}$ where d is the VC dimension of $\mathcal{F}$ and c is a constant (see, e.g., [35]). Thus, the main message of Theorem 19 is that typical fluctuations of $\widehat{L}$ are of the order of $\sqrt{\mathbb{E}[\widehat{L}]/n} + d^{1/4}n^{-3/4}$ (at least when $\mathbb{E}[\widehat{L}] \ge n^{-1}$) which may be significantly smaller than the $n^{-1/2}$ suggested by the bounded difference inequality, since in typical applications $\mathbb{E}[\widehat{L}]$ is small.

In the proof we make use of a recent result of Massart [24] for the concentration of W:

LEMMA 20. *Let* $\phi(\lambda) = \exp(\lambda) - \lambda - 1$. *Then for all* $\lambda \in (0,1)$,

$$\log \mathbb{E}[e^{\lambda(W-\mathbb{E}[W])}] \le \frac{\lambda^2 v}{1-\lambda}\left(1 + \frac{\phi(\lambda)}{\lambda}\right) \le \frac{\lambda^2(e-1)v}{1-\lambda}$$

Proof of Theorem 19. Introduce $Z = n\widehat{L}$. The analysis is based on the second inequality of Proposition 10. Let

$$Z^{(i)} = \min_{f\in\mathcal{F}}\left[\sum_{j\ne i} \ell(f(X_j),Y_j) + \ell(f(X_i'),Y_i')\right]$$

where $\langle X_i', Y_i' \rangle$ is independent of D_n and has the same distribution as $\langle X_i, Y_i \rangle$. Let g denote a (possibly non-unique) minimizer of the empirical risk. The key observation is that

$$\begin{aligned}(Z^{(i)} - Z)^2 \mathbb{1}_{Z^{(i)}>Z} &\le (\ell(g(X_i'), Y_i') - \ell(g(X_i), Y_i))^2 \mathbb{1}_{Z^{(i)}>Z} \\ &= \ell(g(X_i'), Y_i') \mathbb{1}_{\ell(g(X_i),Y_i)=0} \ .\end{aligned}$$

Thus,

$$V_- \le \sum_{i:\ell(g(X_i),Y_i)=0} \mathbb{E}_{X_i',Y_i'}[\ell(g(X_i'), Y_i')] \le nL(g) \ .$$

Since

$$nL(g) = Z + (nL(g) - Z) \le Z + n \sup_{f \in \mathcal{F}} (\mathbb{E}L_n(f) - L_n(f)) \le Z + W \ ,$$

we obtain

$$V_- \le Z + W \ . \tag{7.5}$$

This upper bound fits neither entirely in the framework of Corollary 4 nor completely in the framework of Theorem 5, but a simple modification of their proofs yields the desired result. For the sake of completeness we detail the proof in the sequel.

First we derive the upper-tail inequality, proceeding similarly as in the proof of Theorem 7. Assume that $\lambda \in [0, 4/5)$. By the second inequality in Proposition 10, and the fact that $x^{-2}\psi(x)$ is increasing and that $\psi(4/5) < 1$ we obtain

$$\begin{aligned}\lambda\mathbb{E}[Ze^{\lambda Z}] - \mathbb{E}[e^{\lambda Z}] \log \mathbb{E}[e^{\lambda Z}] &\le \frac{25}{16} \sum_{i=1}^n \lambda^2 \mathbb{E}[e^{\lambda Z} (Z^{(i)} - Z)^2 \mathbb{1}_{Z^{(i)}>Z}] \\ &= \frac{25}{16} \lambda^2 \, \mathbb{E}[V_- e^{\lambda Z}] \\ &\le \frac{25}{16} \lambda^2 \, \mathbb{E}\left[e^{\lambda Z} (Z + W)\right]\end{aligned}$$

The right-hand side may be further bounded as follows.

$$\begin{aligned}\mathbb{E}[e^{\lambda Z} (Z + W)] &= \mathbb{E}[Ze^{\lambda Z}] + \mathbb{E}[W]\mathbb{E}[e^{\lambda Z}] + \mathbb{E}[(W - \mathbb{E}[W])e^{\lambda Z}] \\ &\le 2\mathbb{E}[Ze^{\lambda Z}] + \mathbb{E}[W]\mathbb{E}[e^{\lambda Z}] \\ &\quad + \frac{\mathbb{E}[e^{\lambda Z}]}{\lambda} \log \mathbb{E}[e^{\lambda(W - \mathbb{E}[W])}] - \frac{\mathbb{E}[e^{\lambda Z}]}{\lambda} \log \mathbb{E}[e^{\lambda Z}] \\ &\quad \text{(by (3.4))} \\ &\le 2\mathbb{E}[Ze^{\lambda Z}] + \left(\mathbb{E}[W] + \frac{(e-1)v\lambda}{1-\lambda}\right) \mathbb{E}[e^{\lambda Z}]\end{aligned}$$

