Concentration-of-measure inequalities

Lecture notes by Gábor Lugosi *

January 27, 2004

Abstract

This text contains some of the material presented at the Summer School on Machine Learning at the Australian National University, Canberra, 2003 and at the Workshop on Combinatorics, Probability and Algorithms at the Centre de Recherches Mathématiques, Université de Montréal.

*Department of Economics, Pompeu Fabra University, Ramon Trias Fargas 25-27, 08005 Barcelona, Spain (email: `lugosi@upf.es`).

Contents

1 Introduction

The laws of large numbers of classical probability theory state that sums of independent random variables are, under very mild conditions, close to their expectation with a large probability. Such sums are the most basic examples of random variables concentrated around their mean. More recent results reveal that such a behavior is shared by a large class of general functions of independent random variables. The purpose of these notes is to give an introduction to some of these general concentration inequalities.

The inequalities discussed in these notes 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 [56], [57]), information-theoretic methods (see Alhswede, Gács, and Körner [1], Marton [50], [51],[52], Dembo [22], Massart [53] and Rio [66]), Talagrand's induction method [75],[73],[74] (see also Luczak and McDiarmid [47], McDiarmid [58], Panchenko [61, 62, 63] and the so-called "entropy method", based on logarithmic Sobolev inequalities, developed by Ledoux [43],[42], see also Bobkov and Ledoux [10], Massart [54], Rio [66], Boucheron, Lugosi, and Massart [12], [13], and Bousquet [14]. Also, various problem-specific methods have been worked out in random graph theory, see Janson, Łuczak, and Ruciński [37] for a survey.

2 Basics

To make these notes self-contained, we first briefly introduce some of the basic inequalities of probability theory.

First of all, recall that for any nonnegative random variable X,

$$\mathbb{E}X = \int_0^\infty \mathbb{P}\{X \geq t\}dt \ .$$

This implies *Markov's inequality*: for any nonnegative random variable X, and $t > 0$,

$$\mathbb{P}\{X \geq t\} \leq \frac{\mathbb{E}X}{t}.$$

If follows from Markov's inequality that if ϕ is a strictly monotonically increasing nonnegative-valued function then for any random variable X and real number t,

$$\mathbb{P}\{X \geq t\} = \mathbb{P}\{\phi(X) \geq \phi(t)\} \leq \frac{\mathbb{E}\phi(X)}{\phi(t)}.$$

An application of this with $\phi(x) = x^2$ is *Chebyshev's inequality*: if X is an arbitrary random variable and $t > 0$, then

$$\mathbb{P}\{|X - \mathbb{E}X| \geq t\} = \mathbb{P}\left\{|X - \mathbb{E}X|^2 \geq t^2\right\} \leq \frac{\mathbb{E}\left[|X - \mathbb{E}X|^2\right]}{t^2} = \frac{\mathrm{Var}\{X\}}{t^2}.$$

More generally taking $\phi(x) = x^q$ $(x \geq 0)$, for any $q > 0$ we have

$$\mathbb{P}\{|X - \mathbb{E}X| \geq t\} \leq \frac{\mathbb{E}\left[|X - \mathbb{E}X|^q\right]}{t^q}.$$

In specific examples one may choose the value of q to optimize the obtained upper bound. Such moment bounds often provide with very sharp estimates of the tail probabilities. A related idea is at the basis of *Chernoff's bounding method*. Taking $\phi(x) = e^{sx}$ where s is an arbitrary positive number, for any random variable X, and any $t > 0$, we have

$$\mathbb{P}\{X \geq t\} = \mathbb{P}\{e^{sX} \geq e^{st}\} \leq \frac{\mathbb{E}e^{sX}}{e^{st}}.$$

In Chernoff's method, we find an $s > 0$ that minimizes the upper bound or makes the upper bound small. Even though Chernoff bounds are never as good as the best moment bound (see Exercise 1), in many cases they are easier to handle.

The *Cauchy-Schwarz inequality* states that if the random variables X and Y have finite second moments ($\mathbb{E}[X^2] < \infty$ and $\mathbb{E}[Y^2] < \infty$), then

$$|\mathbb{E}[XY]| \leq \sqrt{\mathbb{E}[X^2]\mathbb{E}[Y^2]}.$$

We may use this to prove a one-sided improvement of Chebyshev's inequality:

Theorem 1 CHEBYSHEV-CANTELLI INEQUALITY. *Let* $t \geq 0$. *Then*

$$\mathbb{P}\{X - \mathbb{E}X \geq t\} \leq \frac{\mathrm{Var}\{X\}}{\mathrm{Var}\{X\} + t^2}.$$

Proof. We may assume without loss of generality that $\mathbb{E}X = 0$. Then for all t

$$t = \mathbb{E}[t - X] \leq \mathbb{E}[(t - X)\mathbb{1}_{\{X<t\}}].$$

(where $\mathbb{1}$ denotes the indicator function). Thus for $t \geq 0$ from the Cauchy-Schwarz inequality,

$$\begin{aligned} t^2 &\leq \mathbb{E}[(t - X)^2]\mathbb{E}[\mathbb{1}^2_{\{X<t\}}] \\ &= \mathbb{E}[(t - X)^2]\mathbb{P}\{X < t\} \\ &= (\mathrm{Var}\{X\} + t^2)\mathbb{P}\{X < t\}, \end{aligned}$$

that is,

$$\mathbb{P}\{X < t\} \geq \frac{t^2}{\mathrm{Var}\{X\} + t^2},$$

and the claim follows. $\square$

We end this section by recalling a simple association inequality due to Chebyshev (see, e.g., [34]). We note here that association properties may often be used to derive concentration properties. We refer the reader to the survey of Dubdashi and Ranjan [28]

Theorem 2 CHEBYSHEV'S ASSOCIATION INEQUALITY. *Let* f *and* g *be nondecreasing real-valued functions defined on the real line. If* X *is a real-valued random variable, then*

$$\mathbb{E}[f(X)g(X)] \geq \mathbb{E}[f(X)]\mathbb{E}[g(X)]] .$$

If f *is nonincreasing and* g *is nondecreasing then*

$$\mathbb{E}[f(X)g(X)] \leq \mathbb{E}[f(X)]\mathbb{E}[g(X)]] .$$

Proof. Let the random variable Y be distributed as X and independent of it. If f and g are nondecreasing, $(f(x) - f(y))(g(x) - g(y)) \geq 0$ so that

$$\mathbb{E}[(f(X) - f(Y))(g(X) - g(Y))] \geq 0 .$$

Expand this expectation to obtain the first inequality. The proof of the second is similar. □

Exercises

Exercise 1 MOMENTS VS. CHERNOFF BOUNDS. *Show that moment bounds for tail probabilities are always better than Chernoff bounds. More precisely, let* X *be a nonnegative random variable and let* $t > 0$*. The best moment bound for the tail probability* $\mathbb{P}\{X \geq t\}$ *is* $\min_q \mathbb{E}[X^q]t^{-q}$ *where the minimum is taken over all positive integers. The best Chernoff bound is* $\inf_{s>0} \mathbb{E}[e^{s(X-t)}]$*. Prove that*

$$\min_q \mathbb{E}[X^q]t^{-q} \leq \inf_{s>0} \mathbb{E}[e^{s(X-t)}].$$

Exercise 2 FIRST AND SECOND MOMENT METHODS. *Show that if* X *is a nonegative integer-valued random variable then* $\mathbb{P}\{X \neq 0\} \leq \mathbb{E}X$*. Show also that*

$$\mathbb{P}\{X = 0\} \leq \frac{\mathrm{Var}(X)}{\mathrm{Var}(X) + (\mathbb{E}X)^2} .$$

Exercise 3 SUBGAUSSIAN MOMENTS. *We say that a random variable* X *has a subgaussian distribution if there exists a constant* $c > 0$ *such that for all* $s > 0$, $\mathbb{E}[e^{sX}] \leq e^{cs^2}$. *Show that there exists a universal constant* K *such that if* X *is subgaussian, then for every positive integer* q,

$$(\mathbb{E}[X_+^q])^{1/q} \leq K\sqrt{cq} \ .$$

Exercise 4 SUBGAUSSIAN MOMENTS–CONVERSE. *Let* X *be a random variable such that there exists a constant* $c > 0$ *such that*

$$(\mathbb{E}[X_+^q])^{1/q} \leq \sqrt{cq}$$

for every positive integer q. *Show that* X *is subgaussian. More precisely, show that for any* $s > 0$,

$$\mathbb{E}[e^{sX}] \leq \sqrt{2}e^{1/6}e^{ces^2/2} \ .$$

Exercise 5 SUBGEXPONENTIAL MOMENTS. *We say that a random variable* X *has a subexponential distribution if there exists a constant* $c > 0$ *such that for all* $0 < s < 1/c$, $\mathbb{E}[e^{sX}] \leq 1/(1 - cs)$. *Show that if* X *is subexponential, then for every positive integer* q,

$$(\mathbb{E}[X_+^q])^{1/q} \leq \frac{4c}{e}q \ .$$

Exercise 6 SUBEXPONENTIAL MOMENTS–CONVERSE. *Let* X *be a random variable such that there exists a constant* $c > 0$ *such that*

$$(\mathbb{E}[X_+^q])^{1/q} \leq cq$$

for every positive integer q. *Show that* X *is subexponential. More precisely, show that for any* $s > 0$,

$$\mathbb{E}[e^{sX}] \leq \frac{1}{1 - ces} \ .$$

3 Sums of independent random variables

In this introductory section we recall some simple inequalities for sums of independent random variables. Here we are primarily concerned with upper bounds for the probabilities of deviations from the mean, that is, to obtain inequalities for $\mathbb{P}\{S_n - \mathbb{E}S_n \geq t\}$, with $S_n = \sum_{i=1}^n X_i$, where $X_1, \ldots, X_n$ are independent real-valued random variables.

Chebyshev's inequality and independence immediately imply

$$\mathbb{P}\{|S_n - \mathbb{E}S_n| \geq t\} \leq \frac{\mathsf{Var}\{S_n\}}{t^2} = \frac{\sum_{i=1}^n \mathsf{Var}\{X_i\}}{t^2}.$$

In other words, writing $\sigma^2 = \frac{1}{n}\sum_{i=1}^n \mathsf{Var}\{X_i\}$,

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}X_i\right| \geq \epsilon\right\} \leq \frac{\sigma^2}{n\epsilon^2}.$$

This simple inequality is at the basis of the *weak law of large numbers*.

To understand why this inequality is unsatisfactory, recall that, under some additional regularity conditions, the central limit theorem states that

$$\mathbb{P}\left\{\sqrt{\frac{n}{\sigma^2}}\left(\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}X_i\right) \geq y\right\} \to 1 - \Phi(y) \leq \frac{1}{\sqrt{2\pi}}\frac{e^{-y^2/2}}{y},$$

from which we would expect, at least in a certain range of the parameters, something like

$$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}X_i \geq \epsilon\right\} \approx e^{-n\epsilon^2/(2\sigma^2)}. \tag{1}$$

Clearly, Chebyshev's inequality is way off mark in this case, so we should look for something better. In the sequel we prove some of the simplest classical exponential inequalities for the tail probabilities of sums of independent random variables which yield significantly sharper estimates.

3.1 Hoeffding's inequality

Chernoff's bounding method, described in Section 2, is especially convenient for bounding tail probabilities of sums of independent random variables. The reason is that since the expected value of a product of independent random variables equals the product of the expected values, Chernoff's bound becomes

$$\begin{aligned} \mathbb{P}\{S_n - \mathbb{E}S_n \geq t\} &\leq e^{-st}\mathbb{E}\left[\exp\left(s\sum_{i=1}^{n}(X_i - \mathbb{E}X_i)\right)\right] \\ &= e^{-st}\prod_{i=1}^{n}\mathbb{E}\left[e^{s(X_i-\mathbb{E}X_i)}\right] \quad \text{(by independence).} \qquad (2) \end{aligned}$$

Now the problem of finding tight bounds comes down to finding a good upper bound for the moment generating function of the random variables $X_i - \mathbb{E}X_i$. There are many ways of doing this. For bounded random variables perhaps the most elegant version is due to Hoeffding [36]:

Lemma 1 HOEFFDING'S INEQUALITY. *Let* X *be a random variable with* $\mathbb{E}X = 0$, $a \leq X \leq b$. *Then for* $s > 0$,

$$\mathbb{E}\left[e^{sX}\right] \leq e^{s^2(b-a)^2/8}.$$

Proof. Note that by convexity of the exponential function

$$e^{sx} \leq \frac{x-a}{b-a}e^{sb} + \frac{b-x}{b-a}e^{sa} \quad \text{for } a \leq x \leq b.$$

Exploiting $\mathbb{E}X = 0$, and introducing the notation $p = -a/(b-a)$ we get

$$\begin{aligned} \mathbb{E}e^{sX} &\leq \frac{b}{b-a}e^{sa} - \frac{a}{b-a}e^{sb} \\ &= \left(1 - p + pe^{s(b-a)}\right)e^{-ps(b-a)} \\ &\stackrel{\text{def}}{=} e^{\phi(u)}, \end{aligned}$$

where $u = s(b-a)$, and $\phi(u) = -pu + \log(1 - p + pe^u)$. But by straightforward calculation it is easy to see that the derivative of ϕ is

$$\phi'(u) = -p + \frac{p}{p + (1-p)e^{-u}},$$

therefore $\phi(0) = \phi'(0) = 0$. Moreover,

$$\phi''(u) = \frac{p(1-p)e^{-u}}{(p+(1-p)e^{-u})^2} \le \frac{1}{4}.$$

Thus, by Taylor's theorem, for some $\theta \in [0, u]$,

$$\phi(u) = \phi(0) + u\phi'(0) + \frac{u^2}{2}\phi''(\theta) \le \frac{u^2}{8} = \frac{s^2(b-a)^2}{8}. \quad \Box$$

Now we may directly plug this lemma into (2):

$$\begin{aligned}
&\mathbb{P}\{S_n - \mathbb{E}S_n \ge t\} \\
&\quad \le \quad e^{-st}\prod_{i=1}^{n} e^{s^2(b_i-a_i)^2/8} \quad \text{(by Lemma 1)} \\
&\quad = \quad e^{-st}e^{s^2\sum_{i=1}^{n}(b_i-a_i)^2/8} \\
&\quad = \quad e^{-2t^2/\sum_{i=1}^{n}(b_i-a_i)^2} \quad \text{(by choosing } s = 4t/\textstyle\sum_{i=1}^{n}(b_i-a_i)^2\text{)}.
\end{aligned}$$

Theorem 3 HOEFFDING'S TAIL INEQUALITY [36]. *Let* $X_1, \ldots, X_n$ *be independent bounded random variables such that* X_i *falls in the interval* $[a_i, b_i]$ *with probability one. Then for any* $t > 0$ *we have*

$$\mathbb{P}\{S_n - \mathbb{E}S_n \ge t\} \le e^{-2t^2/\sum_{i=1}^{n}(b_i-a_i)^2}$$

and

$$\mathbb{P}\{S_n - \mathbb{E}S_n \le -t\} \le e^{-2t^2/\sum_{i=1}^{n}(b_i-a_i)^2}.$$

The theorem above is generally known as *Hoeffding's inequality*. For binomial random variables it was proved by Chernoff [17] and Okamoto [59].

