# A sharp concentration inequality with applications


Stéphane Boucheron
Laboratoire de Recherche en Informatique
Bâtiment 490
CNRS-Université Paris-Sud
91405 Orsay-Cedex

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

Pascal Massart
Mathématiques
Bâtiment 425
Université Paris-Sud
91405 Orsay-Cedex


April 15, 1999


**Abstract**

We present a new general concentration-of-measure inequality and illustrate its power by applications in random combinatorics. The results find direct applications in some problems of learning theory.


*The work of the second author was supported by DGES grant PB96-0300



## 1 Introduction

The phenomenon of measure concentration has recently received distinguished attention due to its much better understanding and its spectacular power and simplicity in applications. The basic methods for proving concentration inequalities have been

(1) martingale methods—see McDiarmid [22], [23] for excellent surveys;

(2) information-theoretic methods, see Alhswede, Gács, and Körner [1], Marton [17], [18],[19], Dembo [5] and Massart [21];

(3) Talagrand's induction method [27],[25],[26], which led to a large variety of powerful new inequalities.

Recently, a new proof technique emerged based on logarithmic Sobolev inequalities, see Ledoux [14],[13]. The method has been shown to provide the sharpest inequalities for empirical processes (Massart [20]). The purpose of this paper is to show that the concentration inequalities obtained by this method have wide applications outside of empirical process theory. In Section 2 we present a general concentration inequality, which is derived from results in Massart [20]. In Section 3 this inequality is applied to prove sharp concentration of certain random combinatorial objects such as the *empirical* VC *dimension* of a family of sets as well as to sharpen earlier concentration inequalities for the length of the longest increasing subsequence and for other configuration functions. In Section 4 we show that the new inequality may be used to prove new concentration inequalities for quantities like the number of increasing subsequences or the empirical Vapnik-Chervonenkis entropy (VC-entropy) of a class of sets. These concentration results have direct applications in learning theory, which are illustrated in Section 5.

## 2 A concentration inequality for nonnegative functionals

The main result of the paper is the following concentration inequality for functionals of independent (not necessarily identically distributed) random variables:

**Theorem 1** *Let $(X_1, ..., X_n)$ be independent random variables taking values in some measurable set $\mathcal{X}$, and let $f : \mathcal{X}^n \to [0, \infty)$ be a function. Assume that there exists another function $g : \mathcal{X}^{n-1} \to \mathbb{R}$ such that for any $x_1, \ldots, x_n \in \mathcal{X}$, the following properties hold:*

$$0 \leq f(x_1, \ldots, x_n) - g(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) \leq 1, \quad \text{for every } 1 \leq i \leq n \tag{1}$$

*and*

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

*Denote $Z = f(X_1, \ldots, X_n)$, and define $h(u) = (1+u)\log(1+u) - u$, for $u \geq -1$. Then for every positive number $t$*

$$\mathbb{P}[Z \geq \mathbb{E}[Z] + t] \leq \exp\left[-\mathbb{E}[Z]\, h\left(\frac{t}{\mathbb{E}[Z]}\right)\right]. \tag{3}$$

*Moreover for every positive number $t \leq \mathbb{E}[Z]$*

$$\mathbb{P}[Z \leq \mathbb{E}[Z] - t] \leq \exp\left[-\mathbb{E}[Z]\, h\left(-\frac{t}{\mathbb{E}[Z]}\right)\right]. \tag{4}$$

The proof of inequality (3) may be obtained by a modification of the proof of an inequality in [20] for the right tail of the supremum of a nonnegative empirical process. The left-tail inequality (4) is new. The proof of a more general version of the theorem is given in the Appendix.