where at the last step we used Lemma 20 and the fact that $\log \mathbb{E}[e^{\lambda Z}] \ge 0$. This leads, for $\lambda < 8/25$, to

$$\left(\frac{1}{\lambda} - \frac{25}{8}\right) \frac{\mathbb{E}[Ze^{\lambda Z}]}{\mathbb{E}[e^{\lambda Z}]} - \frac{1}{\lambda^2} \log \mathbb{E}[e^{\lambda Z}] \le \frac{25}{16} \left(\mathbb{E}[W] + \frac{(e-1)v\lambda}{4(1-\lambda)}\right) \ .$$

Denoting $F(\lambda) = \mathbb{E}[e^{\lambda Z}]$, this translates into the differential inequality

$$\left(\frac{1}{\lambda} - \frac{25}{8}\right)\frac{F'(\lambda)}{F(\lambda)} - \frac{1}{\lambda^2}\log F(\lambda) \le \frac{25}{16}\left(\mathbb{E}[W] + \frac{(e-1)v}{4}\frac{\lambda}{1-\lambda}\right) .$$

Now observe that the left-hand side is just the derivative of $\frac{1-25\lambda/8}{\lambda}\log F(\lambda)$. Thus, by integrating both sides of the inequality, we have

$$\begin{aligned}\log F(\lambda) &\le \frac{\lambda}{1-25\lambda/8}\left(\mathbb{E}[Z] + \frac{25}{16}\lambda\mathbb{E}[W] + \frac{25(e-1)v}{64}\left(\log\frac{1}{1-\lambda} - \lambda\right)\right)\\ &\le \frac{\lambda}{1-25\lambda/8}\left(\mathbb{E}[Z] + \frac{25}{16}\lambda\mathbb{E}[W] + \frac{25(e-1)v}{64}\frac{\lambda^2}{1-\lambda}\right) .\end{aligned}$$

After centering, we get

$$\log\mathbb{E}[e^{\lambda(Z-\mathbb{E}[Z])}] \le \frac{\lambda^2}{1-25\lambda/8}\frac{25}{16}\left(2\mathbb{E}[Z] + \mathbb{E}[W] + \frac{(e-1)v}{4}\frac{\lambda}{1-\lambda}\right)$$

To bring the bound into a manageable form, after elementary calculations, we obtain

$$\log\mathbb{E}[e^{\lambda(Z-\mathbb{E}[Z])}] \le \frac{25}{16}\frac{\lambda^2}{1-\left(\frac{25}{8} + \left((1 \vee \frac{(e-1)v}{4\mathbb{E}[2Z+W]}\right)\right)\lambda}\left(\mathbb{E}[2Z+W]\right) .$$

Introducing $a = \left(\frac{25}{8} + 1 \vee (e-1)v/(4\mathbb{E}[2Z+W])\right) \ge 4/5$ and applying Markov's inequality, we obtain

$$\mathbb{P}[Z > \mathbb{E}[Z] + t] \le \exp\left(-\sup_{\lambda\in[0,1/a)}\left[\lambda t - \frac{\lambda^2}{1-a\lambda}\frac{25}{16}\mathbb{E}[2Z+W]\right]\right) .$$

Lemma 11 now yields the first inequality of the theorem.

The lower-tail inequality also follows from (7.5), but now we use the second inequality of Proposition 10 for negative values of λ. For all $\lambda < 0$, we have

$$\begin{aligned}\lambda\mathbb{E}[e^{\lambda Z}] - \mathbb{E}[e^{\lambda Z}]\log\mathbb{E}[e^{\lambda Z}] &\le \sum_{i=1}^{n}\lambda^2\mathbb{E}[e^{\lambda Z}(Z^{(i)} - Z)^2\mathbb{1}_{Z^{(i)}>Z}]\\ &= \sum_{i=1}^{n}\lambda^2\mathbb{E}[V_- e^{\lambda Z}]\\ &\le \lambda^2\mathbb{E}[e^{\lambda Z}(Z+W)]\end{aligned}$$

Note that for $\lambda < 0$, (3.4) gives:

$$\begin{aligned}(7.6)\quad \mathbb{E}[e^{\lambda Z}W] &= \mathbb{E}[e^{-\lambda(-Z)}W]\\ &\le \frac{-1}{\lambda}\left(\mathbb{E}\big[\lambda Z e^{\lambda Z}\big] - \mathbb{E}\big[e^{\lambda Z}\big]\log\mathbb{E}\big[e^{\lambda Z}\big] + \mathbb{E}\big[e^{\lambda Z}\big]\log\mathbb{E}\big[e^{-\lambda W}\big]\right) .\end{aligned}$$

