Concentration Inequalities

Stéphane Boucheron[1], Gábor Lugosi[2], and Olivier Bousquet[3]

[1] Université de Paris-Sud, Laboratoire d'Informatique
Bâtiment 490, F-91405 Orsay Cedex, France
stephane.boucheron@lri.fr
WWW home page: http://www.lri.fr/~bouchero

[2] Department of Economics, Pompeu Fabra University
Ramon Trias Fargas 25-27, 08005 Barcelona, Spain
lugosi@upf.es
WWW home page: http://www.econ.upf.es/~lugosi

[3] Max-Planck Institute for Biological Cybernetics
Spemannstr. 38, D-72076 Tübingen, Germany
olivier.bousquet@m4x.org
WWW home page: http://www.kyb.mpg.de/~bousquet

Abstract. Concentration inequalities deal with deviations of functions of independent random variables from their expectation. In the last decade new tools have been introduced making it possible to establish simple and powerful inequalities. These inequalities are at the heart of the mathematical analysis of various problems in machine learning and made it possible to derive new efficient algorithms. This text attempts to summarize some of the basic tools.

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 Milman and Schechtman [1] and the surveys of McDiarmid [2, 3]), information-theoretic methods (see Alhswede, Gács, and Körner [4], Marton [5, 6, 7], Dembo [8], Massart [9] and Rio [10]), Talagrand's induction method [11, 12, 13] (see also Luczak and McDiarmid [14], McDiarmid [15] and Panchenko [16, 17, 18]), the decoupling method surveyed by de la Peña and Giné [19], and the so-called "entropy method", based on logarithmic Sobolev inequalities, developed by Ledoux [20, 21], see also Bobkov and Ledoux [22], Massart [23], Rio [10], Klein [24], Boucheron, Lugosi, and Massart [25, 26], Bousquet [27, 28], and Boucheron, Bousquet, Lugosi, and Massart [29]. Also, various problem-specific methods have been worked out in random graph theory, see Janson, Łuczak, and Ruciński [30] for a survey.

First of all we recall some of the essential basic tools needed in the rest of these notes. 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

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

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.

Next 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 $\sigma^2 = \frac{1}{n}\sum_{i=1}^n \mathsf{Var}\{X_i\}$,

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

Chernoff's bounding method 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

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

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 [31] which we state without proof.

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

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

This lemma, combined with (1) immediately implies Hoeffding's tail inequality [31]:

Theorem 1. *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*

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

and

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

The theorem above is generally known as *Hoeffding's inequality*. For binomial random variables it was proved by Chernoff [32] and Okamoto [33].

A disadvantage of Hoeffding's inequality is that it ignores information about the variance of the X_i's. The inequalities discussed next provide an improvement in this respect.

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

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

Also, let $\psi(x) = \exp(x) - x - 1$, and observe that $\psi(x) \le x^2/2$ for $x \le 0$ and $\psi(sx) \le x^2\psi(s)$ for $s \ge 0$ and $x \in [0,1]$. Since $e^{sx} = 1 + sx + \psi(sx)$, we may write

$$
\begin{aligned}
\mathbb{E}\left[e^{sX_i}\right] &= 1 + s\mathbb{E}[X_i] + \mathbb{E}[\psi(sX_i)] \\
&= 1 + \mathbb{E}[\psi(sX_i)] \quad (\text{since } \mathbb{E}[X_i] = 0.) \\
&\le 1 + \mathbb{E}[\psi(s(X_i)_+) + \psi(-s(X_i)_-)] \\
&\qquad (\text{where } x_+ = \max(0,x) \text{ and } x_- = \max(0,-x)) \\
&\le 1 + \mathbb{E}[\psi(s(X_i)_+) + \frac{s^2}{2}(X_i)_-^2] \quad (\text{using } \psi(x) \le x^2/2 \text{ for } x \le 0.\ )\ .
\end{aligned}
$$