**Remarks.** 1. A typical application of the theorem is to the supremum of sums of nonnegative bounded random variables (empirical processes). Indeed, let $X_1, ..., X_n$ be independent $[0,1]^N$-valued random variables and consider $Z = \sup_{t \leq N} \sum_{i=1}^{n} X_{i,t}$. Defining $f(x_1, \ldots, x_n) = \sup_{t \leq N} \sum_{i=1}^{n} x_{i,t}$ and $g(x_1, \ldots, x_{n-1}) = \sup_{t \leq N} \sum_{i=1}^{n-1} x_{i,t}$, and denoting by $\tau \leq n$ some positive integer such that $f(x_1, \ldots, x_n) = \sum_{i=1}^{n} x_{i,\tau}$, one obviously has

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



and therefore

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

2. Using the same argument as in Massart [20], we see that inequality (4) (and similarly (3)) is in some sense unimprovable. Indeed, consider $N = 1$ and suppose that $X_1, ..., X_n$ are independent Bernoulli trials with probability of success $p = 1 - q$. In this case (4) states that for every $0 < t \leq np$,

$$\mathbb{P}[Z \leq np - t] \leq \exp\left[-np\, h\left(\frac{-t}{np}\right)\right].$$

Given $\theta > 0$, taking $p = \theta/n$ and setting $t = \theta\varepsilon$, this inequality may be written as

$$\mathbb{P}[Z \leq \theta - \theta\varepsilon] \leq \exp[-\theta h(-\varepsilon)], \quad \text{for every } \varepsilon \in (0,1). \tag{5}$$

But $Z$ follows the binomial distribution $\mathcal{B}(n, \theta/n)$ and therefore follows asymptotically the Poisson distribution with parameter $\theta$ as $n$ goes to infinity. Moreover, the right-hand side of (5) is known to be the Cramér-Chernoff deviation upper bound for a Poisson random variable with parameter $\theta$ . This implies that the exponent in this upper bound cannot be improved since Cramér's large deviation asymptotic ensures that for for every $\varepsilon \in (0,1)$

$$\liminf_{\theta \to \infty} \lim_{n \to \infty} \frac{1}{\theta} \log \mathbb{P}[Z \leq \theta - \theta\varepsilon] \geq -h(-\varepsilon).$$

3. It is worth noting that (3) and (4) respectively imply the simpler inequalities

$$\mathbb{P}[Z \geq \mathbb{E}[Z] + t] \leq \exp\left[-\frac{t^2}{2\mathbb{E}[Z] + 2t/3}\right] \tag{6}$$

and

$$\mathbb{P}[Z \leq \mathbb{E}[Z] - t] \leq \exp\left[-\frac{t^2}{2\mathbb{E}[Z]}\right] \tag{7}$$

which hold for any $t > 0$. (6) follows immediately from the inequality

$$h(t) \geq \frac{t^2}{2 + \frac{2t}{3}}, \quad t > 0,$$

and (7) is trivial when $t > \mathbb{E}[Z]$ and follows from (4) otherwise since, for every $t \in [0,1]$,

$$h(-t) \geq \frac{t^2}{2}.$$

## 3 Configuration functions

In this and the next section, we show that Theorem 1 has many natural applications outside empirical process theory. More precisely, we apply Theorem 1 to random combinatorial quantities that were called *configuration functions* in [25, section 7].

**Definition 1** *Assume that we have a sequence of spaces $\Omega_1, \Omega_2, \ldots$ and that we have a property $P$ defined over the union of finite products of spaces: $(\Omega_{i_1} \times \Omega_{i_2} \times \cdots \times \Omega_{i_n})$ with $i_j < i_{j+1}$, that is, for any element $(x_{i_1}, \ldots x_{i_n}) \in \Omega_{i_1} \times \Omega_{i_2} \times \cdots \times \Omega_{i_n}$, we may decide whether $(x_{i_1}, \ldots x_{i_n})$ satisfies property $P$. Moreover assume that $P$ is* hereditary *in the following sense: if $(x_{i_1}, \ldots x_{i_n})$ satisfies $P$ then so does any subsequence $(x_{j_1}, \ldots x_{j_m})$ of $(x_{i_1}, \ldots x_{i_n})$ where $\{j_1 \ldots j_m\} \subseteq \{i_1 \ldots i_n\}$ and $j_k$ is increasing. The function $f_n$ that maps any tuple $(x_{i_1}, \ldots x_{i_n})$ to the size of the largest subsequence satisfying $P$ is the* configuration function *associated with property $P$. Any subsequence of maximal length satisfying property $P$ is called a* witness.