This inequality has the same form as the one we hoped for based on (1) except that the average variance σ^2 is replaced by the upper bound $(1/4)\sum_{i=1}^{n}(b_i - a_i)^2$. In other words, Hoeffding's inequality ignores information about the variance of the X_i's. The inequalities discussed next provide an improvement in this respect.

3.2 Bernstein's inequality

Assume now without loss of generality that $\mathbb{E}X_i = 0$ for all $i = 1, \ldots, n$. Our starting point is again (2), that is, we need bounds for $\mathbb{E}\left[e^{sX_i}\right]$. Introduce $\sigma_i^2 = \mathbb{E}[X_i^2]$, and

$$F_i = \sum_{r=2}^{\infty} \frac{s^{r-2}\mathbb{E}[X_i^r]}{r!\sigma_i^2}.$$

Since $e^{sx} = 1 + sx + \sum_{r=2}^{\infty} s^r x^r / r!$, we may write

$$\begin{aligned} \mathbb{E}\left[e^{sX_i}\right] &= 1 + s\mathbb{E}[X_i] + \sum_{r=2}^{\infty} \frac{s^r\mathbb{E}[X_i^r]}{r!} \\ &= 1 + s^2\sigma_i^2 F_i \quad (\text{since } \mathbb{E}[X_i] = 0.) \\ &\le e^{s^2\sigma_i^2 F_i}. \end{aligned}$$

Now assume that the X_i's are bounded such that $|X_i| \le c$. Then for each $r \ge 2$,

$$\mathbb{E}[X_i^r] \le c^{r-2}\sigma_i^2.$$

Thus,

$$F_i \le \sum_{r=2}^{\infty} \frac{s^{r-2}c^{r-2}\sigma_i^2}{r!\sigma_i^2} = \frac{1}{(sc)^2}\sum_{r=2}^{\infty}\frac{(sc)^r}{r!} = \frac{e^{sc} - 1 - sc}{(sc)^2}.$$

Thus, we have obtained

$$\mathbb{E}\left[e^{sX_i}\right] \le e^{s^2\sigma_i^2 \frac{e^{sc}-1-sc}{(sc)^2}}.$$

Returning to (2) and using the notation $\sigma^2 = (1/n)\sum \sigma_i^2$, we get

$$\mathbb{P}\left\{\sum_{i=1}^{n} X_i > t\right\} \le e^{n\sigma^2(e^{sc}-1-sc)/c^2 - st}.$$

Now we are free to choose s. The upper bound is minimized for

$$s = \frac{1}{c}\log\left(1 + \frac{tc}{n\sigma^2}\right).$$

Resubstituting this value, we obtain *Bennett's inequality* [7]:

Theorem 4 BENNETT'S INEQUALITY. *Let* $X_1, \ldots, X_n$ *be independent real-valued random variables with zero mean, and assume that* $|X_i| \le c$ *with probability one. Let*

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} \mathrm{Var}\{X_i\}.$$

Then for any $t > 0$*,*

$$\mathbb{P}\left\{\sum_{i=1}^{n} X_i > t\right\} \le \exp\left(-\frac{n\sigma^2}{c^2} h\left(\frac{ct}{n\sigma^2}\right)\right).$$

where $h(u) = (1+u)\log(1+u) - u$ *for* $u \ge 0$*.*

The message of this inequality is perhaps best seen if we do some further bounding. Applying the elementary inequality $h(u) \ge u^2/(2+2u/3)$, $u \ge 0$ (which may be seen by comparing the derivatives of both sides) we obtain a classical inequality of Bernstein [8]:

Theorem 5 BERNSTEIN'S INEQUALITY. *Under the conditions of the previous theorem, for any* $\epsilon > 0$*,*

$$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n} X_i > \epsilon\right\} \le \exp\left(-\frac{n\epsilon^2}{2\sigma^2 + 2c\epsilon/3}\right).$$

We see that, except for the term $2c\epsilon/3$ in the denominator of the exponent, Bernstein's inequality is qualitatively right when we compare it with the central limit theorem (1). Bernstein's inequality points out one more interesting phenomenon: if $\sigma^2 < \epsilon$, then the upper bound behaves like $e^{-n\epsilon}$ instead of the $e^{-n\epsilon^2}$ guaranteed by Hoeffding's inequality. This might be intuitively explained by recalling that a Binomial$(n, \lambda/n)$ distribution can be approximated, for large n, by a Poisson(λ) distribution, whose tail decreases as $e^{-\lambda}$.

Exercises

Exercise 7 *Let $X_1, \ldots, X_n$ be independent random variables, taking their values from $[0,1]$. Denoting $m = \mathbb{E}S_n$, show that for any $t \geq m$,*

$$\mathbb{P}\{S_n \geq t\} \leq \left(\frac{m}{t}\right)^t \left(\frac{n-m}{n-t}\right)^{n-t}.$$

Hint: Proceed by Chernoff's bounding.

Exercise 8 CONTINUATION. *Use the previous exercise to show that*

$$\mathbb{P}\{S_n \geq t\} \leq \left(\frac{m}{t}\right)^t e^{t-m},$$

and for all $\epsilon > 0$,

$$\mathbb{P}\{S_n \geq m(1+\epsilon)\} \leq e^{-mh(\epsilon)},$$

where h is the function defined in Bennett's inequality. Finally,

$$\mathbb{P}\{S_n \leq m(1-\epsilon)\} \leq e^{-m\epsilon^2/2}.$$

(see, e.g., Karp [38], Hagerup and Rüb [33]).

Exercise 9 *Compare the first bound of the previous exercise with the best Chernoff bound for the tail of a Poisson random variable: let Y be a Poisson(m) random variable. Show that*

$$\mathbb{P}\{Y \geq t\} \leq \inf_{s>0} \frac{\mathbb{E}\left[e^{sY}\right]}{e^{st}} = \left(\frac{m}{t}\right)^t e^{t-m}.$$

Use Stirling's formula to show that

$$\mathbb{P}\{Y \geq t\} \geq \mathbb{P}\{Y = t\} \geq \left(\frac{m}{t}\right)^t e^{t-m} \frac{1}{\sqrt{2\pi t}} e^{-1/(12t+1)},$$

Exercise 10 SAMPLING WITHOUT REPLACEMENT. *Let $\mathcal{X}$ be a finite set with N elements, and let $X_1, \ldots, X_n$ be a random sample without replacement from $\mathcal{X}$ and $Y_1, \ldots, Y_n$ a random sample with replacement from $\mathcal{X}$. Show that for any convex real-valued function f,*

$$\mathbb{E}f\left(\sum_{i=1}^n X_i\right) \leq \mathbb{E}f\left(\sum_{i=1}^n Y_i\right).$$

In particular, by taking $f(x) = e^{sx}$*, we see that all inequalities derived for the sums of independent random variables* Y_i *using Chernoff's bounding remain true for the sum of the* X_i*'s. (This result is due to Hoeffding [36].)*

4 The Efron-Stein inequality

The main purpose of these notes is to show how many of the tail inequalities for sums of independent random variables can be extended to general functions of independent random variables. The simplest, yet surprisingly powerful inequality of this kind is known as the *Efron-Stein inequality*. It bounds the variance of a general function. To obtain tail inequalities, one may simply use Chebyshev's inequality.

Let $\mathcal{X}$ be some set, and let $g : \mathcal{X}^n \to \mathbb{R}$ be a measurable function of n variables. We derive inequalities for the difference between the random variable $Z = g(X_1, \ldots, X_n)$ and its expected value $\mathbb{E}Z$ when $X_1, \ldots, X_n$ are arbitrary independent (not necessarily identically distributed!) random variables taking values in $\mathcal{X}$.

The main inequalities of this section follow from the next simple result. To simplify notation, we write $\mathbb{E}_i$ for the expected value with respect to the variable X_i, that is, $\mathbb{E}_i Z = \mathbb{E}[Z|X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n]$.

Theorem 6

$$\mathsf{Var}(Z) \leq \sum_{i=1}^{n} \mathbb{E}\left[(Z - \mathbb{E}_i Z)^2\right] .$$

Proof. The proof is based on elementary properties of conditional expectation. Recall that if X and Y are arbitrary bounded random variables, then $\mathbb{E}[XY] = \mathbb{E}[\mathbb{E}[XY|Y]] = \mathbb{E}[Y\mathbb{E}[X|Y]]$.

Introduce the notation $V = Z - \mathbb{E}Z$, and define

$$V_i = \mathbb{E}[Z|X_1, \ldots, X_i] - \mathbb{E}[Z|X_1, \ldots, X_{i-1}], \qquad i = 1, \ldots, n.$$

Clearly, $V = \sum_{i=1}^{n} V_i$. (Thus, V is written as a sum of martingale differences.) Then

$$\begin{aligned}
\mathrm{Var}(Z) &= \mathbb{E}\left[\left(\sum_{i=1}^{n} V_i\right)^2\right] \\
&= \mathbb{E}\sum_{i=1}^{n} V_i^2 + 2\mathbb{E}\sum_{i>j} V_i V_j \\
&= \mathbb{E}\sum_{i=1}^{n} V_i^2 \ ,
\end{aligned}$$

since, for any $i > j$,

$$\mathbb{E}V_iV_j = \mathbb{E}\mathbb{E}\left[V_iV_j|X_1,\ldots,X_j\right] = \mathbb{E}\left[V_j\mathbb{E}\left[V_i|X_1,\ldots,X_j\right]\right] = 0 \ .$$

To bound $\mathbb{E}V_i^2$, note that, by Jensen's inequality,

$$\begin{aligned}
V_i^2 &= (\mathbb{E}[Z|X_1,\ldots,X_i] - \mathbb{E}[Z|X_1,\ldots,X_{i-1}])^2 \\
&= \left(\mathbb{E}\left[\mathbb{E}[Z|X_1,\ldots,X_n] - \mathbb{E}[Z|X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_n]\Big|X_1,\ldots,X_i\right]\right)^2 \\
&\le \mathbb{E}\left[(\mathbb{E}[Z|X_1,\ldots,X_n] - \mathbb{E}[Z|X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_n])^2\Big|X_1,\ldots,X_i\right] \\
&= \mathbb{E}\left[(Z-\mathbb{E}_iZ)^2\Big|X_1,\ldots,X_i\right] \ .
\end{aligned}$$

Taking expected values on both sides, we obtain the statement. □

Now the Efron-Stein inequality follows easily. To state the theorem, let $X_1',\ldots,X_n'$ form an independent copy of $X_1,\ldots,X_n$ and write

$$Z_i' = g(X_1,\ldots,X_i',\ldots,X_n) \ .$$

Theorem 7 EFRON-STEIN INEQUALITY (EFRON AND STEIN [30], STEELE [71]).

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

Proof. The stetement follows by Theorem 6 simply by using (conditionally) the elementary fact that if X and Y are independent and identically distributed random variables, then $\mathrm{Var}(X) = (1/2)\mathbb{E}[(X-Y)^2]$, and therefore

$$\mathbb{E}_i\left[(Z - \mathbb{E}_i Z)^2\right] = \frac{1}{2}\mathbb{E}_i\left[(Z - Z_i')^2\right] \ . \ \square$$

Remark. Observe that in the case when $Z = \sum_{i=1}^n X_i$ is a sum of independent random variables (of finite variance) then the inequality in Theorem 7 becomes an equality. Thus, the bound in the Efron-Stein inequality is, in a sense, not improvable. This example also shows that, among all functions of independent random variables, sums, in some sense, are the least concentrated. Below we will see other evidences for this extremal property of sums.

Another useful corollary of Theorem 6 is obtained by recalling that, for any random variable X, $\mathrm{Var}(X) \le \mathbb{E}[(X-a)^2]$ for any constant $a \in \mathbb{R}$. Using this fact conditionally, we have, for every $i = 1, \ldots, n$,

$$\mathbb{E}_i\left[(Z - \mathbb{E}_i Z)^2\right] \le \mathbb{E}_i\left[(Z - Z_i)^2\right]$$

where $Z_i = g_i(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$ for arbitrary measurable functions $g_i : \mathcal{X}^{n-1} \to \mathbb{R}$ of $n-1$ variables. Taking expected values and using Theorem 6 we have the following.

Theorem 8

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

In the next two sections we specialize the Efron-Stein inequality and its variant Theorem 8 to functions which satisfy some simple easy-to-verify properties.

4.1 Functions with bounded differences

We say that a function $g : \mathcal{X}^n \to \mathbb{R}$ has the *bounded differences property* if for some nonnegative constants $c_1, \ldots, c_n$,

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

In other words, if we change the i-th variable of g while keeping all the others fixed, the value of the function cannot change by more than c_i. Then the Efron-Stein inequality implies the following:

Corollary 1 *If g has the bounded differences property with constants $c_1, \ldots, c_n$, then*

$$\mathrm{Var}(Z) \leq \frac{1}{2} \sum_{i=1}^{n} c_i^2 \ .$$

Next we list some interesting applications of this corollary. In all cases the bound for the variance is obtained effortlessly, while a direct estimation of the variance may be quite involved.

Example. BIN PACKING. This is one of the basic operations research problems. Given n numbers $x_1, \ldots, x_n \in [0, 1]$, the question is the following: what is the minimal number of "bins" into which these numbers can be packed such that the sum of the numbers in each bin doesn't exceed one. Let $g(x_1, \ldots, x_n)$ be this minimum number. The behavior of $Z = g(X_1, \ldots, X_n)$, when $X_1, \ldots, X_n$ are independent random variables, has been extensively studied, see, for example, Rhee and Talagrand [65], Rhee [64], Talagrand [73]. Now clearly by changing one of the x_i's, the value of $g(x_1, \ldots, x_n)$ cannot change by more than one, so we have

$$\mathrm{Var}(Z) \leq \frac{n}{2} \ .$$

However, sharper bounds may be proved by using Talagrand's convex distance inequality discussed later.