Now assume that the X_i's are bounded such that $X_i \le 1$. Thus, we have obtained

$$
\mathbb{E}\left[e^{sX_i}\right] \le 1 + \mathbb{E}[\psi(s)(X_i)_+^2 + \frac{s^2}{2}(X_i)_-^2] \le 1 + \psi(s)\mathbb{E}[X_i^2] \le \exp\left(\psi(s)\mathbb{E}[X_i^2]\right)
$$

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

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

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

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

Resubstituting this value, we obtain *Bennett's inequality* [34]:

Theorem 2. BENNETT'S INEQUALITY. *Let $X_1, \ldots, X_n$ be independent real-valued random variables with zero mean, and assume that $X_i \le 1$ with probability one. Let*

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

Then for any $t > 0$,

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

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 [35]:

Theorem 3. BERNSTEIN'S INEQUALITY. *Under the conditions of the previous theorem, for any $\epsilon > 0$,*

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

Bernstein's inequality points out an 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 $e^{-\lambda}$.

2 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 4.

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

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

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

since, for any $i > j$,

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

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

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

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 5. EFRON-STEIN INEQUALITY (EFRON AND STEIN [36], STEELE [37]).

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

Proof. The statement follows by Theorem 4 simply by using (conditionally) the elementary fact that if X and Y are independent and identically distributed random variables, then $\mathsf{Var}(X) = (1/2)\mathbb{E}[(X - Y)^2]$, and therefore

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

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 5 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 4 is obtained by recalling that, for any random variable X, $\mathsf{Var}(X) \leq \mathbb{E}[(X-a)^2]$ for any constant $a \in \mathbb{R}$. Using this fact conditionally, we have, for every $i = 1, \ldots, n$,

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

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 4 we have the following.

Theorem 6.

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

In the next two sections we specialize the Efron-Stein inequality and its variant Theorem 6 to functions which satisfy some simple easy-to-verify properties.

2.1 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 $c_1, \ldots, c_n$,

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

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*

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

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. 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 [38] 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 [39], Vapnik [40], van der Vaart and Wellner [41], Dudley [42].

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 and define $Z = g(X_1, \ldots, X_n) = \sup_{f\in\mathcal{F}} \sum_{j=1}^n f(X_j)$. Assume that the functions $f \in \mathcal{F}$ are such that $\mathbb{E}[f(X_i)] = 0$ and take values in $[-1, 1]$. Let Z_i be defined as

$$Z_i = \sup_{f\in\mathcal{F}} \sum_{j\neq i} f(X_j) .$$

Let $\hat{f}$ be the function achieving the supremum[4] in the definition of Z, that is $Z = \sum_{i=1}^n \hat{f}(X_i)$ and similarly $\hat{f}_i$ be such that $Z_i = \sum_{j\neq i} \hat{f}_i(X_j)$. We have

$$\hat{f}_i(X_i) \le Z - Z_i \le \hat{f}(X_i) ,$$

and thus $\sum_{i=1}^n Z - Z_i \le Z$. As $\hat{f}_i$ and X_i are independent, $\mathbb{E}_i[\hat{f}_i(X_i)] = 0$. On the other hand,

$$\begin{aligned}(Z - Z_i)^2 - \hat{f}_i^2(X_i) &= (Z - Z_i + \hat{f}_i(X_i))(Z - Z_i - \hat{f}_i(X_i)) \\ &\le 2(Z - Z_i + \hat{f}_i(X_i)) .\end{aligned}$$

Summing over all i and taking expectations,

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^n (Z - Z_i)^2\right] &\le \mathbb{E}\left[\sum_{i=1}^n \hat{f}_i^2(X_i) + 2(Z - Z_i) + 2\hat{f}_i(X_i)\right] \\ &\le n \sup_{f\in\mathcal{F}} \mathbb{E}[f^2(X_1)] + 2\mathbb{E}[Z]\end{aligned}$$

[4] If the supremum is not attained the proof can be modified to yield the same result. We omit the details here.