When $\Omega_1, \ldots \Omega_n$ is provided with a product probability measure, results about concentration around the median for configurations functions were proved by Talagrand using the *convex distance* approach [25, 27]. Here we provide concentration around the mean and (slightly) better constants. Moreover the proof is completely straightforward from Theorem 1.

**Theorem 2** *Let $f_n$ be a configuration function, and let $Z = f_n(X_1, \ldots, X_n)$, where $X_1, \ldots, X_n$ are independent random variables. Then for an $t \geq 0$,*

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

*and*

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

**Proof.** Let $f(x_1, \ldots, x_n) = f_n(x_1, \ldots, x_n)$ and $g(x_1, \ldots, x_{n-1}) = f_{n-1}(x_1, \ldots, x_{n-1})$. It suffices to show that $f$ and $g$ satisfy the conditions of Theorem 1. Condition (1) is trivially satisfied. On the other hand, let $\{x_{i_1}, \ldots, x_{i_k}\} \subset \{x_1, \ldots, x_n\}$ be a subsequence of cardinality $k$ witnessing the fact that $f(x_{i_1}, \ldots, x_{i_k}) = \ell$. (Note that such a set exists.) Observing that for any $i \leq n$ such that $x_i \notin \{x_{i_1}, \ldots, x_{i_k}\}$, $f(x_1, \ldots, x_n) = g(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$, we see that (2) is also satisfied, which concludes the proof. $\square$



To illustrate the fact that configuration functions are rather natural objects, let us describe three of them originating from different fields.

1. Increasing subsequences. Consider a vector $x = (x_1, \ldots, x_n)$ 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 $i_1 \geq 1$ and $i_m \leq n$). Let $L(x)$ denote the length of the longest increasing subsequence. $f_n(x) = L(x)$ is a clearly a configuration function (taking $\Omega_i = [0, 1]$ for all $i$), and therefore $Z = L(X_1, \ldots, X_n)$ satisfies the inequalities of Theorem 2, where the $X_i$'s are independent uniform random variables on $[0, 1]$. This improves the constants of the inequalities obtained and Talagrand [25] for the same random variable. See also Frieze [8] for early work on the concentration on $L(X)$.

2. Independent sets in random graphs. 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 subset of vertices $A$ is independent in $G$ if and only if no two vertices from $A$ are adjacent in $G$. Independence is an hereditary property. The size of the largest independent set is the independence number of the graph and it is denoted by $\alpha(G)$.

To show that the independence number can be regarded as a configuration function, we merely have to show that the $\mathcal{G}(n, p)$ model may be regarded as a product probability space. This is well known (see, e.g., [22]): the $i^{\text{th}}$ component of the probability space just defines the set of edges between vertex $i$ and vertices with index $j < i$. Thus, $Z = \alpha(G)$ satisfies the inequalities of Theorem 2, regardless of the values of $n$ and $p$.

Results concerning the average value of $\alpha(G)$ and concentration of $\alpha(G)$ around it have been known for a while for both sparse ($p = d/n$ with $d$ constant) and dense ($p$ constant) random graphs. It is well-known that the independence number of dense random graphs is nearly deterministic (see Bollobás [3]). In this case Theorem 2 does not provide anything new. On the other hand, in the sparse case Frieze [7] proved, that for any $\epsilon > 0$, for sufficiently large $d$ and $n$:

$$|\alpha(G) - \frac{2n}{d}\big(\log d - \log\log d - \log 2 + 1\big)| \leq \frac{\epsilon n}{d} \tag{8}$$

with probability going to 1 as $n \to \infty$. Frieze uses the method of bounded-differences, but gives no explicit concentration inequality for the independence number. It does not seem obvious to get sharp concentration inequalities independent of $d$, from the construction presented in [7]. Such issues are handled in an effortless way by viewing the independence number as a configuration function.