Example. LONGEST COMMON SUBSEQUENCE. This problem has been studied intensively for about 20 years now, see Chvátal and Sankoff[18], Deken [21], Dančík and Paterson [20], Steele [70, 72], The simplest version is the following: Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ be two sequences of coin flips. Define Z as the longest subsequence which appears in both sequences, that is,

$$\begin{aligned} Z &= \max\{k : X_{i_1} = Y_{j_1}, \ldots, X_{i_k} = Y_{j_k}, \\ &\qquad \text{where } 1 \leq i_1 < \cdots < i_k \leq n \text{ and } 1 \leq j_1 < \cdots < j_k \leq n\}. \end{aligned}$$

The behavior of $\mathbb{E}Z$ has been investigated in many papers. It is known that $\mathbb{E}[Z]/n$ converges to some number γ, whose value is unknown. It is conjectured to be $2/(1+\sqrt{2})$, and it is known to fall between 0.75796 and 0.83763. Here we are concerned with the concentration of Z. A moment's thought reveals that changing one bit can't change the length of the longest common subsequence by more than one, so Z satisfies the bounded differences property with $c_i = 1$. Consequently,

$$\mathsf{Var}\{Z\} \le \frac{n}{2},$$

(see Steele [71]). Thus, by Chebyshev's inequality, with large probability, Z is within a constant times $\sqrt{n}$ of its expected value. In other words, it is strongly concentrated around the mean, which means that results about $\mathbb{E}Z$ really tell us about the behavior of the longest common subsequence of two random strings.

Example. UNIFORM DEVIATIONS. One of the central quantities of statistical learning theory and empirical process theory is the following: let $X_1, \ldots, X_n$ be i.i.d. random variables taking their values in some set $\mathcal{X}$, and let $\mathcal{A}$ be a collection of subsets of $\mathcal{X}$. Let μ denote the distribution of X_1, that is, $\mu(A) = \mathbb{P}\{X_1 \in A\}$, and let μ_n denote the empirical distribution:

$$\mu_n(A) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_n \in A\}} \ .$$

The quantity of interest is

$$Z = \sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)|.$$

If $\lim_{n\to\infty} \mathbb{E}Z = 0$ for every distribution of the X_i's, then $\mathcal{A}$ is called a *uniform Glivenko-Cantelli class*, and Vapnik and Chervonenkis [79] gave a beautiful combinatorial characterization of such classes. But regardless of what $\mathcal{A}$ is, by changing one X_i, Z can change by at most $1/n$, so regardless of the behavior of $\mathbb{E}Z$, we always have

$$\mathsf{Var}(Z) \le \frac{1}{2n} \ .$$

For more information on the behavior of Z and its role in learning theory see, for example, Devroye, Györfi, and Lugosi [26], Vapnik [78], van der Vaart and Wellner [76], Dudley [29].

Next we show how a closer look at the the Efron-Stein inequality implies a significantly better bound for the variance of Z. We do this in a slightly more general framework of empirical processes. Let $\mathcal{F}$ be a class of real-valued functions (no boundedness is assumed!) and define $Z = g(X_1, \dots, x_n) = \sup_{f\in\mathcal{F}} \sum_{j=1}^n f(X_j)$. Observe that, by symmetry, the Efron-Stein inequality may be rewritten as

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

Let $f^* \in \mathcal{F}$ denote the (random) function which achieves the supremum in the definition of Z, that is, $Z = \sum_{j=1}^n f^*(X_j)$. Then clearly,

$$(Z - Z_i')^2 \mathbb{1}_{Z_i' < Z} \le \left(f^*(X_i) - f^*(X_i')\right)^2$$

and therefore

$$\begin{aligned}
\mathrm{Var}(Z) &\le \mathbb{E}\left[\sup_{f\in\mathcal{F}} \sum_{i=1}^n \left(f(X_i) - f(X_i')\right)^2\right] \\
&\le \mathbb{E}\left[\sup_{f\in\mathcal{F}} \sum_{i=1}^n \left(2f(X_i)^2 + 2f(X_i')^2\right)\right] \\
&\le 4\mathbb{E}\left[\sup_{f\in\mathcal{F}} \sum_{i=1}^n f(X_i)^2\right] .
\end{aligned}$$

For functions $f \in \mathcal{F}$ are taking values in the interval $[-1, 1]$, then from just the bounded differences property we derived $\mathrm{Var}(Z) \le 2n$. The new bound may be a significant improvement whenever the maximum of the variances $\sum_{i=1}^n f(X_i)^2$ of the functions in $\mathcal{F}$ is small. More importantly, in deriving the new bound we have not assumed any boundedness of the functions f. The exponential tail inequality due to Talagrand [74] extends this variance inequality, and is one of the most important recent results of the theory of empirical processes, see also Ledoux [43], Massart [54], Rio [66], and Bousquet [14].

Example. MINIMUM OF THE EMPIRICAL LOSS. Concentration inequalities have been used as a key tool in recent developments of model selection methods in statistical learning theory. For the background we refer to the the recent work of Koltchinskii Panchenko [40], Massart [55], Bartlett, Boucheron, and Lugosi [4], Lugosi and Wegkamp [49], Bousquet [15].

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 $D_n = (\langle X_i, Y_i \rangle)_{i \le n}$ 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 loss

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

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

$$\ell(y, y') = \mathbb{1}_{y \neq y'} \ .$$

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

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

Corollary 1 immediately implies that $\mathsf{Var}(\widehat{L}) \le 1/(2n)$. However, a more careful application of the Efron-Stein inequality reveals that $\widehat{L}$ may be much more concentrated than predicted by this simple inequality. Getting tight results for the fluctuations of $\widehat{L}$ provides better insight into the calibration of penalties in certain model selection methods.

Let $Z = n\widehat{L}$ and let Z_i' be defined as in Theorem 7, that is,

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

where $\langle X_i{}', Y_i{}' \rangle$ is independent of D_n and has the same distribution as $\langle X_i, Y_i \rangle$. Now the convenient form of the Efron-Stein inequality is the following:

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

Let f^* denote a (possibly non-unique) minimizer of the empirical risk so that $Z = \sum_{j=1}^{n}\ell(f^*(X_j), Y_j)$. The key observation is that

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

Thus,

$$\sum_{i=1}^{n}\mathbb{E}\left[(Z-Z_i')^2\mathbb{1}_{Z_i'>Z}\right] \le \mathbb{E}\sum_{i:\ell(f^*(X_i),Y_i)=0}\mathbb{E}_{X_i',Y_i'}[\ell(f^*(X_i'), Y_i')] \le n\mathbb{E}L(f^*)$$

where $\mathbb{E}_{X_i',Y_i'}$ denotes expectation with respect to the variables X_i', Y_i' and for each $f \in \mathcal{F}$, $L(f) = \mathbb{E}\ell(f(X), Y)$ is the true (expected) loss of f. Therefore, the Efron-Stein inequality implies that

$$\mathrm{Var}(\widehat{L}) \le \frac{\mathbb{E}L(f^*)}{n} \ .$$

This is a significant improvement over the bound $1/(2n)$ whenever $\mathbb{E}L(f^*)$ is much smaller than $1/2$. This is very often the case. For example, we have

$$L(f^*) = \widehat{L} - (L_n(f^*) - L(f^*)) \le \frac{Z}{n} + \sup_{f\in\mathcal{F}}(L(f) - L_n(f))$$

so that we obtain

$$\mathrm{Var}(\widehat{L}) \le \frac{\mathbb{E}\widehat{L}}{n} + \frac{\mathbb{E}\sup_{f\in\mathcal{F}}(L(f) - L_n(f))}{n} \ .$$

In most cases of interest, $\mathbb{E}\sup_{f\in\mathcal{F}}(L(f) - L_n(f))$ may be bounded by a constant (depending on $\mathcal{F}$) times $n^{-1/2}$ (see, e.g., Lugosi [48]) and then the second term on the right-hand side is of the order of $n^{-3/2}$. For exponential concentration inequalities for $\widehat{L}$ we refer to Boucheron, Lugosi, and Massart [13].

Example. KERNEL DENSITY ESTIMATION. Let $X_1, \ldots, X_n$ be i.i.d. samples drawn according to some (unknown) density f on the real line. The density is estimated by the kernel estimate

$$f_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right),$$

where $h > 0$ is a smoothing parameter, and K is a nonnegative function with $\int K = 1$. The performance of the estimate is measured by the L_1 error

$$Z = g(X_1, \ldots, X_n) = \int |f(x) - f_n(x)| dx.$$

It is easy to see that

$$\begin{aligned} |g(x_1, \ldots, x_n) - g(x_1, \ldots, x_i', \ldots, x_n)| &\le \frac{1}{nh} \int \left| K\left(\frac{x - x_i}{h}\right) - K\left(\frac{x - x_i'}{h}\right) \right| dx \\ &\le \frac{2}{n}, \end{aligned}$$

so without further work we get

$$\mathrm{Var}(Z) \le \frac{2}{n}.$$

It is known that for every f, $\sqrt{n}\mathbb{E}g \to \infty$ (see Devroye and Györfi [25]) which implies, by Chebyshev's inequality, that for every $\epsilon > 0$

$$\mathbb{P}\left\{ \left| \frac{Z}{\mathbb{E}Z} - 1 \right| \ge \epsilon \right\} = \mathbb{P}\left\{ |Z - \mathbb{E}Z| \ge \epsilon \mathbb{E}Z \right\} \le \frac{\mathrm{Var}(Z)}{\epsilon^2 (\mathbb{E}Z)^2} \to 0$$

as $n \to \infty$. That is, $Z/\mathbb{E}Z \to 0$ in probability, or in other words, Z is *relatively stable*. This means that the random L_1-error behaves like its expected value. This result is due to Devroye [23], [24]. For more on the behavior of the L_1 error of the kernel density estimate we refer to Devroye and Györfi [25], Devroye and Lugosi [27].

4.2 Self-bounding functions

Another simple property which is satisfied for many important examples is the so-called *self-bounding* property. We say that a nonnegative function

$g : \mathcal{X}^n \to \mathbb{R}$ has the self-bounding property if there exist functions $g_i : \mathcal{X}^{n-1} \to \mathbb{R}$ such that for all $x_1, \ldots, x_n \in \mathcal{X}$ and all $i = 1, \ldots, n$,

$$0 \leq g(x_1, \ldots, x_n) - g_i(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) \leq 1$$

and also

$$\sum_{i=1}^{n} \left(g(x_1, \ldots, x_n) - g_i(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) \right) \leq g(x_1, \ldots, x_n) \ .$$

Concentration properties for such functions have been studied by Boucheron, Lugosi, and Massart [12], Rio [66], and Bousquet [14]. For self-bounding functions we clearly have

$$\sum_{i=1}^{n} \left(g(x_1, \ldots, x_n) - g_i(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) \right)^2 \leq g(x_1, \ldots, x_n) \ .$$

and therefore Theorem 8 implies

Corollary 2 *If* g *has the self-bounding property, then*

$$\mathrm{Var}(Z) \leq \mathbb{E}Z \ .$$

Next we mention some applications of this simple corollary. It turns out that in many cases the obtained bound is a significant improvement over what we would obtain by using simply Corollary 1.

Remark. RELATIVE STABILITY. Bounding the variance of Z by its expected value implies, in many cases, the relative stability of Z. A sequence of nonnegative random variables (Z_n) is said to be relatively stable if $Z_n/\mathbb{E}Z_n \to 1$ in probability. This property guarantees that the random fluctuations of Z_n around its expectation are of negligible size when compared to the expectation, and therefore most information about the size of Z_n is given by $\mathbb{E}Z_n$. If Z_n has the self-bounding property, then, by Chebyshev's inequality, for all $\epsilon > 0$,

$$\mathbb{P}\left\{ \left| \frac{Z_n}{\mathbb{E}Z_n} - 1 \right| > \epsilon \right\} \leq \frac{\mathrm{Var}(Z_n)}{\epsilon^2 (\mathbb{E}Z_n)^2} \leq \frac{1}{\epsilon^2 \mathbb{E}Z_n} \ .$$

Thus, for relative stability, it suffices to have $\mathbb{E}Z_n \to \infty$.

Example. EMPIRICAL PROCESSES. A typical example of self-bounding functions is the supremum of nonnegative empirical processes. Let $\mathcal{F}$ be a class of functions taking values in the interval $[0,1]$ and consider $Z = g(X_1, \ldots, x_n) = \sup_{f\in\mathcal{F}} \sum_{j=1}^n f(X_j)$. (A special case of this is mentioned above in the example of uniform deviations.) Defining $g_i = g'$ for $i = 1, \ldots, n$ with $g'(x_1, \ldots, x_{n-1}) = \sup_{f\in\mathcal{F}} \sum_{j=1}^{n-1} f(X_j)$ (so that $Z_i = \sup_{f\in\mathcal{F}} \sum_{\substack{j=1 \\ j\neq i}}^n f(X_j)$) and letting $f^* \in \mathcal{F}$ be a function for which $Z = \sum_{j=1}^n f^*(X_j)$, one obviously has

$$0 \leq Z - Z_i \leq f^*(X_i) \leq 1$$

and therefore

$$\sum_{i=1}^n (Z - Z_i) \leq \sum_{i=1}^n f^*(X_i) = Z.$$

(Here we have assumed that the supremum is always achieved. The modification of the argument for the general case is straightforward.) Thus, by Corollary 2 we obtain $\mathsf{Var}(Z) \leq \mathbb{E}Z$. Note that Corollary 1 implies $\mathsf{Var}(Z) \leq n/2$. In some important applications $\mathbb{E}Z$ may be significantly smaller than $n/2$ and the improvement is essential.

Example. RADEMACHER AVERAGES. A less trivial example for self-bounding functions is the one of Rademacher averages. Let $\mathcal{F}$ be a class of functions with values in $[-1,1]$. If $\sigma_1, \ldots, \sigma_n$ denote independent symmetric $\{-1,1\}$-valued random variables, independent of the X_i's (the so-called Rademacher random variables), then we define the *conditional Rademacher average* as

$$Z = \mathbb{E}\left[\sup_{f\in\mathcal{F}} \sum_{j=1}^n \sigma_j f(X_j) | X_1^n\right] .$$

(Thus, the expected value is taken with respect to the Rademacher variables and Z is a function of the X_i's.) Quantities like Z have been known to measure effectively the complexity of model classes in statistical learning theory, see, for example, Koltchinskii [39], Bartlett, Boucheron, and Lugosi [4], Bartlett and Mendelson [5], Bartlett, Bousquet, and Mendelson [60]. It is immediate that Z has the bounded differences property and Corollary 1 implies $\mathsf{Var}(Z) \le n/2$. However, this bound may be improved by observing that Z also has the self-bounding property, and therefore $\mathsf{Var}(Z) \le \mathbb{E}Z$. Indeed, defining

$$Z_i = \mathbb{E}\left[\sup_{f\in\mathcal{F}} \sum_{\substack{j=1 \\ j\neq i}}^{n} \sigma_j f(X_j) | X_1^n\right]$$

it is easy to see that $0 \le Z - Z_i \le 1$ and $\sum_{i=1}^n (Z - Z_i) \le Z$ (the details are left as an exercise). The improvement provided by Lemma 2 is essential since it is well-known in empirical process theory and statistical learning theory that in many cases when $\mathcal{F}$ is a relatively small class of functions, $\mathbb{E}Z$ may be bounded by something like $Cn^{1/2}$ where the constant C depends on the class $\mathcal{F}$, see, e.g., Vapnik [78], van der Vaart and Wellner [76], Dudley [29].