where at the last step we used the facts that $\mathbb{E}[\hat{f}_i(X_i)^2] \leq \sup_{f\in\mathcal{F}} \mathbb{E}[f^2(X_1)]$, $\sum_{i=1}^n (Z - Z_i) \leq Z$, and $\mathbb{E}\hat{f}_i(X_i) = 0$. Thus, by the Efron-Stein inequality

$$\mathsf{Var}(Z) \leq n \sup_{f\in\mathcal{F}} \mathbb{E}[f^2(X_1)] + 2\mathbb{E}[Z]$$

¿From just the bounded differences property we derived $\mathsf{Var}(Z) \leq 2n$. The new bound may be a significant improvement whenever the maximum of $\mathbb{E}f(X_i)^2$ over $f \in \mathcal{F}$ is small. (Note that if the class $\mathcal{F}$ is not too large, $\mathbb{E}Z$ is typically of the order of $\sqrt{n}$.) The exponential tail inequality due to Talagrand [12] extends this variance inequality, and is one of the most important recent results of the theory of empirical processes, see also Ledoux [20], Massart [23], Rio [10], Klein [24], and Bousquet [27, 28].

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 and Panchenko [43], Massart [44], Bartlett, Boucheron, and Lugosi [45], Lugosi and Wegkamp [46], Bousquet [47].

Let $\mathcal{F}$ denote a class of $\{0,1\}$-valued functions on some space $\mathcal{X}$. For simplicity of the exposition we assume that $\mathcal{F}$ is finite. The results remain true for general classes as long as the measurability issues are taken care of. Given an i.i.d. sample $D_n = (\langle X_i, Y_i\rangle)_{i\leq n}$ of n pairs of random variables $\langle X_i, Y_i\rangle$ taking values in $\mathcal{X} \times \{0,1\}$, for each $f \in \mathcal{F}$ we define the empirical 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 $\mathsf{Var}(\widehat{L}) \leq 1/(2n)$. 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 5, 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:

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

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

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

Thus,

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

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

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

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

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

so that we obtain

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

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 [26].

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

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

so without further work we get

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

It is known that for every f, $\sqrt{n}\mathbb{E}g \to \infty$ (see Devroye and Györfi [49]) which implies, by Chebyshev's inequality, that for every $\epsilon > 0$

$$\mathbb{P}\left\{ \left| \frac{Z}{\mathbb{E}Z} - 1 \right| \geq \epsilon \right\} = \mathbb{P}\left\{ |Z - \mathbb{E}Z| \geq \epsilon \mathbb{E}Z \right\} \leq \frac{\mathsf{Var}(Z)}{\epsilon^2 (\mathbb{E}Z)^2} \to 0$$

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 [50], [51]. For more on the behavior of the L_1 error of the kernel density estimate we refer to Devroye and Györfi [49], Devroye and Lugosi [52].

2.2 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 [25], Rio [10], and Bousquet [27, 28]. 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 6 implies

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

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

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 non-negative 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 $\epsilon > 0$,

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

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

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 $\sigma_1, \dots, \sigma_n$ 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

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

where the notation X_1^n is a shorthand for $X_1, \dots, X_n$. 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 [53], Bartlett, Boucheron, and Lugosi [45], Bartlett and Mendelson [54], Bartlett, Bousquet, and Mendelson [55]. 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

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

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 [40], van der Vaart and Wellner [41], Dudley [42].

Configuration functions. An important class of functions satisfying the self-bounding property consists of the so-called *configuration functions* defined by Talagrand [11, section 7]. Our definition, taken from [25] 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 $t \geq 0$,*

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

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 $0 \leq Z - Z_i \leq 1$ 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

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

is also satisfied, which concludes the proof. □

To illustrate the fact that configuration functions appear rather naturally in various applications, we describe a prototypical example:

Example. VC DIMENSION. One of the central quantities in statistical learning theory is the *Vapnik-Chervonenkis dimension*, see Vapnik and Chervonenkis [38, 56], Blumer, Ehrenfeucht, Haussler, and Warmuth [57], Devroye, Györfi, and Lugosi [39], Anthony and Bartlett [58], Vapnik [40], 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