Once again, using (3.4) and Lemma 20, we have, for all $\lambda \in (-1,0)$,

$$\begin{aligned}
\mathbb{E}[e^{\lambda Z}(Z+W)] &\le \mathbb{E}[Ze^{\lambda Z}] + \mathbb{E}[W]\mathbb{E}[e^{\lambda Z}] + \mathbb{E}[(W-\mathbb{E}[W])e^{\lambda Z}] \\
&\le \mathbb{E}[Ze^{\lambda Z}] + \mathbb{E}[W]\mathbb{E}[e^{\lambda Z}] \\
&\quad -\mathbb{E}[Ze^{\lambda Z}] + \frac{\mathbb{E}[e^{\lambda Z}]}{\lambda}\log \mathbb{E}[e^{\lambda Z}] - \frac{\mathbb{E}[e^{\lambda Z}]}{\lambda}\log \mathbb{E}[e^{-\lambda(W-\mathbb{E}[W])}] \\
&\le \mathbb{E}[W]\mathbb{E}[e^{\lambda Z}] + \frac{\mathbb{E}[e^{\lambda Z}]}{\lambda}\log \mathbb{E}[e^{\lambda Z}] - \mathbb{E}[e^{\lambda Z}]\frac{\lambda(e-1)v}{1+\lambda}
\end{aligned}$$

which leads to

$$\frac{1}{\lambda}\frac{\mathbb{E}[Ze^{\lambda Z}]}{\mathbb{E}[e^{\lambda Z}]} - \frac{1+\lambda}{\lambda^2}\log \mathbb{E}[e^{\lambda Z}] \le \mathbb{E}[W] - \frac{\lambda(e-1)v}{1+\lambda} \; .$$

Now letting $H(\lambda) = \frac{e^{-\lambda}}{\lambda}\log F(\lambda)$ with $F(\lambda) = \log \mathbb{E}[\exp(\lambda Z)]$, the preceding inequality translates into the differential inequality

$$e^{\lambda}H'(\lambda) \le \mathbb{E}[W] - \frac{\lambda(e-1)v}{1+\lambda}$$

which, after integrating both sides, gives

$$\mathbb{E}[Z] - H(\lambda) \le -\mathbb{E}[W](1-e^{-\lambda}) + \frac{v(e-1)}{2}\frac{\lambda^2 e^{-\lambda}}{1+\lambda} \; .$$

Finally, for all $\lambda \in (-1,0)$,

$$\begin{aligned}
\log \mathbb{E}[e^{\lambda(Z-\mathbb{E}[Z])}] &\le \lambda(e^{\lambda}-1)\mathbb{E}[Z+W] - \frac{v(e-1)}{2}\frac{\lambda^3}{1+\lambda} \\
&\le \lambda^2\mathbb{E}[Z+W]\left(1 - \frac{v(e-1)}{2\mathbb{E}[Z+W]}\frac{\lambda}{1+\lambda}\right) \; ,
\end{aligned}$$

where in the last line we used the fact that $\lambda(\exp(\lambda)-1) \le \lambda^2$ for $\lambda \le 0$. Now, assuming $\lambda \in (-1/2 \wedge \mathbb{E}[Z+W]/(2v(e-1)), 0)$ we may deduce the following upper bound:

$$\log \mathbb{E}[e^{\lambda(Z-\mathbb{E}[Z])}] \le \frac{\lambda^2\mathbb{E}[Z+W]}{1+\frac{\lambda v(e-1)}{\mathbb{E}[Z+W]}} \; . \tag{7.7}$$

Once again, Markov's inequality, for $\lambda \in (-1/2 \wedge 2\mathbb{E}[Z+W]/(v(e-1)), 0]$ and invoking Lemma 11 finishes the proof of the second inequality of Theorem 19. □

Acknowledgments. The authors want to thank M. Zani, G. Blanchard and O. Bousquet for many interesting conversations. They gratefully acknowledge the care and patience of an anonymous referee.

REFERENCES

[1] R. Ahlswede, P. Gács, and J. Kőrner. Bounds on conditional probabilities with applications in multi-user communication. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 34:157–177, 1976. (correction in 39:353–354,1977).

[2] P. Bartlett, S. Boucheron, and G. Lugosi. Model selection and error estimation. *Machine Learning*, 48:85–113, 2002.

[3] S.G. Bobkov and M. Ledoux. On modified logarithmic Sobolev inequalities for Bernouilli and Poisson measures. *J. Funct. Anal.*, 156:347-365, 1998.