Configuration functions

An important class of functions satisfying the self-bounding property consists of the so-called *configuration functions* defined by Talagrand [73, section 7]. Our definition, taken from [12] is a slight modification of Talagrand's.

Assume that we have a property P defined over the union of finite products of a set $\mathcal{X}$, that is, a sequence of sets $P_1 \in \mathcal{X}, P_2 \in \mathcal{X}\times\mathcal{X}, \ldots, P_n \in \mathcal{X}^n$. We say that $(x_1, \ldots x_m) \in \mathcal{X}^m$ satisfies the property P if $(x_1, \ldots x_m) \in P_m$. We assume that P is *hereditary* in the sense that if $(x_1, \ldots x_m)$ satisfies P then so does any subsequence $(x_{i_1}, \ldots x_{i_k})$ of $(x_1, \ldots x_m)$. The function g_n that maps any tuple $(x_1, \ldots x_n)$ to the size of the largest subsequence satisfying P is the *configuration function* associated with property P.

Corollary 2 implies the following result:

Corollary 3 *Let g_n be a configuration function, and let $Z = g_n(X_1, \ldots, X_n)$, where $X_1, \ldots, X_n$ are independent random variables. Then for any $t \ge 0$,*

$$\mathsf{Var}(Z) \le \mathbb{E}Z\ .$$

Proof. By Corollary 2 it suffices to show that any configuration function is self bounding. Let $Z_i = g_{n-1}(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$. The condition $0 \le Z - Z_i \le 1$ is trivially satisfied. On the other hand, assume that $Z = k$ and let $\{X_{i_1}, \ldots, X_{i_k}\} \subset \{X_1, \ldots, X_n\}$ be a subsequence of cardinality k such that $f_k(X_{i_1}, \ldots, X_{i_k}) = k$. (Note that by the definition of a configuration function such a subsequence exists.) Clearly, if the index i is such that $i \notin \{i_1, \ldots, i_k\}$ then $Z = Z_i$, and therefore

$$\sum_{i=1}^{n} (Z - Z_i) \le Z$$

is also satisfied, which concludes the proof. □

To illustrate the fact that configuration functions appear rather naturally in various applications, we describe some examples originating from different fields.

Example. NUMBER OF DISTINCT VALUES IN A DISCRETE SAMPLE. Let $X_1, \ldots, X_n$ be independent, identically distributed random variables taking their values on the set of positive integers such that $\mathbb{P}\{X_1 = k\} = p_k$, and let Z denote the number of distinct values taken by these n random variables. Then we may write

$$Z = \sum_{i=1}^{n} \mathbb{1}_{\{X_i \neq X_1, \ldots, X_i \neq X_{i-1}\}},$$

so the expected value of Z may be computed easily:

$$\mathbb{E}Z = \sum_{i=1}^{n} \sum_{j=1}^{\infty} (1 - p_j)^{i-1} p_j.$$

It is easy to see that $\mathbb{E}[Z]/n \to 0$ as $n \to \infty$ (see Exercise 13). But how concentrated is the distribution of Z? Clearly, Z satisfies the bounded differences property with $c_i = 1$, so Corollary 1 implies $\mathrm{Var}(Z) \le n/2$ so $Z/n \to 0$ *in probability* by Chebyshev's inequality. On the other hand, it is obvious that Z is a configuration function associated to the property of "distinctness", and by Corollary 3 we have

$$\mathrm{Var}(Z) \le \mathbb{E}Z$$

which is a significant improvement since $\mathbb{E}Z = o(n)$.

Example. VC DIMENSION. One of the central quantities in statistical learning theory is the *Vapnik-Chervonenkis dimension*, see Vapnik and Chervonenkis [79, 80], Blumer, Ehrenfeucht, Haussler, and Warmuth [9], Devroye, Györfi, and Lugosi [26], Anthony and Bartlett [2], Vapnik [78], etc.

Let $\mathcal{A}$ be an arbitrary collection of subsets of $\mathcal{X}$, and let $x_1^n = (x_1, \ldots, x_n)$ be a vector of n points of $\mathcal{X}$. Define the *trace* of $\mathcal{A}$ on x_1^n by

$$\mathrm{tr}(x_1^n) = \{A \cap \{x_1, \ldots, x_n\} : A \in \mathcal{A}\}.$$

The *shatter coefficient*, (or *Vapnik-Chervonenkis growth function*) of $\mathcal{A}$ in x_1^n is $T(x_1^n) = |\mathrm{tr}(x_1^n)|$, the size of the trace. $T(x_1^n)$ is the number of different subsets of the n-point set $\{x_1, \ldots, x_n\}$ generated by intersecting it with elements of $\mathcal{A}$. A subset $\{x_{i_1}, \ldots, x_{i_k}\}$ of $\{x_1, \ldots, x_n\}$ is said to be *shattered* if $2^k = T(x_{i_1}, \ldots, x_{i_k})$. The VC *dimension* $D(x_1^n)$ of $\mathcal{A}$ (with respect to x_1^n) is the cardinality k of the largest shattered subset of x_1^n. From the definition it is obvious that $g_n(x_1^n) = D(x_1^n)$ is a configuration function (associated to the property of "shatteredness", and therefore if $X_1, \ldots, X_n$ are independent random variables, then

$$\mathrm{Var}(D(X_1^n)) \le \mathbb{E}D(X_1^n) \ .$$

Example. INCREASING SUBSEQUENCES. Consider a vector $x_1^n = (x_1, \ldots, x_n)$ of n different numbers in $[0, 1]$. The positive integers $i_1 < i_2 < \cdots < i_m$ form an *increasing subsequence* if $x_{i_1} < x_{i_2} < \cdots < x_{i_m}$ (where $i_1 \ge 1$ and $i_m \le n$). Let $L(x_1^n)$ denote the length of the longest increasing subsequence. $g_n(x_1^n) = L(x_1^n)$ is a clearly a configuration function (associated with the "increasing sequence" property), and therefore if $X_1, \ldots, X_n$ are independent random variables such that they are different with probability one (it suffices if every X_i has an absolutely continuous distribution) then $\mathrm{Var}(L(X_1^n)) \le \mathbb{E}L(X_1^n)$. If the X_i's are uniformly distributed in $[0, 1]$ then it is known that $\mathbb{E}L(X_1^n) \sim 2\sqrt{n}$, see Logan and Shepp [46], Groeneboom [32]. The obtained bound for the variance is apparently loose. A difficult result of Baik, Deift, and Johansson [3] implies that $\mathrm{Var}(L(X_1^n)) = O(n^{1/3})$.

For early work on the concentration on $L(X)$ we refer to Frieze [31], Bollobás and Brightwell [11], and Talagrand [73].

Exercises

Exercise 11 *Assume that the random variables* $X_1, \ldots, X_n$ *are independent and binary {0,1}-valued with* $\mathbb{P}\{X_i = 1\} = p_i$ *and that* g *has the bounded differences property with constants* $c_1, \ldots, c_n$. *Show that*

$$\mathrm{Var}(Z) \leq \sum_{i=1}^{n} c_i^2 p_i (1 - p_i).$$

Exercise 12 *Complete the proof of the fact that the conditional Rademacher average has the self-bounding property.*

Exercise 13 *Consider the example of the number of distinct values in a discrete sample described in the text. Show that* $\mathbb{E}[Z]/n \to 0$ *as* $n \to \infty$. *Calculate explicitely* $\mathrm{Var}(Z)$ *and compare it with the upper bound obtained by Theorem 8.*

Exercise 14 *Let* Z *be the number of triangles in a random graph* $\mathcal{G}(n, p)$. *Calculate the variance of* Z *and compare it with what you get by using the Efron-Stein inequality to estimate it. (In the* $\mathcal{G}(n, p)$ *model for random graphs, the random graph* $G = (V, E)$ *with vertex set* V *(*$|V| = n$*) and edge set* E *is generated by starting from the complete graph with* n *vertices and deleting each edge independently from the others with probability* $1 - p$. *A triangle is a complete three-vertex subgraph.)*

5 The entropy method

In the previous section we saw that the Efron-Stein inequality serves as a powerful tool for bounding the variance of general functions of independent random variables. Then, via Chebyshev's inequality, one may easily bound the tail probabilities of such functions. However, just as in the case of sums of independent random variables, tail bounds based on inequalities for the variance are often not satisfactory, and essential improvements are possible. The purpose of this section is to present a methodology which allows one to obtain exponential tail inequalities in many cases. The pursuit of such inequalities has been an important topics in probability theory in the last few decades. Originally, martingale methods dominated the research (see, e.g., McDiarmid [56], [57], Rhee and Talagrand [65], Shamir and Spencer [68]) but independently information-theoretic methods were also used with success (see Alhswede, Gács, and Körner [1], Marton [50], [51],[52], Dembo [22], Massart [53], Rio [66], and Samson [67]). Talagrand's induction method [75],[73],[74] caused an important breakthrough both in the theory and applications of exponential concentration inequalities. In this section we focus on so-called "entropy method", based on logarithmic Sobolev inequalities developed by Ledoux [43],[42], see also Bobkov and Ledoux [10], Massart [54], Rio [66], Boucheron, Lugosi, and Massart [12], [13], and Bousquet [14]. This method makes it possible to derive exponential analogues of the Efron-Stein inequality perhaps the simplest way.

The method is based on an appropriate modification of the "tensorization" inequality Theorem 6. In order to prove this modification, we need to recall some of the basic notions of information theory. To keep the material at an elementary level, we prove the modified tensorization inequality for discrete random variables only. The extension to arbitrary distributions is straightforward.

5.1 Basic information theory

In this section we summarize some basic properties of the entropy of a discrete-valued random variable. For a good introductory book on information theory we refer to Cover and Thomas [19].

Let X be a random variable taking values in the countable set $\mathcal{X}$ with distribution $\mathbb{P}\{X = x\} = p(x)$, $x \in \mathcal{X}$. The *entropy* of X is defined by

$$H(X) = \mathbb{E}[-\log p(X)] = -\sum_{x \in \mathcal{X}} p(x) \log p(x)$$

(where log denotes natural logarithm and $0 \log 0 = 0$). If X, Y is a pair of discrete random variables taking values in $X \times \mathcal{Y}$ then the *joint entropy* $H(X, Y)$ of X and Y is defined as the entropy of the pair (X, Y). The *conditional entropy* $H(X|Y)$ is defined as

$$H(X|Y) = H(X, Y) - H(Y) .$$

Observe that if we write $p(x, y) = \mathbb{P}\{X = x, Y = y\}$ and $p(x|y) = \mathbb{P}\{X = x|Y = y\}$ then

$$H(X|Y) = -\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x, y) \log p(x|y)$$

from which we see that $H(X|Y) \geq 0$. It is also easy to see that the defining identity of the conditional entropy remains true conditionally, that is, for any three (discrete) random variables X, Y, Z,

$$H(X, Y|Z) = H(Y|Z) + H(X|Y, Z) .$$

(Just add $H(Z)$ to both sides and use the definition of the conditional entropy.) A repeated application of this yields the *chain rule for entropy*: for arbitrary disctrete random variables $X_1, \ldots, X_n$,

$$H(X_1, \ldots, X_n) = H(X_1) + H(X_2|X_1) + H(X_3|X_1, X_2) + \cdots + H(X_n|X_1, \ldots, X_{n-1}) .$$

Let P and Q be two probability distributions over a countable set $\mathcal{X}$ with probability mass functions p and q. Then the *Kullback-Leibler divergence* or *relative entropy* of P and Q is

$$D(P\|Q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} .$$

Since $\log x \leq x - 1$,

$$D(P\|Q) = -\sum_{x \in \mathcal{X}} p(x) \log \frac{q(x)}{p(x)} \geq -\sum_{x \in \mathcal{X}} p(x) \left(\frac{q(x)}{p(x)} - 1 \right) = 0 ,$$

so that the relative entropy is always nonnegative, and equals zero if and only if $P = Q$. This simple fact has some interesting consequences. For example, if $\mathcal{X}$ is a finite set with N elements and X is a random variable with distribution P and we take Q to be the uniform distribution over $\mathcal{X}$ then $D(P\|Q) = \log N - H(X)$ and therefore the entropy of X never exceeds the logarithm of the cardinality of its range.

Consider a pair of random variables X, Y with joint distribution $P_{X,Y}$ and marginal distributions P_X and P_Y. Noting that $D(P_{X,Y}\|P_X \times P_Y) = H(X) - H(X|Y)$, the nonnegativity of the relative entropy implies that $H(X) \geq H(X|Y)$, that is, conditioning reduces entropy. It is similarly easy to see that this fact remains true for conditional entropies as well, that is,

$$H(X|Y) \geq H(X|Y, Z) .$$

Now we may prove the following inequality of Han [35]

Theorem 9 HAN'S INEQUALITY. *Let $X_1, \ldots, X_n$ be discrete random variables. Then*

$$H(X_1, \ldots, X_n) \leq \frac{1}{n-1} \sum_{i=1}^{n} H(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$$

Proof. For any $i = 1, \ldots, n$, by the definition of the conditional entropy and the fact that conditioning reduces entropy,

$$\begin{aligned}
&H(X_1, \ldots, X_n) \\
&\quad = H(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) + H(X_i|X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) \\
&\quad \leq H(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) + H(X_i|X_1, \ldots, X_{i-1}) \qquad i = 1, \ldots, n .
\end{aligned}$$

Summing these n inequalities and using the chain rule for entropy, we get

$$nH(X_1, \ldots, X_n) \leq \sum_{i=1}^{n} H(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) + H(X_1, \ldots, X_n)$$

which is what we wanted to prove. □

We finish this section by an inequality which may be regarded as a version of Han's inequality for relative entropies. As it was pointed out by

Massart [55], this inequality may be used to prove the key tensorization inequality of the next section.

To this end, let $\mathcal{X}$ be a countable set, and let P and Q be probability distributions on $\mathcal{X}^n$ such that $P = P_1 \times \cdots \times P_n$ is a product measure. We denote the elements of $\mathcal{X}^n$ by $x_1^n = (x_1, \ldots, x_n)$ and write $x^{(i)} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$ for the $(n-1)$-vector obtained by leaving out the i-th component of x_1^n. Denote by $Q^{(i)}$ and $P^{(i)}$ the marginal distributions of x_1^n according to Q and P, that is,

$$Q^{(i)}(x) = \sum_{x \in \mathcal{X}} Q(x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_n)$$

and

$$\begin{aligned} P^{(i)}(x) &= \sum_{x \in \mathcal{X}} P(x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_n) \\ &= \sum_{x \in \mathcal{X}} P_1(x_1) \cdots P_{i-1}(x_{i-1}) P_i(x) P_{i+1}(x_{i+1}) \cdots P_n(x_n) \ . \end{aligned}$$

Then we have the following.