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

The *shatter coefficient*, (or *Vapnik-Chervonenkis growth function*) of $\mathcal{A}$ in x_1^n is $T(x_1^n) = |\operatorname{tr}(x_1^n)|$, 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

$$\mathsf{Var}(D(X_1^n)) \leq \mathbb{E}D(X_1^n) \ .$$

3 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 [2, 3], Rhee and Talagrand [59], Shamir and Spencer [60]) but independently information-theoretic methods were also used with success (see Alhswede, Gács, and Körner [4], Marton [5, 6, 7], Dembo [8], Massart [9], Rio [10], and Samson [61]). Talagrand's induction method [11, 12, 13] 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 [20, 21], see also Bobkov and Ledoux [22], Massart [23], Rio [10], Boucheron, Lugosi, and Massart [25], [26], and Bousquet [27, 28]. 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 4. 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.

3.1 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 [62].

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 $\mathcal{X} \times \mathcal{Y}$ 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 discrete 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 [63]

Theorem 7. HAN'S INEQUALITY. *Let $X_1, \ldots, X_n$ be discrete random variables. Then*

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

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

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

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

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

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 [44], 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 8. HAN'S INEQUALITY FOR RELATIVE ENTROPIES.

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

or equivalently,

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

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 $Q^{(i)}$. □

3.2 Tensorization of the Entropy

We are now prepared to prove the main exponential concentration inequalities of these notes. Just as in Section 2, 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 4. Note that Theorem 4 may be rewritten as

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

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 [64], Latała and Oleszkiewicz [65], Ledoux [20], Boucheron, Bousquet, Lugosi, and Massart [29], and Chafaï [66]. 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 [20]).

Theorem 9. *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 8, does not exceed $\sum_{i=1}^{n}\left(D(Q\|P) - D(Q^{(i)}\|P^{(i)})\right)$. However, straightforward calculation shows that

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

and the statement follows. $\square$

The main idea in Ledoux's entropy method for proving concentration inequalities is to apply Theorem 9 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 9 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 9. This is the purpose of the next section.

3.3 Logarithmic Sobolev Inequalities

Recall from Section 2 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 9 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 [20, 67, 68], Massart [23].

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 10. 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 9. 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 9. □

The following symmetrized version, due to Massart [23], 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 11. SYMMETRIZED LOGARITHMIC SOBOLEV INEQUALITY. *If ψ is defined as in Theorem 10 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 $s \in \mathbb{R}$,

$$s\mathbb{E}\left[Ze^{sZ}\right] - \mathbb{E}\left[e^{sZ}\right] \log \mathbb{E}\left[e^{sZ}\right] \le \sum_{i=1}^{n} \mathbb{E}\left[e^{sZ} \tau(-s(Z - Z'_i)) \mathbb{1}_{Z > Z'_i}\right],$$

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

Proof. The first inequality is proved exactly as Theorem 10, 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. □

3.4 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 is prototypical, in the sense that it shows, in a transparent way, the main ideas.

Theorem 12. *Assume that there exists a positive constant C such that, almost surely,*

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

Then for all $t > 0$,

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

Proof. Observe that for $x > 0$, $\tau(-x) \leq x^2$, and therefore, for any $s > 0$, Theorem 11 implies

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

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

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

After dividing both sides by $s^2F(s)$, 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,

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

Now by Markov's inequality,

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

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

Remark. It is easy to see that the condition of Theorem 12 may be relaxed in the following way: 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

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

An immediate corollary of Theorem 12 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*

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

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 [2] showed that under the conditions of Corollary 4,

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

(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

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

and therefore the improvement is essential. Thus the applications of Corollary 1 in all the examples shown in Section 2.1 are now improved in an essential way without further work.

However, Theorem 12 is much stronger than Corollary 4. To understand why, just observe that the conditions of Theorem 12 do not require that g has bounded differences. All that's required is that

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

an obviously much milder requirement.