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

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

The *shatter coefficient*, (or *Vapnik-Chervonenkis growth function*) of $\mathcal{A}$ in $x$ is $T(x) = |\mathrm{tr}(x)|$, the size of the trace. $T(x)$ 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* $D(x)$ of $\mathcal{A}$ (with respect to $x$) is the cardinality $k$ of the largest shattered subset of $x$. From the definition, it is obvious that $f_n(x) = D(x)$ is a configuration function, and therefore satisfies the conditions of Theorem 2.

**Remarks.** 1. To illustrate that the constants of Theorem 2 are optimal, consider the VC-dimension and the following example: let $X = (X_1, \ldots, X_n)$ be a vector of i.i.d. random variables taking values in the set of nonnegative integers. Let the common distribution of the $X_i$'s be such that $\mathbb{P}\{X_1 = 0\} = 1 - c/n$ for some positive constant $c$, and $\mathbb{P}\{X_1 = i\} = c/n^3$ for $i = 1, \ldots, n^2$. Then it is easy to see that, for large $n$, $D(X)$ is approximately distributed as a Poisson$(c)$ random variable, so the optimality of the bounds are seen as in Remark 2 following Theorem 1 above.

2. It is likely that Theorem 2 does not provide the sharpest possible answer for the longest increasing subsequence problem, and it does not provide the right answer for the independence number in dense random graphs [2, chapter XI]. The longest increasing subsequence has already been commented by Talagrand in [27]: empirical evidence suggests that the longest increasing subsequence is more concentrated than suggested by Theorem 2.

# 4 Combinatorial entropies

The analysis of combinatorial optipization problems has been often complemented by the analysis of counting versions (see, e.g., [12]): rather than determining the largest independent set or increasing subsequence one may be interested in estimating the number $I(x)$ of independent sets or the number $N(x)$ of increasing subsequences. In statistical pattern recognition, the shatter coefficient $T(x)$ is of primary interest.

The next result shows sharp concentration of combinatorial entropies and particularly of the VC entropy (log-shattering coefficient). These bounds are completely new, we do not know whether any of the previously known concentration inequalities may be used to derive similar bounds. The constants are again optimal by the same example as above.



A combinatorial entropy is defined as follows:

**Definition 2** *let $x = (x_1, \ldots, x_n)$ be an $n$-vector of elements of $\mathcal{X}$ to which we associate a set $\mathrm{Tr}(x) \subset \mathcal{Y}^n$ 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 $i \le n$, the set $\mathrm{Tr}(x^{(i)}) = \mathrm{Tr}(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$ is the the projection of $\mathrm{Tr}(x)$ along the $i^{th}$ coordinate, that is,*

$$\mathrm{Tr}(x^{(i)}) = \Big\{ y^{(i)} = (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n) \in \mathcal{Y}^{n-1} : \tag{9}$$

$$\exists y_i \in \mathcal{Y} \text{ such that } (y_1, \ldots, y_n) \in \mathrm{Tr}(x) \Big\}. \tag{10}$$

*The associated combinatorial entropy is $H(x) = \log_b |\mathrm{Tr}(x)|$ where $b$ is an arbitrary positive number.*

**Remark.** The logarithm of the number of subsequences that satisfy an hereditary property is obviously a combinatorial entropy.

The key property of combinatorial entropies is that they all satisfy condition (2) of Theorem 1:

**Lemma 1** *Let $H(x)$ be a combinatorial entropy. Then for any $x \in \mathcal{X}^n$,*

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

**Proof.** Recall that the *Shannon entropy* (of base $b$) of a discrete random variable $Y$ is

$$h_b(Y) = -\sum_y \mathbb{P}\{Y = y\} \log_b \mathbb{P}\{Y = y\},$$

where the sum is taken over all possible values of $Y$. The key of our proof is the following inequality of Han [9] (see also Cover and Thomas [4, page 491]): for any $n$ discrete random variables $Y_1, \ldots, Y_n$,