Theorem 10 HAN'S INEQUALITY FOR RELATIVE ENTROPIES.

$$D(Q\|P) \geq \frac{1}{n-1} \sum_{i=1}^{n} D(Q^{(i)}\|P^{(i)})$$

or equivalently,

$$D(Q\|P) \leq \sum_{i=1}^{n} \left(D(Q\|P) - D(Q^{(i)}\|P^{(i)}) \right) \ .$$

Proof. The statement is a straightforward consequence of Han's inequality. Indeed, Han's inequality states that

$$\sum_{x_1^n \in \mathcal{X}^n} Q(x_1^n) \log Q(x_1^n) \geq \frac{1}{n-1} \sum_{i=1}^{n} \sum_{x^{(i)} \in \mathcal{X}^{n-1}} Q^{(i)}(x^{(i)}) \log Q^{(i)}(x^{(i)}) \ .$$

Since

$$D(Q\|P) = \sum_{x_1^n \in \mathcal{X}^n} Q(x_1^n) \log Q(x_1^n) - \sum_{x_1^n \in \mathcal{X}^n} Q(x_1^n) \log P(x_1^n)$$

and

$$D(Q^{(i)}\|P^{(i)}) = \sum_{x^{(i)}\in\mathcal{X}^{n-1}} \left(Q^{(i)}(x^{(i)})\log Q^{(i)}(x^{(i)}) - Q^{(i)}(x^{(i)})\log P^{(i)}(x^{(i)})\right) \ ,$$

it suffices to show that

$$\sum_{x_1^n\in\mathcal{X}^n} Q(x_1^n)\log P(x_1^n) = \frac{1}{n-1}\sum_{i=1}^{n}\sum_{x^{(i)}\in\mathcal{X}^{n-1}} Q^{(i)}(x^{(i)})\log P^{(i)}(x^{(i)}) \ .$$

This may be seen easily by noting that by the product property of P, we have $P(x_1^n) = P^{(i)}(x^{(i)})P_i(x_i)$ for all i, and also $P(x_1^n) = \prod_{i=1}^{n} P_i(x_i)$, and therefore

$$\begin{aligned}
\sum_{x_1^n\in\mathcal{X}^n} Q(x_1^n)\log P(x_1^n) &= \frac{1}{n}\sum_{i=1}^{n}\sum_{x_1^n\in\mathcal{X}^n} Q(x_1^n)\left(\log P^{(i)}(x^{(i)}) + \log P_i(x_i)\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{x_1^n\in\mathcal{X}^n} Q(x_1^n)\log P^{(i)}(x^{(i)}) + \frac{1}{n}Q(x_1^n)\log P(x_i^n) \ .
\end{aligned}$$

Rearranging, we obtain

$$\begin{aligned}
\sum_{x_1^n\in\mathcal{X}^n} Q(x_1^n)\log P(x_1^n) &= \frac{1}{n-1}\sum_{i=1}^{n}\sum_{x_1^n\in\mathcal{X}^n} Q(x_1^n)\log P^{(i)}(x^{(i)}) \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\sum_{x^{(i)}\in\mathcal{X}^{n-1}} Q^{(i)}(x^{(i)})\log P^{(i)}(x^{(i)})
\end{aligned}$$

where we used the defining property of $Q^{(i)}$. $\square$

5.2 Tensorization of the entropy

We are now prepared to prove the main exponential concentration inequalities of these notes. Just as in Section 4, we let $X_1,\ldots,X_n$ be independent random variables, and investigate concentration properties of $Z = g(X_1,\ldots,X_n)$. The basis of Ledoux's entropy method is a powerful extension of Theorem 6. Note that Theorem 6 may be rewritten as

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

or, putting $\phi(x) = x^2$,

$$\mathbb{E}\phi(Z) - \phi(\mathbb{E}Z) \leq \sum_{i=1}^{n} \mathbb{E}\left[\mathbb{E}_i\phi(Z) - \phi(\mathbb{E}_i(Z))\right] .$$

As it turns out, this inequality remains true for a large class of convex functions ϕ, see Beckner [6], Latała and Oleszkiewicz [41], Ledoux [43], and Chafaï [16]. The case of interest in our case is when $\phi(x) = x \log x$. In this case, as seen in the proof below, the left-hand side of the inequality may be written as the relative entropy between the distribution induced by Z on $\mathcal{X}^n$ and the distribution of X_1^n. Hence the name "tensorization inequality of the entropy", (see, e.g., Ledoux [43]).

Theorem 11 *Let $\phi(x) = x \log x$ for $x > 0$. Let $X_1 \ldots, X_n$ be independent random variables taking values in $\mathcal{X}$ and let f be a positive-valued function on $\mathcal{X}^n$. Letting $Y = f(X_1, \ldots, X_n)$, we have*

$$\mathbb{E}\phi(Y) - \phi(\mathbb{E}Y) \leq \sum_{i=1}^{n} \mathbb{E}\left[\mathbb{E}_i\phi(Y) - \phi(\mathbb{E}_i(Y))\right] .$$

Proof. We only prove the statement for discrete random variables $X_1 \ldots, X_n$. The extension to the general case is technical but straightforward. The theorem is a direct consequence of Han's inequality for relative entropies. First note that if the inequality is true for a random variable Y then it is also true for cY where c is a positive constant. Hence we may assume that $\mathbb{E}Y = 1$. Now define the probability measure Q on $\mathcal{X}^n$ by

$$Q(x_1^n) = f(x_1^n)P(x_1^n)$$

where P denotes the distribution of $X_1^n = X_1, \ldots, X_n$. Then clearly,

$$\mathbb{E}\phi(Y) - \phi(\mathbb{E}Y) = \mathbb{E}[Y \log Y] = D(Q\|P)$$

which, by Theorem 10, does not exceed $\sum_{i=1}^{n}\left(D(Q\|P) - D(Q^{(i)}\|P^{(i)})\right)$. However, straightforward calculation shows that

$$\sum_{i=1}^{n}\left(D(Q\|P) - D(Q^{(i)}\|P^{(i)})\right) = \sum_{i=1}^{n} \mathbb{E}\left[\mathbb{E}_i\phi(Y) - \phi(\mathbb{E}_i(Y))\right]$$

and the statement follows. □

The main idea in Ledoux's entropy method for proving concentration inequalities is to apply Theorem 11 to the positive random variable $Y = e^{sZ}$. Then, denoting the moment generating function of Z by $F(s) = \mathbb{E}[e^{sZ}]$, the left-hand side of the inequality in Theorem 11 becomes

$$s\mathbb{E}\left[Ze^{sZ}\right] - \mathbb{E}\left[e^{sZ}\right]\log\mathbb{E}\left[e^{sZ}\right] = sF'(s) - F(s)\log F(s) \ .$$

Our strategy, then is to derive upper bounds for the derivative of $F(s)$ and derive tail bounds via Chernoff's bounding. To do this in a convenient way, we need some further bounds for the right-hand side of the inequality in Theorem 11. This is the purpose of the next section.

5.3 Logarithmic Sobolev inequalities

Recall from Section 4 that we denote $Z_i = g_i(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$ where g_i is some function over $\mathcal{X}^{n-1}$. Below we further develop the right-hand side of Theorem 11 to obtain important inequalities which serve as the basis in deriving exponential concentration inequalities. These inequalities are closely related to the so-called *logarithmic Sobolev inequalities* of analysis, see Ledoux [43, 44, 45], Massart [54].

First we need the following technical lemma:

Lemma 2 *Let Y denote a positive random variable. Then for any $u > 0$,*

$$\mathbb{E}[Y\log Y] - (\mathbb{E}Y)\log(\mathbb{E}Y) \le \mathbb{E}[Y\log Y - Y\log u - (Y - u)] \ .$$

Proof. As for any $x > 0$, $\log x \le x - 1$, we have

$$\log\frac{u}{\mathbb{E}Y} \le \frac{u}{\mathbb{E}Y} - 1 \ ,$$

hence

$$\mathbb{E}Y\log\frac{u}{\mathbb{E}Y} \le u - \mathbb{E}Y$$

which is equivalent to the statement. □

Theorem 12 A LOGARITHMIC SOBOLEV INEQUALITY. *Denote* $\psi(x) = e^x - x - 1$*. Then*

$$s\mathbb{E}\left[Ze^{sZ}\right] - \mathbb{E}\left[e^{sZ}\right]\log\mathbb{E}\left[e^{sZ}\right] \le \sum_{i=1}^{n}\mathbb{E}\left[e^{sZ}\psi\left(-s(Z - Z_i)\right)\right].$$

Proof. We bound each term on the right-hand side of Theorem 11. Note that Lemma 2 implies that if Y_i is a positive function of $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n$, then

$$\mathbb{E}_i(Y\log Y) - \mathbb{E}_i(Y)\log\mathbb{E}_i(Y) \le \mathbb{E}_i\left[Y(\log Y - \log Y_i) - (Y - Y_i)\right]$$

Applying the above inequality to the variables $Y = e^{sZ}$ and $Y_i = e^{sZ_i}$, one gets

$$\mathbb{E}_i(Y\log Y) - \mathbb{E}_i(Y)\log\mathbb{E}_i(Y) \le \mathbb{E}_i\left[e^{sZ}\psi(-s(Z - Z^{(i)}))\right]$$

and the proof is completed by Theorem 11. $\square$

The following symmetrized version, due to Massart [54], will also be useful. Recall that $Z_i' = g(X_1, \ldots, X_i', \ldots, X_n)$ where the X_i' are independent copies of the X_i.

Theorem 13 SYMMETRIZED LOGARITHMIC SOBOLEV INEQUALITY. *If* ψ *is defined as in Theorem 12 then*

$$s\mathbb{E}\left[Ze^{sZ}\right] - \mathbb{E}\left[e^{sZ}\right]\log\mathbb{E}\left[e^{sZ}\right] \le \sum_{i=1}^{n}\mathbb{E}\left[e^{sZ}\psi\left(-s(Z - Z_i')\right)\right].$$

Moreover, denote $\tau(x) = x(e^x - 1)$*. Then for all* $s \in \mathbb{R}$*,*

$$\begin{aligned}
s\mathbb{E}\left[Ze^{sZ}\right] - \mathbb{E}\left[e^{sZ}\right]\log\mathbb{E}\left[e^{sZ}\right] &\le \sum_{i=1}^{n}\mathbb{E}\left[e^{sZ}\tau(-s(Z - Z_i'))\mathbb{1}_{Z>Z_i'}\right], \\
s\mathbb{E}\left[Ze^{sZ}\right] - \mathbb{E}\left[e^{sZ}\right]\log\mathbb{E}\left[e^{sZ}\right] &\le \sum_{i=1}^{n}\mathbb{E}\left[e^{sZ}\tau(s(Z_i' - Z))\mathbb{1}_{Z<Z_i'}\right].
\end{aligned}$$

Proof. The first inequality is proved exactly as Theorem 12, just by noting that, just like Z_i, Z_i' is also independent of X_i. To prove the second and third inequalities, write

$$e^{sZ}\psi\left(-s(Z-Z_i')\right) = e^{sZ}\psi\left(-s(Z-Z_i')\right)\mathbb{1}_{Z>Z_i'} + e^{sZ}\psi\left(s(Z_i'-Z)\right)\mathbb{1}_{Z<Z_i'} \ .$$

By symmetry, the conditional expectation of the second term may be written as

$$\begin{aligned}\mathbb{E}_i\left[e^{sZ}\psi\left(s(Z_i'-Z)\right)\mathbb{1}_{Z<Z_i'}\right] &= \mathbb{E}_i\left[e^{sZ_i'}\psi\left(s(Z-Z_i')\right)\mathbb{1}_{Z>Z_i'}\right] \\ &= \mathbb{E}_i\left[e^{sZ}e^{-s(Z-Z_i')}\psi\left(s(Z-Z_i')\right)\mathbb{1}_{Z>Z_i'}\right] \ .\end{aligned}$$

Summarizing, we have

$$\begin{aligned}&\mathbb{E}\left[e^{sZ}\psi\left(-s(Z-Z_i')\right)\right] \\ &= \mathbb{E}_i\left[\left(\psi\left(-s(Z-Z_i')\right) + e^{-s(Z-Z_i')}\psi\left(s(Z-Z_i')\right)\right)e^{sZ}\mathbb{1}_{Z>Z_i'}\right] \ .\end{aligned}$$

The second inequality of the theorem follows simply by noting that $\psi(x) + e^x\psi(-x) = x(e^x-1) = \tau(x)$. The last inequality follows similarly. $\square$

5.4 First example: bounded differences and more

The purpose of this section is to illustrate how the logarithmic Sobolev inequalities shown in the previous section may be used to obtain powerful exponential concentration inequalities. The first result is rather easy to obtain, yet it turns out to be very useful. Also, its proof prototypical, in the sense that it shows, in a transparent way, the main ideas.

Theorem 14 *Assume that there exists a positive constant* C *such that, almost surely,*

$$\sum_{i=1}^{n}(Z-Z_i')^2 \le C \ .$$

Then for all $t > 0$*,*

$$\mathbb{P}\left[|Z-\mathbb{E}Z| > t\right] \le 2e^{-t^2/4C} \ .$$

Proof. Observe that for $x > 0$, $\tau(-x) \leq x^2$, and therefore, for any $s > 0$, Theorem 13 implies

$$\begin{aligned} s\mathbb{E}\left[Ze^{sZ}\right] - \mathbb{E}\left[e^{sZ}\right] \log \mathbb{E}\left[e^{sZ}\right] &\leq \mathbb{E}\left[e^{sZ} \sum_{i=1}^{n} s^2 (Z - Z_i')^2 \mathbb{1}_{Z > Z_i'}\right] \\ &\leq s^2 \mathbb{E}\left[e^{sZ} \sum_{i=1}^{n} (Z - Z_i')^2\right] \\ &\leq s^2 C \mathbb{E}\left[e^{sZ}\right] , \end{aligned}$$

where at the last step we used the assumption of the theorem. Now denoting the moment generating function of Z by $F(s) = \mathbb{E}\left[e^{sZ}\right]$, the above inequality may be re-written as

$$sF'(s) - F(s) \log F(s) \leq Cs^2 F(s) .$$

After dividing both sides by $s^2 F(s)$, we observe that the left-hand side is just the derivative of $H(s) = s^{-1} \log F(s)$, that is, we obtain the inequality

$$H'(s) \leq C .$$

By l'Hospital's rule we note that $\lim_{s \to 0} H(s) = F'(0)/F(0) = \mathbb{E}Z$, so by integrating the above inequality, we get $H(s) \leq \mathbb{E}Z + sC$, or in other words,

$$F(s) \leq e^{s\mathbb{E}Z + s^2 C} .$$

Now by Markov's inequality,

$$\mathbb{P}\left[Z > \mathbb{E}Z + t\right] \leq F(s) e^{-s\mathbb{E}Z - st} \leq e^{s^2 C - st} .$$

Choosing $s = t/2C$, the upper bound becomes $e^{-t^2/4C}$. Replace Z by $-Z$ to obtain the same upper bound for $\mathbb{P}\left[Z < \mathbb{E}Z - t\right]$. □

An immediate corollary of Theorem 14 is a subgaussian tail inequality for functions of bounded differences.