3.5 Exponential Inequalities for Self-bounding Functions

In this section we prove exponential concentration inequalities for self-bounding functions discussed in Section 2.2. Recall that a variant of the Efron-Stein inequality (Theorem 2) implies that for self-bounding functions $\mathsf{Var}(Z) \le \mathbb{E}(Z)$. Based on the logarithmic Sobolev inequality of Theorem 10 we may now obtain exponential concentration bounds. The theorem appears in Boucheron, Lugosi, and Massart [25] and builds on techniques developed by Massart [23].

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

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

Theorem 13. *Assume that g satisfies the self-bounding property. Then for every $s \in \mathbb{R}$,*

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

Moreover, for every $t > 0$,

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

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

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

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

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

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

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

Proof. We apply Lemma 10. Since the function ψ is convex with $\psi(0) = 0$, for any s and any $u \in [0,1]$, $\psi(-su) \le u\psi(-s)$. Thus, since $Z - Z_i \in [0,1]$, we have that for every s, $\psi(-s(Z - Z_i)) \le (Z - Z_i)\psi(-s)$ and therefore, Lemma 10 and the condition $\sum_{i=1}^n (Z - Z_i) \le Z$ imply that

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

Introduce $\widetilde{Z} = Z - \mathbb{E}[Z]$ and define, for any s, $\tilde{F}(s) = \mathbb{E}\left[e^{s\widetilde{Z}}\right]$. 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 $(1 - e^{-s}) G'(s) - G(s) = \mathbb{E}Z\psi(-s)$. We want to show that $G \leq G_0$. In fact, if $G_1 = G - G_0$, then

$$\left(1 - e^{-s}\right) G_1'(s) - G_1(s) \leq 0. \tag{2}$$

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 $\widetilde{Z}$ 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:

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

and

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

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

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

□

3.6 VC Entropy

Theorems 2 and 13 provide concentration inequalities for functions having the self-bounding property. In Section 2.2 several examples of such functions are discussed. The purpose of this section is to show that the so-called VC entropy is a self-bounding function.

The Vapnik-Chervonenkis (or VC) entropy is closely related to the VC dimension discussed in Section 2.2. 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 7).

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 $i \le n$, that

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

Since by Han's inequality

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

we have

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

as desired. □

The above lemma, together with Theorems 2 and 12 immediately implies 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* $\mathsf{Var}(Z) \le \mathbb{E}[Z]$*, for* $t > 0$

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

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

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

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

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

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 13. 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 [69, 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 [69].

3.7 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 [26]. 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 $\mathsf{Var}(Z)$. The first inequality may be regarded as a generalization of the upper tail inequality in Theorem 13.

Theorem 14. *Assume that there exist positive constants a and b such that*

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

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

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

and for all $t > 0$*,*

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

Proof. Let $s > 0$. Just like in the first steps of the proof of Theorem 12, we use the fact that for $x > 0$, $\tau(-x) \leq x^2$, and therefore, by Theorem 11 we have

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

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

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

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

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

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

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

or, if $s < 1/a$,

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

proving the first inequality. The inequality for the upper tail now follows by Markov's inequality and the following technical lemma whose proof is left as an exercise. □

Lemma 4. *Let C and a denote two positive real numbers and denote $h_1(x) = 1 + x - \sqrt{1+2x}$. Then*

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

and the supremum is attained at

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

Also,

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

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

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

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 *Chebyshev's association inequality* which states that if X is a real-valued random variable and f is a nonincreasing and g is a nondecreasing function, then

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

Theorem 15. *Assume that for some nondecreasing function g,*

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

Then for all $t > 0$,

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

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 11,

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

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

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

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

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

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

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

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

The next result is useful when one is 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 16. *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 11 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 15, 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 15. Here we took $K = 1$. □

Finally we give, without proof, an inequality (due to Bousquet [28]) for functions satisfying conditions similar but weaker than the self-bounding conditions. This is very useful for suprema of empirical processes for which the non-negativity assumption does not hold.