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

[5] O. Bousquet. A Bennett Concentration Inequality and Its Application to Suprema of Empirical Processes, *C.R.Acad.Sci. Paris*, to appear.

[6] V.H. de la Peña and E. Giné. *Decoupling: from Dependence to Independence.* Springer, New York, 1999.

[7] A. Dembo. Information inequalities and concentration of measure. *Annals of Probability*, 25:927–939, 1997.

[8] B. Efron and C. Stein. The jackknife estimate of variance. *Annals of Statistics*, 9:586–596, 1981.

[9] G.H. Hall, J.E. Littlewood, and G. Pólya. *Inequalities.* Cambridge University Press, London, 1952.

[10] S. Janson. Poisson approximation for large deviations. *Random Structures and Algorithms,* 1:221–230, 2000.

[11] S. Janson, T. Łuczak, and A. Ruciński. *Random graphs.* John Wiley, New York, 2000.

[12] S. Janson and A. Ruciński. The deletion method for upper tail estimates. Technical Report, 20, Uppsala University, 2000.

[13] S. Janson and A. Ruciński. The infamous upper tail. *Random Structures and Algorithms,* to appear.

[14] J. H. Kim and V. Vu. Concentration of multivariate polynomials and applications. *Combinatorica*, 20:417–434, 2000.

[15] V. Koltchinskii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics*, 30, 2002.

[16] M. Ledoux. Isoperimetry and Gaussian analysis. In P. Bernard, editor, *Lectures on Probability Theory and Statistics*, pages 165–294. Ecole d'Eté de Probabilités de St-Flour XXIV-1994, 1996.

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

[18] M. Ledoux and M. Talagrand. *Probability in Banach Space.* Springer-Verlag, New York, 1991.

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

[20] K. Marton. A simple proof of the blowing-up lemma. *IEEE Transactions on Information Theory*, 32:445–446, 1986.

[21] K. Marton. Bounding $\bar{d}$-distance by informational divergence: a way to prove measure concentration. *Annals of Probability*, 24:857–866, 1996.

[22] K. Marton. A measure concentration inequality for contracting Markov chains. *Geometric and Functional Analysis*, 6:556–571, 1996. Erratum: 7:609–613, 1997.

[23] P. Massart. Optimal constants for Hoeffding type inequalities. Technical report, Mathematiques, Université de Paris-Sud, Report 98.86, 1998.

[24] P. Massart. About the constants in Talagrand's concentration inequalities for empirical processes. *Annals of Probability*, 28:863–884, 2000.

[25] P. Massart. Some applications of concentration inequalities to statistics. *Annales de la Faculté des Sciencies de Toulouse*, IX:245–303, 2000.

[26] C. McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics 1989*, pages 148–188. Cambridge University Press, Cambridge, 1989.

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

[28] E. Rio. Inégalités de concentration pour les processus empiriques de classes de parties. *Probability Theory and Related Fields*, 119:163–175, 2001.

[29] M. Sion. On general minimax theorems. *Pacific Journal of Mathematics.* 8:171-176. 1958. links

[30] J.M. Steele. An Efron-Stein inequality for nonsymmetric statistics. *Annals of Statistics*, 14:753–758, 1986.

[31] J.M. Steele. *Probability Theory and Combinatorial Optimization.* SIAM, CBMS-NSF Regional Conference Series in Applied Mathematics 69, 3600 University City Science Center, Philadelphia, PA 19104, 1996.

[32] M. Talagrand. Concentration of measure and isoperimetric inequalities in product spaces. *Publications Mathématiques de l'I.H.E.S.*, 81:73–205, 1995.

[33] M. Talagrand. New concentration inequalities in product spaces. *Invent. Math.*, 126:505–563, 1996.

[34] M. Talagrand. A new look at independence. *Annals of Probability*, 24:1–34, 1996. (Special Invited Paper).

[35] A.W. van der Waart and J.A. Wellner. *Weak convergence and empirical processes.* Springer-Verlag, New York, 1996.

[36] V. Vu. On the concentration of multivariate polynomials with small expectation. *Random Structures and Algorithms*, 16:344–363, 2000.

[37] V. Vu. A large deviation result on the number of small subgraphs of a random graph. *Combinatorics, Probability, and Computation*, to appear, 2001.

LRI, UMR 8623 CNRS
Bâtiment 490
CNRS-Université Paris-Sud
91405 Orsay-Cedex

Department of Economics,
Pompeu Fabra University
Ramon Trias Fargas 25-27,
08005 Barcelona, Spain,

Mathématiques
Bâtiment 425
Université Paris-Sud
91405 Orsay-Cedex