Corollary 4 BOUNDED DIFFERENCES INEQUALITY. *Assume the function* g *satisfies the bounded differences assumption with constants* $c_1, \ldots, c_n$, *then*

$$\mathbb{P}\left[|Z - \mathbb{E}Z| > t\right] \leq 2e^{-t^2/4C}$$

where $C = \sum_{i=1}^{n} c_i^2$.

We remark here that the constant appearing in this corollary may be improved. Indeed, using the martingale method, McDiarmid [56] showed that under the conditions of Corollary 4,

$$\mathbb{P}\left[|Z - \mathbb{E}Z| > t\right] \leq 2e^{-2t^2/C}$$

(see the exercises). Thus, we have been able to extend Corollary 1 to an exponential concentration inequality. Note that by combining the variance bound of Corollary 1 with Chebyshev's inequality, we only obtained

$$\mathbb{P}\left[|Z - \mathbb{E}Z| > t\right] \leq \frac{C}{2t^2}$$

and therefore the improvement is essential. Thus the applications of Corollary 1 in all the examples shown in Section 4.1 are now improved in an essential way without further work.

However, Theorem 14 is much stronger than Corollary 4. To understand why, just observe that the conditions of Theorem 14 do not require that g has bounded differences. All that's required is that

$$\sup_{\substack{x_1,\ldots,x_n,\\ x_1',\ldots,x_n' \in \mathcal{X}}} \sum_{i=1}^{n} |g(x_1,\ldots,x_n) - g(x_1,\ldots,x_{i-1},x_i',x_{i+1},\ldots,x_n)|^2 \leq \sum_{i=1}^{n} c_i^2 \ ,$$

an obviously much milder requirement.

5.5 Exponential inequalities for self-bounding functions

In this section we prove exponential concentration inequalities for self-bounding functions discussed in Section 4.2. Recall that a variant of the Efron-Stein inequality (Theorem 2) implies that for self-bounding functions $\mathrm{Var}(Z) \leq \mathbb{E}(Z)$. Based on the logarithmic Sobolev inequality of Theorem 12 we may now obtain exponential concentration bounds. The theorem appears in Boucheron, Lugosi, and Massart [12] and builds on techniques developed by Massart [54].

Recall the definition of following two functions that we have already seen in Bennett's inequality and in the logarithmic Sobolev inequalities above:

$$\begin{aligned} h(u) &= (1+u)\log(1+u) - u \quad (u \geq -1), \\ \text{and} \quad \psi(v) &= \sup_{u \geq -1} [uv - h(u)] = e^v - v - 1 \ . \end{aligned}$$

Theorem 15 *Assume that g satisfies the self-bounding property. Then for every $s \in \mathbb{R}$,*

$$\log \mathbb{E}\left[e^{s(Z-\mathbb{E}Z)}\right] \leq \mathbb{E}Z\psi(s) \ .$$

Moreover, for every $t > 0$,

$$\mathbb{P}[Z \geq \mathbb{E}Z + t] \leq \exp\left[-\mathbb{E}Zh\left(\frac{t}{\mathbb{E}Z}\right)\right]$$

and for every $0 < t \leq \mathbb{E}Z$,

$$\mathbb{P}[Z \leq \mathbb{E}Z - t] \leq \exp\left[-\mathbb{E}Zh\left(-\frac{t}{\mathbb{E}Z}\right)\right]$$

By recalling that $h(u) \geq u^2/(2+2u/3)$ for $u \geq 0$ (we have already used this in the proof of Bernstein's inequality) and observing that $h(u) \geq u^2/2$ for $u \leq 0$, we obtain the following immediate corollaries: for every $t > 0$,

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

and for every $0 < t \leq \mathbb{E}Z$,

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

Proof. We apply Lemma 12. Since the function ψ is convex with $\psi(0) = 0$, for any s and any $u \in [0,1]$, $\psi(-su) \leq u\psi(-s)$. Thus, since $Z - Z_i \in [0,1]$, we have that for every s, $\psi(-s(Z-Z_i)) \leq (Z-Z_i)\psi(-s)$ and therefore, Lemma 12 and the condition $\sum_{i=1}^n (Z - Z_i) \leq Z$ implies that

$$\begin{aligned} s\mathbb{E}\left[Ze^{sZ}\right] - \mathbb{E}\left[e^{sZ}\right]\log\mathbb{E}\left[e^{sZ}\right] &\leq \mathbb{E}\left[\psi(-s)e^{sZ}\sum_{i=1}^n (Z-Z_i)\right] \\ &\leq \psi(-s)\mathbb{E}\left[Ze^{sZ}\right]. \end{aligned}$$

Introduce $\tilde{Z} = Z - \mathbb{E}[Z]$ and define, for any s, $\tilde{F}(s) = \mathbb{E}\left[e^{s\tilde{Z}}\right]$. Then the inequality above becomes

$$[s - \psi(-s)] \frac{\tilde{F}'(s)}{\tilde{F}(s)} - \log \tilde{F}(s) \leq \mathbb{E}Z\psi(-s) \ ,$$

which, writing $G(s) = \log F(s)$, implies

$$\left(1 - e^{-s}\right) G'(s) - G(s) \leq \mathbb{E}Z\psi(-s) \ .$$

Now observe that the function $G_0 = \mathbb{E}Z\psi$ is a solution of the ordinary differential equation $\left(1 - e^{-s}\right) G'(s) - G(s) = \mathbb{E}Z\psi(-s)$. We want to show that $G \leq G_0$. In fact, if $G_1 = G - G_0$, then

$$\left(1 - e^{-s}\right) G_1'(s) - G_1(s) \leq 0. \tag{3}$$

Hence, defining $\tilde{G}(s) = G_1(s) / (e^s - 1)$, we have

$$\left(1 - e^{-s}\right)(e^s - 1)\, \tilde{G}'(s) \leq 0.$$

Hence $\tilde{G}'$ is non-positive and therefore $\tilde{G}$ is non-increasing. Now, since $\tilde{Z}$ is centered $G_1'(0) = 0$. Using the fact that $s(e^s - 1)^{-1}$ tends to 1 as s goes to 0, we conclude that $\tilde{G}(s)$ tends to 0 as s goes to 0. This shows that $\tilde{G}$ is non-positive on $(0, \infty)$ and non-negative over $(-\infty, 0)$, hence G_1 is everywhere non-positive, therefore $G \leq G_0$ and we have proved the first inequality of the theorem. The proof of inequalities for the tail probabilities may be completed by Chernoff's bounding:

$$\mathbb{P}[Z - \mathbb{E}[Z] \geq t] \leq \exp\left[-\sup_{s>0}\left(ts - \mathbb{E}Z\psi(s)\right)\right]$$

and

$$\mathbb{P}[Z - \mathbb{E}[Z] \leq -t] \leq \exp\left[-\sup_{s<0}\left(-ts - \mathbb{E}Z\psi(s)\right)\right].$$

The proof is now completed by using the easy-to-check (and well-known) relations

$$\begin{aligned} \sup_{s>0}\left[ts - \mathbb{E}Z\psi(s)\right] &= \mathbb{E}Zh(t/\mathbb{E}Z) \quad \text{for } t > 0 \\ \sup_{s<0}\left[-ts - \mathbb{E}Z\psi(s)\right] &= \mathbb{E}Zh(-t/\mathbb{E}Z) \quad \text{for } 0 < t \leq \mathbb{E}Z. \end{aligned}$$

□

5.6 Combinatorial entropies

Theorems 2 and 15 provide concentration inequalities for functions having the self-bounding property. In Section 4.2 several examples of such functions are discussed. The purpose of this section is to show a whole new class of self-bounding functions that we call *combinatorial entropies*.

Example. VC ENTROPY. In this first example we consider the so-called Vapnik-Chervonenkis (or VC) entropy, a quantity closely related to the VC dimension discussed in Section 4.2. Let $\mathcal{A}$ be an arbitrary collection of subsets of $\mathcal{X}$, and let $x_1^n = (x_1, \ldots, x_n)$ be a vector of n points of $\mathcal{X}$. Recall that the *shatter coefficient* is defined as the size of the trace of $\mathcal{A}$ on x_1^n, that is,

$$T(x_1^n) = |\mathrm{tr}(x_1^n)| = |\{A \cap \{x_1, \ldots, x_n\} : A \in \mathcal{A}\}| \ .$$

The VC *entropy* is defined as the logarithm of the shatter coefficient, that is,

$$h(x_1^n) = \log_2 T(x_1^n) \ .$$

Lemma 3 *The* VC *entropy has the self-bounding property.*

Proof. We need to show that there exists a function h' of $n-1$ variables such that for all $i = 1, \ldots, n$, writing $x^{(i)} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$, $0 \le h(x_1^n) - h'(x^{(i)}) \le 1$ and

$$\sum_{i=1}^{n} \left(h(x_1^n) - h'(x^{(i)}) \right) \le h(x_1^n).$$

We define h' the natural way, that is, as the entropy based on the $n-1$ points in its arguments. Then clearly, for any i, $h'(x^{(i)}) \le h(x_1^n)$, and the difference cannot be more than one. The nontrivial part of the proof is to show the second property. We do this using Han's inequality (Theorem 9).

Consider the uniform distribution over the set $\mathrm{tr}(x_1^n)$. This defines a random vector $Y = (Y_1, \ldots, Y_n) \in \mathcal{Y}^n$. Then clearly,

$$h(x_1^n) = \log_2 |\mathrm{tr}(x_1^n)(x)| = \frac{1}{\ln 2} H(Y_1, \ldots, Y_n)$$

where $H(Y_1, \ldots, Y_n)$ is the (joint) entropy of $Y_1, \ldots, Y_n$. Since the uniform distribution maximizes the entropy, we also have, for all $i \leq n$, that

$$h'(x^{(i)}) \geq \frac{1}{\ln 2} H(Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_n).$$

Since by Han's inequality

$$H(Y_1, \ldots, Y_n) \leq \frac{1}{n-1} \sum_{i=1}^{n} H(Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_n)|,$$

we have

$$\sum_{i=1}^{n} \left(h(x_1^n) - h'(x^{(i)}) \right) \leq h(x_1^n)$$

as desired. □

The above lemma, together with Theorems 2 and 14 immediately imply the following:

Corollary 5 *Let $X_1, \ldots, X_n$ be independent random variables taking their values in $\mathcal{X}$ and let $Z = h(X_1^n)$ denote the random* VC *entropy. Then* $\mathrm{Var}(Z) \leq \mathbb{E}[Z]$, *for* $t > 0$

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

and for every $0 < t \leq \mathbb{E}Z$,

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

Moreover, for the random shatter coefficient $T(X_1^n)$, *we have*

$$\mathbb{E} \log_2 T(X_1^n) \leq \log_2 \mathbb{E} T(X_1^n) \leq \log_2 e \mathbb{E} \log_2 T(X_1^n) .$$

Note that the left-hand side of the last statement follows from Jensen's inequality, while the right-hand side by taking $s = \ln 2$ in the first inequality of Theorem 15. This last statement shows that the expected VC entropy $\mathbb{E} \log_2 T(X_1^n)$ and the *annealed* VC entropy are tightly connected, regardless of the class of sets $\mathcal{A}$ and the distribution of the X_i's. We note here that this fact answers, in a positive way, an open question raised by Vapnik [77, pages 53–54]: the empirical risk minimization procedure is *non-trivially consistent* and *rapidly convergent* if and only if the annealed entropy rate $(1/n)\log_2 \mathbb{E}[T(X)]$ converges to zero. For the definitions and discussion we refer to [77].

The proof of concentration of the VC entropy may be generalized, in a straightforward way, to a class of functions we call *combinatorial entropies* defined as follows.

Let $x_1^n = (x_1, \ldots, x_n)$ be an n-vector of elements with $x_i \in \mathcal{X}_i$ to which we associate a set $\mathrm{tr}(x_1^n) \subset \mathcal{Y}^n$ of n-vectors whose components are elements of a possibly different set $\mathcal{Y}$. We assume that for each $x \in \mathcal{X}^n$ and $i \leq n$, the set $\mathrm{tr}(x^{(i)}) = \mathrm{tr}(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$ is the projection of $\mathrm{tr}(x_1^n)$ along the i^{th} coordinate, that is,

$$\mathrm{tr}(x^{(i)}) = \Big\{ y^{(i)} = (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n) \in \mathcal{Y}^{n-1} : \\ \exists y_i \in \mathcal{Y} \text{ such that } (y_1, \ldots, y_n) \in \mathrm{tr}(x_1^n) \Big\}.$$

The associated combinatorial entropy is $h(x_1^n) = \log_b |\mathrm{tr}(x_1^n)|$ where b is an arbitrary positive number.

Just like in the case of VC entropy, combinatorial entropies may be shown to have the self-bounding property. (The details are left as an exercise.) Then we immediately obtain the following generalization:

Theorem 16 *Assume that $h(x_1^n) = \log_b |tr(x_1^n)|$ is a combinatorial entropy such that for all $x \in \mathcal{X}^n$ and $i \leq n$,*

$$h(x_1^n) - h(x^{(i)}) \leq 1 .$$

If $X_1^n = (X_1, \ldots, X_n)$ is a vector of n independent random variables taking values in $\mathcal{X}$, then the random combinatorial entropy $Z = h(X_1^n)$ satisfies

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

and

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

Moreover,

$$\mathbb{E}\left[\log_b |tr(X_1^n)|\right] \leq \log_b \mathbb{E}[|tr(X_1^n)|] \leq \frac{b-1}{\log b}\mathbb{E}\left[\log_b |tr(X_1^n)|\right].$$

Example. INCREASING SUBSEQUENCES. Recall the setup of the example of increasing subsequences of Section 4.2, and let $N(x_1^n)$ denote the number of different increasing subsequences of x_1^n. Observe that $\log_2 N(x_1^n)$ is a combinatorial entropy. This is easy to see by considering $\mathcal{Y} = \{0, 1\}$, and by assigning, to each increasing subsequence $i_1 < i_2 < \cdots < i_m$ of x_1^n, a binary n-vector $y_1^n = (y_1, \ldots, y_n)$ such that $y_j = 1$ if and only if $j = i_k$ for some $k = 1, \ldots, m$ (i.e., the indices appearing in the increasing sequence are marked by 1). Now the conditions of Theorem 16 are obviously met, and therefore $Z = \log_2 N(X_1^n)$ satisfies all three inequalities of Theorem 16. This result significantly improves a concentration inequality obtained by Frieze [31] for $\log_2 N(X_1^n)$.