Theorem 17. *Assume Z satisfies $\sum_{i=1}^n Z - Z_i \le Z$, and there exist random variables Y_i such that for all $i = 1, \ldots, n$, $Y_i \le Z - Z_i \le 1$, $Y_i \le a$ for some $a > 0$ and $\mathbb{E}_i Y_i \ge 0$. Also, let σ^2 be a real number such that*

$$\sigma^2 \ge \frac{1}{n}\sum_{i=1}^n \mathbb{E}_i[Y_i^2] \, .$$

We obtain for all $t > 0$,

$$\mathbb{P}\left\{Z \geq \mathbb{E}Z + t\right\} \leq \exp\left(-vh\left(\frac{t}{v}\right)\right),$$

where $v = (1+a)\mathbb{E}Z + n\sigma^2$.

An important application of the above theorem is the following version of Talagrand's concentration inequality for empirical processes. The constants appearing here were obtained by Bousquet [27].

Corollary 6. *Let $\mathcal{F}$ be a set of functions that satisfy $\mathbb{E}f(X_i) = 0$ and $\sup_{f\in\mathcal{F}} \sup f \leq 1$. We denote*

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

Let σ be a positive real number such that $n\sigma^2 \geq \sum_{i=1}^{n} \sup_{f\in\mathcal{F}} \mathbb{E}[f^2(X_i)]$, then for all $t \geq 0$, we have

$$\mathbb{P}\left\{Z \geq \mathbb{E}Z + t\right\} \leq \exp\left(-vh\left(\frac{t}{v}\right)\right),$$

with $v = n\sigma^2 + 2\mathbb{E}Z$.

References

1. Milman, V., Schechman, G.: Asymptotic theory of finite-dimensional normed spaces. Springer-Verlag, New York (1986)
2. McDiarmid, C.: On the method of bounded differences. In: Surveys in Combinatorics 1989, Cambridge University Press, Cambridge (1989) 148–188
3. McDiarmid, C.: Concentration. In Habib, M., McDiarmid, C., Ramirez-Alfonsin, J., Reed, B., eds.: Probabilistic Methods for Algorithmic Discrete Mathematics, Springer, New York (1998) 195–248
4. Ahlswede, R., Gács, P., Körner, J.: Bounds on conditional probabilities with applications in multi-user communication. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete **34** (1976) 157–177 (correction in 39:353–354,1977).
5. Marton, K.: A simple proof of the blowing-up lemma. IEEE Transactions on Information Theory **32** (1986) 445–446
6. Marton, K.: Bounding $\bar{d}$-distance by informational divergence: a way to prove measure concentration. Annals of Probability **24** (1996) 857–866
7. Marton, K.: A measure concentration inequality for contracting Markov chains. Geometric and Functional Analysis **6** (1996) 556–571 Erratum: 7:609–613, 1997.
8. Dembo, A.: Information inequalities and concentration of measure. Annals of Probability **25** (1997) 927–939
9. Massart, P.: Optimal constants for Hoeffding type inequalities. Technical report, Mathematiques, Université de Paris-Sud, Report 98.86 (1998)
10. Rio, E.: Inégalités de concentration pour les processus empiriques de classes de parties. Probability Theory and Related Fields **119** (2001) 163–175