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

Now we are prepared to prove the lemma. Consider the uniform distribution over the set $\mathrm{Tr}(x)$. This defines a random vector vector $Y = (Y_1, \ldots, Y_n) \in \mathcal{Y}^n$. Then clearly,

$$H(x) = \log_b |\mathrm{Tr}(x)| = h_b(Y_1, \ldots, Y_n).$$

Since the uniform distribution maximizes the Shannon entropy, we also have, for all $i \le n$, that

$$H(x^{(i)}) \ge h_b(Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_n).$$

The statement now follows from Han's inequality. □

**Remark.** The relationship between isoperimetrical issues and concentration of measure has been underlined many times: concentration-of-measure results may be used to prove or replace isoperimetric inequalities (see for example Ledoux [13]). Here, isoperimetry comes at the rescue of concentration: Han's inequality may be viewed as a weak but general isoperimetric inequality. A geometric version of it had been known for decades before they were formulated in the language of information theory, see Loomis and Whitney [15].

**Remark.** Let us notice that Han's inequality and the tensorization inequality for entropies (inequality (15) in the Appendix) that play a key role in the proof of theorem 1 are both consequences of the subadditivity of the Shannon entropy and the fact that the Shannon entropy decreases with conditioning. Moreover it is easy to check that in the discrete case, Han's inequality and the tensorization inequality (15) can be derived from each other.

**Theorem 3** *Assume that $H(x) = \log_b |\operatorname{Tr}(x)|$ is a combinatorial entropy such that for all $x \in \mathcal{X}^n$ and $i \le n$,*

$$H(x) - H(x^{(i)}) \le 1.$$

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

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

*and*

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

*Moreover,*

$$\mathbb{E}\left[\log_b |\operatorname{Tr}(X)|\right] \le \log_b \mathbb{E}[|\operatorname{Tr}(X)|] \le \frac{1}{\ln 2}\mathbb{E}\left[\log_b |\operatorname{Tr}(X)|\right]. \tag{11}$$



**Remark.** Borrowing the terminology of statistical physics, we may call $\log_b \mathbb{E}[|\operatorname{Tr}(X)|]$ the *annealed combinatorial entropy*. This quantity is often much easier to handle than the expected combinatorial entropy. (11) shows that the two quantities are always closely linked together.

**Proof.** The first two inequalities follow from a straighforward combination of Lemma 1 with Theorem 1. The first inequality of (11) is an obvious consequence of Jensen's inequality. As $\log_b |\operatorname{Tr}(X)|$ satisfies the conditions of Theorem 1, we may use (18) in the Appendix and find that, for all $\lambda > 0$,

$$\mathbb{E}\left[e^{\lambda(\log_b |\operatorname{Tr}(X)| - \mathbb{E}\log_b |\operatorname{Tr}(X)|)}\right] \le e^{(e^\lambda - \lambda - 1)\mathbb{E}[\log_b |\operatorname{Tr}(X)|]}.$$

The choice $\lambda = \log b$ yields the desired result. □

Next we discuss some of the applications of Theorem 3.

1. VC ENTROPIES. The VC *entropy* is defined as $H(x) = \log_2 T(x)$, where $T(x)$ is the shatter coefficient defined in the previous section. The VC entropy is a simple example of a combinatorial entropy. It may be generalized to a class of functions with a finite range. More precisely, let $k > 1$ be a positive integer, and let $\mathcal{F}$ be a class of functions $\mathcal{X} \to \{1, \ldots, k\}$. Given a vector $x = (x_1, \ldots, x_n) \in \mathcal{X}^n$, define $\operatorname{Tr}(x) \subset \{1, \ldots, k\}^n$ as the set of all different $n$-vectors $(f(x_1), \ldots, f(x_n))$ with $f \in \mathcal{F}$. Then it is immediate to see that $H_k(x) = \log_k |\operatorname{Tr}(x)|$ is a combinatorial entropy satisfying the condition of Theorem 3.