5.7 Variations on the theme

In this section we show how the techniques of the entropy method for proving concentration inequalities may be used in various situations not considered so far. The versions differ in the assumptions on how $\sum_{i=1}^n (Z - Z_i')^2$ is controlled by different functions of Z. For various other versions with applications we refer to Boucheron, Lugosi, and Massart [13]. In all cases the upper bound is roughly of the form e^{-t^2/σ^2} where σ^2 is the corresponding Efron-Stein upper bound on $\mathrm{Var}(Z)$. The first inequality may be regarded as a generalization of the upper tail inequality in Theorem 15.

Theorem 17 *Assume that there exist positive constants* a *and* b *such that*

$$\sum_{i=1}^n (Z - Z_i')^2 \mathbb{1}_{Z>Z_i'} \leq aZ + b\ .$$

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

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

and for all $t > 0$,

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

Proof. Let $s > 0$. Just like in the first steps of the proof of Theorem 14, we use the fact that for $x > 0$, $\tau(-x) \le x^2$, and therefore, by Theorem 13 we have

$$\begin{aligned} s\mathbb{E}\left[Ze^{sZ}\right] - \mathbb{E}\left[e^{sZ}\right]\log\mathbb{E}\left[e^{sZ}\right] &\le \mathbb{E}\left[e^{sZ}\sum_{i=1}^{n}(Z - Z_i')^2\mathbb{1}_{Z>Z_i'}\right] \\ &\le s^2\left(a\mathbb{E}\left[Ze^{sZ}\right] + b\mathbb{E}\left[e^{sZ}\right]\right) , \end{aligned}$$

where at the last step we used the assumption of theorem.

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

$$sF'(s) - F(s)\log F(s) \le as^2F'(s) + bs^2F(s) .$$

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

$$H'(s) \le a(\log F(s))' + b .$$

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

$$H(s) \le \mathbb{E}Z + a\log F(s) + bs ,$$

or, if $s < 1/a$,

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

proving the first inequality. The inequality for the upper tail now follows by Markov's inequality and Exercise 17. $\square$

There is a subtle difference between upper and lower tail bounds. Bounds for the lower tail $\mathbb{P}\{Z < \mathbb{E}Z - t\}$ may be easily derived, due to the association inequality of Theorem 2, under much more general conditions on $\sum_{i=1}^{n}(Z - Z_i')^2\mathbb{1}_{Z<Z_i'}$ (note the difference between this quantity and $\sum_{i=1}^{n}(Z - Z_i')^2\mathbb{1}_{Z>Z_i'}$ appearing in the theorem above!).

Theorem 18 *Assume that for some nondecreasing function* g,

$$\sum_{i=1}^{n}(Z-Z_i')^2\mathbb{1}_{Z<Z_i'} \le g(Z)\ .$$

Then for all $t>0$,

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

Proof. To prove lower-tail inequalities we obtain upper bounds for $F(s) = \mathbb{E}[\exp(sZ)]$ with $s<0$. By the third inequality of Theorem 13,

$$\begin{aligned}
& s\mathbb{E}\left[Ze^{sZ}\right] - \mathbb{E}\left[e^{sZ}\right]\log\mathbb{E}\left[e^{sZ}\right] \\
\le\ & \sum_{i=1}^{n}\mathbb{E}\left[e^{sZ}\tau(s(Z_i'-Z))\mathbb{1}_{Z<Z_i'}\right] \\
\le\ & \sum_{i=1}^{n}\mathbb{E}\left[e^{sZ}s^2(Z_i'-Z)^2\mathbb{1}_{Z<Z_i'}\right] \\
& \text{(using } s<0 \text{ and that } \tau(-x)\le x^2 \text{ for } x>0\text{)} \\
=\ & s^2\mathbb{E}\left[e^{sZ}\sum_{i=1}^{n}(Z-Z_i')^2\mathbb{1}_{Z<Z_i'}\right] \\
\le\ & s^2\mathbb{E}\left[e^{sZ}g(Z)\right]\ .
\end{aligned}$$

Since $g(Z)$ is a nondecreasing and e^{sZ} is a decreasing function of Z, Chebyshev's association inequality (Theorem 2) implies that

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

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

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

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

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

Markov's inequality and optimizing in s now implies the theorem. $\square$

The next result is useful when one in interested in lower-tail bounds but $\sum_{i=1}^n (Z-Z_i')^2 \mathbb{1}_{Z<Z_i'}$ is difficult to handle. In some cases $\sum_{i=1}^n (Z-Z_i')^2 \mathbb{1}_{Z>Z_i'}$ is easier to bound. In such a situation we need the additional guarantee that $|Z-Z_i'|$ remains bounded. Without loss of generality, we assume that the bound is 1.

Theorem 19 *Assume that there exists a nondecreasing function* g *such that* $\sum_{i=1}^n (Z-Z_i')^2 \mathbb{1}_{Z>Z_i'} \le g(Z)$ *and for any value of* X_1^n *and* X_i', $|Z-Z_i'| \le 1$. *Then for all* $K>0$, $s \in [0, 1/K)$

$$\log \mathbb{E}\Big[\exp(-s(Z-\mathbb{E}[Z]))\Big] \le s^2 \frac{\tau(K)}{K^2} \mathbb{E}[g(Z)] \ ,$$

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

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

Proof. The key observation is that the function $\tau(x)/x^2 = (e^x-1)/x$ is increasing if $x>0$. Choose $K>0$. Thus, for $s \in (-1/K, 0)$, the second inequality of Theorem 13 implies that

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

where at the last step we used the assumption of the theorem.

Just like in the proof of Theorem 18, we bound $\mathbb{E}\left[g(Z)e^{sZ}\right]$ by $\mathbb{E}[g(Z)]\mathbb{E}\left[e^{sZ}\right]$. The rest of the proof is identical to that of Theorem 18. Here we took $K=1$.
$\square$

Exercises

Exercise 15 *Relax the condition of Theorem 14 in the following way. Show that if*

$$\mathbb{E}\left[\sum_{i=1}^{n}(Z-Z_i')^2\mathbb{1}_{Z>Z_i'}\middle|\mathbf{X}\right]\le c$$

then for all $t>0$,

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

and if

$$\mathbb{E}\left[\sum_{i=1}^{n}(Z-Z_i')^2\mathbb{1}_{Z_i'>Z}\middle|\mathbf{X}\right]\le c\ ,$$

then

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

Exercise 16 MCDIARMID'S BOUNDED DIFFERENCES INEQUALITY. *Prove that under the conditions of Corollary 4, the following improvement holds:*

$$\mathbb{P}\left[|Z-\mathbb{E}Z|>t\right]\le 2e^{-2t^2/C}$$

(McDiarmid [56]). Hint: Write Z *as a sum of martingale differences as in the proof of Theorem 6. Use Chernoff's bounding and proceed as in the proof of Hoeffding's inequality, noting that the argument works for sums of martingale differences.*

Exercise 17 *Let* C *and* a *denote two positive real numbers and denote* $h_1(x)=1+x-\sqrt{1+2x}$. *Show that*

$$\sup_{\lambda\in[0,1/a)}\left(\lambda t-\frac{C\lambda^2}{1-a\lambda}\right)=\frac{2C}{a^2}h_1\left(\frac{at}{2C}\right)\ge\frac{t^2}{2\big(2C+at\big)}$$

and that the supremum is attained at

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

Also,

$$\sup_{\lambda\in[0,\infty)}\left(\lambda t-\frac{C\lambda^2}{1+a\lambda}\right)=\frac{2C}{a^2}h_1\left(\frac{-at}{2C}\right)\ge\frac{t^2}{4C}$$

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

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

Exercise 18 *Assume that* $h(x_1^n) = \log_b |tr(x)|$ *is a combinatorial entropy such that for all* $x \in \mathcal{X}^n$ *and* $i \leq n$,

$$h(x_1^n) - h(x^{(i)}) \leq 1$$

Show that h *has the self-bounding property.*

6 Concentration of measure

In this section we address the "isoperimetric" approach to concentration inequalities, promoted and developed, in large part, by Talagrand [73, 74, 75]. First we give an equivalent formulation of the bounded differences inequality (Corollary 4) which shows that any not too small set in a product probability space has the property that the probability of those points whose Hamming distance from the set is much larger than $\sqrt{n}$ is exponentially small. Then, using the full power of Theorem 14, we provide a significant improvement of this concentration-of-measure result, known as Talagrand's *convex distance inequality*.

6.1 Bounded differences inequality revisited

Consider independent random variables $X_1, \ldots, X_n$ taking their values in a (measurable) set $\mathcal{X}$ and denote the vector of these variables by $X_1^n = (X_1, \ldots, X_n)$ taking its value in $\mathcal{X}^n$.

Let $A \subset \mathcal{X}^n$ be an arbitrary (measurable) set and write $\mathbb{P}[A] = \mathbb{P}[X_1^n \in A]$. The *Hamming distance* $d(x_1^n, y_1^n)$ between the vectors $x_1^n, y_1^n \in \mathcal{X}^n$ is defined as the number of coordinates in which x_1^n and y_1^n differ. Introduce

$$d(x_1^n, A) = \min_{y_1^n \in A} d(x_1^n, y_1^n),$$

the Hamming distance between the set A and the point x_1^n. The basic result is the following:

Theorem 20 *For any* $t > 0$*,*

$$\mathbb{P}\left[d(X_1^n, A) \geq t + \sqrt{\frac{n}{2} \log \frac{1}{\mathbb{P}[A]}}\right] \leq e^{-2t^2/n} .$$

Observe that on the right-hand side we have the measure of the complement of the t-*blowup* of the set A, that is, the measure of the set of points whose Hamming distance from A is at least t. If we consider a set with $\mathbb{P}[A] \approx 1/2$, we see something very surprising: the measure of the set of points whose Hamming distance to A is more than $10\sqrt{n}$ is smaller than

e^{-177}! In other words, product measures are concentrated on extremely small sets—hence the name "concentration of measure".

Proof. Observe that the function $g(x_1^n) = d(x_1^n, A)$ cannot change by more than 1 by altering one component of x_1^n, that is, it has the bounded differences property with constants $c_1 = \cdots = c_n = 1$. Thus, by the bounded differences inequality (Theorem 4 with the optimal constants given in Exercise 16),

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

But by taking $t = \mathbb{E}d(X_1^n, A)$, the left-hand side becomes $\mathbb{P}[d(X_1^n, A) \leq 0] = \mathbb{P}[A]$, so the above inequality implies

$$\mathbb{E}[d(X_1^n, A)] \leq \sqrt{\frac{n}{2}\log\frac{1}{\mathbb{P}[A]}}.$$

Then, by using the bounded differences inequality again, we obtain

$$\mathbb{P}\left[d(X_1^n, A) \geq t + \sqrt{\frac{n}{2}\log\frac{1}{\mathbb{P}\{A\}}}\right] \leq e^{-2t^2/n}$$

as desired. $\square$

Observe that the bounded differences inequality may also be derived from the above theorem. Indeed, if we consider a function g on $\mathcal{X}^n$ having the bounded differences property with constants $c_i = 1$ (for simplicity), then we may let $A = \{x_1^n \in \mathcal{X}^n : g(x_1^n) \geq \mathbb{M}[Z]\}$, where $\mathbb{M}[Z]$ denotes the median of the random variable $Z = g(X_1, \ldots, X_n)$. Then clearly $\mathbb{P}[A] \geq 1/2$, so the above theorem implies

$$\mathbb{P}[Z - \mathbb{M}Z \geq t + \sqrt{\frac{n}{2}\log 2}] \leq e^{-2t^2/n}\ .$$

This has the same form as the bounded differences inequality except that the expected value of Z is replaced by its median. This difference is usually negligible, since

$$|\mathbb{E}Z - \mathbb{M}Z| \leq \mathbb{E}|Z - \mathbb{M}Z| = \int_0^\infty \mathbb{P}[|Z - \mathbb{M}Z| \geq t]dt,$$

so whenever the deviation of Z from its mean is small, its expected value must be close to its median (see Exercise 19).

6.2 Convex distance inequality

In a remarkable series of papers (see [75],[73],[74]), Talagrand developed an induction method to prove powerful concentration results in many cases when the bounded differences inequality fails. Perhaps the most widely used of these is the so-called "convex-distance inequality", see also Steele [72], McDiarmid [57] for surveys with several interesting applications. Here we use Theorem 14 to derive a version of the convex distance inequality. For several extensions and variations we refer to Talagrand [75],[73],[74].

To understand Talagrand's inequality, we borrow a simple argument from [57]. First observe that Theorem 20 may be easily generalized by allowing the distance of the point X_1^n from the set A to be measured by a *weighted Hamming distance*

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

where $\alpha = (\alpha_1, \ldots, \alpha_n)$ is a vector of nonnegative numbers. Repeating the argument of the proof of Theorem 20, we obtain, for all α,

$$\mathbb{P}\left[d_\alpha(X_1^n, A) \geq t + \sqrt{\frac{\|\alpha\|^2}{2} \log \frac{1}{\mathbb{P}[A]}} \right] \leq e^{-2t^2/\|\alpha\|^2} ,$$

where $\|\alpha\| = \sqrt{\sum_{i=1}^n \alpha_i^2}$ denotes the euclidean norm of α. Thus, for example, for all vectors α with unit norm $\|\alpha\| = 1$,

$$\mathbb{P}\left[d_\alpha(X_1^n, A) \geq t + \sqrt{\frac{1}{2} \log \frac{1}{\mathbb{P}[A]}} \right] \leq e^{-2t^2} .$$

Thus, denoting $u = \sqrt{\frac{1}{2} \log \frac{1}{\mathbb{P}[A]}}$, for any $t \geq u$,

$$\mathbb{P}\left[d_\alpha(X_1^n, A) \geq t \right] \leq e^{-2(t-u)^2} .$$

On the one hand, if $t \leq \sqrt{-2 \log \mathbb{P}[A]}$, then $\mathbb{P}[A] \leq e^{-t^2/2}$. On the other hand, since $(t-u)^2 \geq t^2/4$ for $t \geq 2u$, for any $t \geq \sqrt{2 \log \frac{1}{\mathbb{P}[A]}}$ the inequality above implies $\mathbb{P}\left[d_\alpha(X_1^n, A) \geq t \right] \leq e^{-t^2/2}$. Thus, for all $t > 0$, we have

$$\sup_{\alpha: \|\alpha\|=1} \mathbb{P}[A] \cdot \mathbb{P}\left[d_\alpha(X_1^n, A) \geq t \right] \leq \sup_{\alpha: \|\alpha\|=1} \min\left(\mathbb{P}[A], \mathbb{P}\left[d_\alpha(X_1^n, A) \geq t \right] \right) \leq e^{-t^2/2} .$$

The main message of Talagrand's inequality is that the above inequality remains true even if the supremum is taken within the probability. To make this statement precise, introduce, for any $x_1^n = (x_1, \dots, x_n) \in \mathcal{X}^n$, the *convex distance* of x_1^n from the set A by

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

The next result is a prototypical result from Talagrand's important paper [73]. For an even stronger concentration-of-measure result we refer to [74].