11. Talagrand, M.: Concentration of measure and isoperimetric inequalities in product spaces. Publications Mathématiques de l'I.H.E.S. **81** (1995) 73–205
12. Talagrand, M.: New concentration inequalities in product spaces. Inventiones Mathematicae **126** (1996) 505–563
13. Talagrand, M.: A new look at independence. Annals of Probability **24** (1996) 1–34 (Special Invited Paper).
14. Luczak, M.J., McDiarmid, C.: Concentration for locally acting permutations. Discrete Mathematics (2003) to appear
15. McDiarmid, C.: Concentration for independent permutations. Combinatorics, Probability, and Computing **2** (2002) 163–178
16. Panchenko, D.: A note on Talagrand's concentration inequality. Electronic Communications in Probability **6** (2001)
17. Panchenko, D.: Some extensions of an inequality of Vapnik and Chervonenkis. Electronic Communications in Probability **7** (2002)
18. Panchenko, D.: Symmetrization approach to concentration inequalities for empirical processes. Annals of Probability **to appear** (2003)
19. de la Peña, V., Giné, E.: Decoupling: from Dependence to Independence. Springer, New York (1999)
20. Ledoux, M.: On Talagrand's deviation inequalities for product measures. ESAIM: Probability and Statistics **1** (1997) 63–87 `http://www.emath.fr/ps/`.
21. Ledoux, M.: Isoperimetry and Gaussian analysis. In Bernard, P., ed.: Lectures on Probability Theory and Statistics, Ecole d'Eté de Probabilités de St-Flour XXIV-1994 (1996) 165–294
22. Bobkov, S., Ledoux, M.: Poincaré's inequalities and Talagrands's concentration phenomenon for the exponential distribution. Probability Theory and Related Fields **107** (1997) 383–400
23. Massart, P.: About the constants in Talagrand's concentration inequalities for empirical processes. Annals of Probability **28** (2000) 863–884
24. Klein, T.: Une inégalité de concentration à gauche pour les processus empiriques. C. R. Math. Acad. Sci. Paris **334** (2002) 501–504
25. Boucheron, S., Lugosi, G., Massart, P.: A sharp concentration inequality with applications. Random Structures and Algorithms **16** (2000) 277–292
26. Boucheron, S., Lugosi, G., Massart, P.: Concentration inequalities using the entropy method. The Annals of Probability **31** (2003) 1583–1614
27. Bousquet, O.: A Bennett concentration inequality and its application to suprema of empirical processes. C. R. Acad. Sci. Paris **334** (2002) 495–500
28. Bousquet, O.: Concentration inequalities for sub-additive functions using the entropy method. In Giné, E., C.H., Nualart, D., eds.: Stochastic Inequalities and Applications. Volume 56 of Progress in Probability. Birkhauser (2003) 213–247
29. Boucheron, S., Bousquet, O., Lugosi, G., Massart, P.: Moment inequalities for functions of independent random variables. The Annals of Probability (2004) to appear.
30. Janson, S., Łuczak, T., Ruciński, A.: Random graphs. John Wiley, New York (2000)
31. Hoeffding, W.: Probability inequalities for sums of bounded random variables. Journal of the American Statistical Association **58** (1963) 13–30
32. Chernoff, H.: A measure of asymptotic efficiency of tests of a hypothesis based on the sum of observations. Annals of Mathematical Statistics **23** (1952) 493–507
33. Okamoto, M.: Some inequalities relating to the partial sum of binomial probabilities. Annals of the Institute of Statistical Mathematics **10** (1958) 29–35