The case $k = 2$ (i.e., the case of the VC entropy $H(x)$ is of particular interest, as it plays a key role in some applications in pattern recognition and machine learning (see, e.g., [6], [28]). In this case we obtain the following:

**Corollary 1** *For any class of sets $\mathcal{A}$ and for all $t > 0$, the random VC entropy satisfies*

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

*and*

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

*Also,*

$$\mathbb{E}\left[\log_2 T(X)\right] \le \log_2 \mathbb{E}[T(X)] \le \frac{1}{\ln 2}\mathbb{E}\left[\log_2 T(X)\right].$$

In [28], Vapnik considers the limit of the average VC-entropy rate $\mathbb{E}[\log_2 T(X)]/n$ and the limit of the annealed VC-entropy rate $\log_2 \mathbb{E}[T(X)]/n$ as criteria for consistency and fast convergence of an inference rule called Empirical Risk Minimization. The last statement of the corollary shows that either these two quantities converge to zero or none of them. This answers, in a positive way, an open question raised by Vapnik [28, pages 53–54]: the empirical risk minimization procedure is *non-trivially consistent* and *rapidly convergent* if and only if the annealed entropy rate $(1/n)\log \mathbb{E}[T(X)]$ converges to zero. For the definitions and discussion we refer to [28].

2. Increasing subsequences. Recall the setup of the first example of Section 3, and let $N(x)$ denote the number of different increasing subsequences of $x$. Observe that $\log_2 N(x)$ 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 $x$, a binary $n$-vector $y = (y_1, \ldots, y_n)$ 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 condition (9) as well as the condition of Theorem 3 are obviously met, and therefore $H(X) = \log_2 N(X)$ satisfies all three inequalities of Theorem 3. This result significantly improves a concentration inequality obtained by Frieze [8] for $\log_2 N(X)$.

3. Independent sets in random graphs. The logarithm of the number of independent sets was considered by Zuckerman [29]. This logarithm can also be regarded as a combinatorial entropy: Tr is the set of bitvectors of length $n$, such that the vertices corresponding to coordinates equal to 1 form an independent set.



# 5 Learning and model selection

In this section we point out some immediate applications of the results of Section 3 to some problems emerging in learning theory and pattern recognition. The results presented here are not necessarily optimal, they merely intend to illustrate how some of the new concentration inequalities may be applied in a virtually effortless manner to re-prove and improve some previously obtained results in statistical learning theory.

Let the data $D_n = ((X_1, Y_1), \ldots, (X_n, Y_n))$ consist of independent, identically distributed copies of the random variable pair $(X, Y)$ taking values in $\mathbb{R}^d \times \{0,1\}$. The goal of pattern classification (see [6], [28]) is to construct a function $f_n : \mathbb{R}^d \to \{0,1\}$ that mimimizes $L(f_n)$ where the loss functional $L(\cdot)$ is defined by

$$L(f) = \mathbb{P}[f(X) \neq Y]$$

for all $f : \mathbb{R}^d \to \{0,1\}$. The *empirical loss* of such a function is defined simply by

$$\widehat{L}(f_n) = \frac{1}{n}\sum_{i=1}^{n} I_{\{f_n(X_i)\neq Y_i\}},$$

where $I_A$ denotes the indicator function of an event $A$. Often the data are used to select a set $A_n \subset \mathbb{R}^d$ from a given class $\mathcal{A}$ of subsets of $\mathbb{R}^d$, and the classifier is defined as

$$f_n(x) = I_{\{x \in A_n\}}.$$

The following theorem shows how the loss of such a classifier may be bounded by some purely empirical quantities. It involves the random shatter coefficient of the class $\mathcal{A}$ defined in Section 3. Introduce the notation $X_1^n = (X_1, \ldots, X_n)$.