Theorem 21 CONVEX DISTANCE INEQUALITY. *For any subset* $A \subseteq \mathcal{X}^n$ *with* $\mathbb{P}[X_1^n \in A] \ge 1/2$ *and* $t > 0$,

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

Even though at the first sight it is not obvious how Talagrand's result can be used to prove concentration for general functions g of X_1^n, apparently with relatively little work, the theorem may be converted into very useful inequalities. Talagrand [73], Steele [72], and McDiarmid [57] survey a large variety of applications. Instead of reproducing Talagrand's original proof here we show how Theorem 14 and 19 imply the convex distance inequality. (This proof gives a slightly worse exponent than the one obtained by Talagrand's method stated above.)

Proof. Define the random variable $Z = d_T(X_1^n, A)$. First we observe 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 [69] which states that if $f(x, y)$ denotes 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, where $\mathcal{X}$ is convex and compact, then

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

(We omit the details of checking the conditions of Sion's theorem, see [13].)

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

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

where $X_j^{(i)} = X_j$ if $j \neq i$ and $X_i^{(i)} = X_i'$. Let $\tilde{\nu}$ denote the distribution on A that achieves the infimum in the latter expression. Now we have

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

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 $\sum_{i=1}^n (Z - Z_i')^2 \mathbb{1}_{Z > Z_i'} \leq \sum_i \widehat{\alpha}_i^2 = 1$. Thus by Theorem 14 (more precisely, by its generalization in Exercise 15), for any $t > 0$,

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

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. □

6.3 Examples

In what follows we describe an application of the convex distance inequality for the bin packing discussed in Section 4.1, appearing in Talagrand [73]. Let $g(x_1^n)$ denote the minimum number of bins of size 1 into which the numbers $x_1, \ldots, x_n \in [0,1]$ can be packed. We consider the random variable $Z = g(X_1^n)$ where $X_1, \ldots, X_n$ are independent, taking values in $[0,1]$.

Corollary 6 *Denote $\Sigma = \sqrt{\mathbb{E}\sum_{i=1}^n X_i^2}$. Then for each $t > 0$,*

$$\mathbb{P}[|Z - \mathbb{M}Z| \geq t+1] \leq 8e^{-t^2/(16(2\Sigma^2+t))} .$$

Proof. First observe (and this is the only specific property of g we use in the proof!) that for any $x_1^n, y_1^n \in [0,1]^n$,

$$g(x_1^n) \leq g(y_1^n) + 2\sum_{i:x_i \neq y_i} x_i + 1 .$$

To see this it suffices to show that the x_i for which $x_i \neq y_i$ can be packed into at most $\left\lfloor 2\sum_{i:x_i\neq y_i} x_i \right\rfloor + 1$ bins. For this it enough to find a packing such that at most one bin is less than half full. But such a packing must exist because we can always pack the contents of two half-empty bins into one.

Denoting by $\alpha = \alpha(x_1^n) \in [0,\infty)^n$ the unit vector $x_1^n/\|x_1^n\|$, we clearly have

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

Let a be a positive number and define the set $A_a = \{y_1^n : g(y_1^n) \le a\}$. Then, by the argument above and by the definition of the convex distance, for each $x_1^n \in [0,1]^n$ there exists $y_1^n \in A_a$ such that

$$g(x_1^n) \le g(y_1^n) + 2\sum_{i:x_i \ne y_i} x_i + 1 \le a + 2\|x_1^n\| d_T(x_1^n, A_a) + 1$$

from which we conclude that for each $a > 0$, $Z \le a + 2\|x_1^n\| d_T(X_1^n, A_a) + 1$. Thus, writing $\Sigma = \sqrt{\mathbb{E}\sum_{i=1}^n X_i^2}$ for any $t \ge 0$,

$$\begin{aligned}
&\mathbb{P}[Z \ge a + 1 + t] \\
&\le \mathbb{P}\left[Z \ge a + 1 + t\frac{2\|X_1^n\|}{2\sqrt{2\Sigma^2 + t}}\right] + \mathbb{P}\left[\|X_1^n\| \ge \sqrt{2\Sigma^2 + t}\right] \\
&\le \mathbb{P}\left[d_T(X_1^n, A_a) \ge \frac{t}{2\sqrt{2\Sigma^2 + t}}\right] + e^{-(3/8)(\Sigma^2 + t)}
\end{aligned}$$

where the bound on the second term follows by a simple application of Bernstein's inequality, see Exercise 20.

To obtain the desired inequality, we use the obtained bound with two different choices of a. To derive a bound for the upper tail of Z, we take $a = \mathbb{M}Z$. Then $\mathbb{P}[A_a] \ge 1/2$ and the convex distance inequality yields

$$\mathbb{P}[Z \ge \mathbb{M}Z + 1 + t] \le 2\left(e^{-t^2/(16(2\Sigma^2+t))} + e^{-(3/8)(\Sigma^2+t)}\right) \le 4e^{-t^2/(16(2\Sigma^2+t))} .$$

We obtain a similar inequality in the same way for $\mathbb{P}[Z \le \mathbb{M}Z - 1 - t]$ by taking $a = \mathbb{M}Z - t - 1$. $\square$

Exercises

Exercise 19 *Let X be a random variable with median $\mathbb{M}X$ such that there exist positive constants a and b such that for all $t > 0$,*

$$\mathbb{P}[|X - \mathbb{M}X| > t] \le ae^{-t^2/b} .$$

Show that $|\mathbb{M}X - \mathbb{E}X| \le a\sqrt{b\pi}/2$.

Exercise 20 *Let* $X_1, \dots, X_n$ *be independent random variables taking values is* $[0,1]$. *Show that*

$$\mathbb{P}\left[\sqrt{\sum_{i=1}^{n} X_i^2} \geq \sqrt{2\mathbb{E}\sum_{i=1}^{n} X_i^2 + t}\right] \leq e^{-(3/8)(\mathbb{E}\sum_{i=1}^{n} X_i^2 + t)} .$$

References

[1] R. Ahlswede, P. Gács, and J. Körner. 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] M. Anthony and P. L. Bartlett. *Neural Network Learning: Theoretical Foundations.* Cambridge University Press, Cambridge, 1999.

[3] J. Baik, P. Deift, and K. Johansson. On the distribution of the length of the second row of a Young diagram under Plancherel measure. *Geometric and Functional Analysis*, 10:702–731, 2000.

[4] P. Bartlett, S. Boucheron, and G. Lugosi. Model selection and error estimation. *Machine Learning*, 48:85–113, 2001.

[5] P.L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

[6] W. Beckner. A generalized Poincaré inequality for Gaussian measures. *Proceedings of the American Mathematical Society*, 105:397–400, 1989.

[7] G. Bennett. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57:33–45, 1962.

[8] S.N. Bernstein. *The Theory of Probabilities.* Gastehizdat Publishing House, Moscow, 1946.

[9] A. Blumer, A. Ehrenfeucht, D. Haussler, and M.K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36:929–965, 1989.

[10] S. Bobkov and M. Ledoux. Poincaré's inequalities and Talagrands's concentration phenomenon for the exponential distribution. *Probability Theory and Related Fields*, 107:383–400, 1997.

[11] B. Bollobás and G. Brightwell. The height of a random partial order: Concentration of measure. *Annals of Applied Probability*, 2:1009–1018, 1992.

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

[13] S. Boucheron, G. Lugosi, and P. Massart. Concentration inequalities using the entropy method. *Annals of Probability*, to appear, 2003.

[14] O. Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *C. R. Acad. Sci. Paris*, 334:495–500, 2002.

[15] O. Bousquet. New approaches to statistical learning theory. *Annals of the Institute of Statistical Mathematics*, 2003.

[16] D. Chafaï. On ϕ-entropies and ϕ-Sobolev inequalities. Technical report, arXiv.math.PR/0211103, 2002.

[17] H. Chernoff. A measure of asymptotic efficiency of tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics*, 23:493–507, 1952.

[18] V. Chvátal and D. Sankoff. Longest common subsequences of two random sequences. *Journal of Applied Probability*, 12:306–315, 1975.

[19] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. John Wiley, New York, 1991.

[20] V. Dančík and M. Paterson. Upper bound for the expected. In *Proceedings of STACS'94. Lecture notes in Computer Science, 775*, pages 669–678. Springer, New York, 1994.

[21] J.P. Deken. Some limit results for longest common subsequences. *Discrete Mathematics*, 26:17–31, 1979.

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

[23] L. Devroye. The kernel estimate is relatively stable. *Probability Theory and Related Fields*, 77:521–536, 1988.

[24] L. Devroye. Exponential inequalities in nonparametric estimation. In G. Roussas, editor, *Nonparametric Functional Estimation and Related Topics*, pages 31–44. NATO ASI Series, Kluwer Academic Publishers, Dordrecht, 1991.

[25] L. Devroye and L. Györfi. *Nonparametric Density Estimation: The* L_1 *View*. John Wiley, New York, 1985.

[26] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, New York, 1996.

[27] L. Devroye and G. Lugosi. *Combinatorial Methods in Density Estimation*. Springer-Verlag, New York, 2000.

[28] D. Dubdashi and D. Ranjan. Balls and bins: a study in negative dependence. *Random Structures and Algorithms*, pages 99–124, 1998.

[29] R.M. Dudley. *Uniform Central Limit Theorems*. Cambridge University Press, Cambridge, 1999.

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

[31] A.M. Frieze. On the length of the longest monotone subsequence in a random permutation. *Annals of Applied Probability*, 1:301–305, 1991.

[32] P. Groeneboom. Hydrodynamical methods for analyzing longest increasing subsequences. probabilistic methods in combinatorics and combinatorial optimization. *Journal of Computational and Applied Mathematics*, 142:83–105, 2002.

[33] T. Hagerup and C. Rüb. A guided tour of Chernoff bounds. *Information Processing Letters*, 33:305–308, 1990.

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

[35] T.S. Han. Nonnegative entropy measures of multivariate symmetric correlations. *Information and Control*, 36, 1978.

[36] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

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

[38] R.M. Karp. *Probabilistic Analysis of Algorithms.* Class Notes, University of California, Berkeley, 1988.

[39] V. Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47:1902–1914, 2001.

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

[41] R. Latała and C. Oleszkiewicz. Between Sobolev and Poincaré. In *Geometric Aspects of Functional Analysis, Israel Seminar (GAFA), 1996-2000*, pages 147–168. Springer, 2000. Lecture Notes in Mathematics, 1745.

[42] 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.

[43] M. Ledoux. On Talagrand's deviation inequalities for product measures. *ESAIM: Probability and Statistics*, 1:63–87, 1997. `http://www.emath.fr/ps/`.

[44] M. Ledoux. Concentration of measure and logarithmic sobolev inequalities. In *Séminaire de Probabilités XXXIII. Lecture Notes in Mathematics 1709*, pages 120–216. Springer, 1999.

[45] M. Ledoux. *The concentration of measure phenomenon*. American Mathematical Society, Providence, RI, 2001.

[46] B.F. Logan and L.A. Shepp. A variational problem for Young tableaux. *Advances in Mathematics*, 26:206–222, 1977.

[47] Malwina J. Luczak and Colin McDiarmid. Concentration for locally acting permutations. *Discrete Mathematics*, page to appear, 2003.

[48] G. Lugosi. Pattern classification and learning theory. In L. Györfi, editor, *Principles of Nonparametric Learning*, pages 5–62. Springer, Viena, 2002.

[49] G. Lugosi and M. Wegkamp. Complexity regularization via localized random penalties. *submitted*, 2003.

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

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

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

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

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

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

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

[57] 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.

[58] C. McDiarmid. Concentration for independent permutations. *Combinatorics, Probability, and Computing*, 2:163–178, 2002.

[59] M. Okamoto. Some inequalities relating to the partial sum of binomial probabilities. *Annals of the Institute of Statistical Mathematics*, 10:29–35, 1958.

[60] S. Mendelson P. Bartlett, O. Bousquet. Localized rademacher complexities. In *COLT 2002*. to appear, 2002.

[61] D. Panchenko. A note on Talagrand's concentration inequality. *Electronic Communications in Probability*, 6, 2001.

[62] D. Panchenko. Some extensions of an inequality of Vapnik and Chervonenkis. *Electronic Communications in Probability*, 7, 2002.

[63] D. Panchenko. Symmetrization approach to concentration inequalities for empirical processes. *Annals of Probability*, to appear, 2003.

[64] W. Rhee. A matching problem and subadditive Euclidean functionals. *Annals of Applied Probability*, 3:794–801, 1993.

[65] W. Rhee and M. Talagrand. Martingales, inequalities, and NP-complete problems. *Mathematics of Operations Research*, 12:177–181, 1987.

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

[67] P.-M. Samson. Concentration of measure inequalities for Markov chains and ϕ-mixing processes. *Annals of Probability*, 28:416–461, 2000.

[68] E. Shamir and J. Spencer. Sharp concentration of the chromatic number on random graphs $g_{n,p}$. *Combinatorica*, 7:374–384, 1987.

[69] M. Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8:171–176, 1958.

[70] J.M. Steele. Long common subsequences and the proximity of two random strings. *SIAM Journal of Applied Mathematics*, 42:731–737, 1982.

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

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

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

[74] M. Talagrand. New concentration inequalities in product spaces. *Inventiones Mathematicae*, 126:505–563, 1996.

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

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

[77] V.N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 1995.

[78] V.N. Vapnik. *Statistical Learning Theory*. John Wiley, New York, 1998.

[79] V.N. Vapnik and A.Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16:264–280, 1971.

[80] V.N. Vapnik and A.Ya. Chervonenkis. *Theory of Pattern Recognition*. Nauka, Moscow, 1974. (in Russian); German translation: *Theorie der Zeichenerkennung*, Akademie Verlag, Berlin, 1979.