34. Bennett, G.: Probability inequalities for the sum of independent random variables. Journal of the American Statistical Association **57** (1962) 33–45
35. Bernstein, S.: The Theory of Probabilities. Gastehizdat Publishing House, Moscow (1946)
36. Efron, B., Stein, C.: The jackknife estimate of variance. Annals of Statistics **9** (1981) 586–596
37. Steele, J.: An Efron-Stein inequality for nonsymmetric statistics. Annals of Statistics **14** (1986) 753–758
38. Vapnik, V., Chervonenkis, A.: On the uniform convergence of relative frequencies of events to their probabilities. Theory of Probability and its Applications **16** (1971) 264–280
39. Devroye, L., Györfi, L., Lugosi, G.: A Probabilistic Theory of Pattern Recognition. Springer-Verlag, New York (1996)
40. Vapnik, V.: Statistical Learning Theory. John Wiley, New York (1998)
41. van der Waart, A., Wellner, J.: Weak convergence and empirical processes. Springer-Verlag, New York (1996)
42. Dudley, R.: Uniform Central Limit Theorems. Cambridge University Press, Cambridge (1999)
43. Koltchinskii, V., Panchenko, D.: Empirical margin distributions and bounding the generalization error of combined classifiers. Annals of Statistics **30** (2002)
44. Massart, P.: Some applications of concentration inequalities to statistics. Annales de la Faculté des Sciencies de Toulouse **IX** (2000) 245–303
45. Bartlett, P., Boucheron, S., Lugosi, G.: Model selection and error estimation. Machine Learning **48** (2001) 85–113
46. Lugosi, G., Wegkamp, M.: Complexity regularization via localized random penalties. submitted (2003)
47. Bousquet, O.: New approaches to statistical learning theory. Annals of the Institute of Statistical Mathematics **55** (2003) 371–389
48. Lugosi, G.: Pattern classification and learning theory. In Györfi, L., ed.: Principles of Nonparametric Learning, Springer, Viena (2002) 5–62
49. Devroye, L., Györfi, L.: Nonparametric Density Estimation: The L_1 View. John Wiley, New York (1985)
50. Devroye, L.: The kernel estimate is relatively stable. Probability Theory and Related Fields **77** (1988) 521–536
51. Devroye, L.: Exponential inequalities in nonparametric estimation. In Roussas, G., ed.: Nonparametric Functional Estimation and Related Topics, NATO ASI Series, Kluwer Academic Publishers, Dordrecht (1991) 31–44
52. Devroye, L., Lugosi, G.: Combinatorial Methods in Density Estimation. Springer-Verlag, New York (2000)
53. Koltchinskii, V.: Rademacher penalties and structural risk minimization. IEEE Transactions on Information Theory **47** (2001) 1902–1914
54. Bartlett, P., Mendelson, S.: Rademacher and Gaussian complexities: risk bounds and structural results. Journal of Machine Learning Research **3** (2002) 463–482
55. Bartlett, P., Bousquet, O., Mendelson, S.: Localized Rademacher complexities. In: Proceedings of the 15th annual conference on Computational Learning Theory. (2002) 44–48
56. Vapnik, V., Chervonenkis, A.: Theory of Pattern Recognition. Nauka, Moscow (1974) (in Russian); German translation: *Theorie der Zeichenerkennung*, Akademie Verlag, Berlin, 1979.
57. Blumer, A., Ehrenfeucht, A., Haussler, D., Warmuth, M.: Learnability and the Vapnik-Chervonenkis dimension. Journal of the ACM **36** (1989) 929–965

58. Anthony, M., Bartlett, P.L.: Neural Network Learning: Theoretical Foundations. Cambridge University Press, Cambridge (1999)
59. Rhee, W., Talagrand, M.: Martingales, inequalities, and NP-complete problems. Mathematics of Operations Research **12** (1987) 177–181
60. Shamir, E., Spencer, J.: Sharp concentration of the chromatic number on random graphs $g_{n,p}$. Combinatorica **7** (1987) 374–384
61. Samson, P.M.: Concentration of measure inequalities for Markov chains and ϕ-mixing processes. Annals of Probability **28** (2000) 416–461
62. Cover, T., Thomas, J.: Elements of Information Theory. John Wiley, New York (1991)
63. Han, T.: Nonnegative entropy measures of multivariate symmetric correlations. Information and Control **36** (1978)
64. Beckner, W.: A generalized Poincaré inequality for Gaussian measures. Proceedings of the American Mathematical Society **105** (1989) 397–400
65. Latała, R., Oleszkiewicz, C.: Between Sobolev and Poincaré. In: Geometric Aspects of Functional Analysis, Israel Seminar (GAFA), 1996-2000, Springer (2000) 147–168 Lecture Notes in Mathematics, 1745.
66. Chafaï, D.: On ϕ-entropies and ϕ-Sobolev inequalities. Technical report, arXiv.math.PR/0211103 (2002)
67. Ledoux, M.: Concentration of measure and logarithmic sobolev inequalities. In: Séminaire de Probabilités XXXIII. Lecture Notes in Mathematics 1709, Springer (1999) 120–216
68. Ledoux, M.: The concentration of measure phenomenon. American Mathematical Society, Providence, RI (2001)
69. Vapnik, V.: The Nature of Statistical Learning Theory. Springer-Verlag, New York (1995)