**Theorem 4** *Let $\mathcal{A}$ be an arbitrary class of subsets of $\mathbb{R}^d$, and let $f_n$ be defined as above. Then for any $\delta > 0$, the probability that*

$$L(f_n) \leq \widehat{L}(f_n) + \sqrt{\frac{6\log T(X_1^n)}{n}} + 5\sqrt{\frac{\log\frac{2}{\delta}}{n}}$$

*is greater than $1 - \delta$.*

**Proof.** For any $t > 0$, and $u > 0$,

$$\begin{aligned}
&\mathbb{P}\left[L(f_n) - \widehat{L}(f_n) > \sqrt{\frac{6\log T(X_1^n)}{n}} + \sqrt{\frac{6t}{n}} + u\right] \\
&\le \mathbb{P}\left[L(f_n) - \widehat{L}(f_n) > \sqrt{\frac{6\log T(X_1^n) + 6t}{n}} + u\right] \\
&\le \mathbb{P}\left[L(f_n) - \widehat{L}(f_n) > \sqrt{\frac{3\mathbb{E}\log T(X_1^n)}{n}} + u\right] \\
&\quad + \mathbb{P}\left[\mathbb{E}\log T(X_1^n) > 2\log T(X_1^n) + 2t\right].
\end{aligned}$$

The first term on the right-hand side may be bounded by recalling the Vapnik-Chervonenkis inequality (see [28, Theorem 3.1]):

$$\mathbb{P}\left[\sup_{f=I_A : A\in\mathcal{A}} L(f) - \widehat{L}(f) > \epsilon\right] \le \mathbb{E}T(X_1^{2n})e^{-n\epsilon^2},$$

which is true for any $\epsilon > 0$. Now it is easy to see that for any $x_1^{2n} = (x_1, \ldots, x_{2n})$, $T(x_1^{2n}) \le T(x_1^n)T(x_{n+1}^{2n})$, so by independence we have $\log\mathbb{E}T(X_1^{2n}) \le 2\log\mathbb{E}T(X_1^n)$. Therefore,

$$\begin{aligned}
&\mathbb{P}\left[L(f_n) - \widehat{L}(f_n) > \sqrt{\frac{3\mathbb{E}\ \log T(X_1^n)}{n}} + u\right] \\
&\le \mathbb{P}\left[L(f_n) - \widehat{L}(f_n) > \sqrt{\frac{2\log\mathbb{E}T(X_1^n)}{n}} + u\right] \\
&\quad \text{(by inequality 11 from Theorem 3)} \\
&\le \mathbb{P}\left[L(f_n) - \widehat{L}(f_n) > \sqrt{\frac{\log\mathbb{E}T(X_1^{2n})}{n}} + u\right] \\
&\quad \text{(by the argument above)} \\
&\le e^{-nu^2},
\end{aligned}$$

where at the last step we used the Vapnik-Chervonenkis inequality.

To bound the second term, we use the lower-tail inequality of the concentration result Theorem 3 for the VC entropy $\log T(X_1^n)$:

$$\begin{aligned}
\mathbb{P}\left[\mathbb{E}\log T(X_1^n) > 2\log T(X_1^n) + 2t\right] &\le \exp\left(-\frac{\left(\frac{1}{2}\mathbb{E}\log T(X_1^n) + t\right)^2}{2\mathbb{E}\log T(X_1^n)}\right) \\
&\le e^{-t/2}.
\end{aligned}$$

Summarizing, we have

$$\mathbb{P}\left[L(f_n) - \widehat{L}(f_n) > \sqrt{\frac{6\log T(X_1^n)}{n}} + \sqrt{\frac{6t}{n}} + u\right] \le e^{-nu^2} + e^{-t/2}.$$



Choosing $u = \sqrt{\frac{1}{n}\log\frac{2}{\delta}}$ and $t = 2\log\frac{2}{\delta}$ concludes the proof. □

The above result is important because the (unknown) loss may be controlled by a purely empirical quantity. Such a result is useful in automatic model selection. Assume now that a sequence of classes of sets $\mathcal{A}_1, \mathcal{A}_2, \ldots$ is given, and for each class there is a classifier $\widehat{f}_k$ which chooses its hypothesis from class $\mathcal{A}_k$ (i.e., $\widehat{f}_k(x) = I_{\{x\in A_k\}}$ for some $A_k \in \mathcal{A}_k$). Then it is immediate from the above Theorem that with probability greater than $1-\delta$, simultaneously for all $k \ge 1$,

$$L(\widehat{f}_k) < \widehat{L}(\widehat{f}_k) + \sqrt{\frac{6\log T_k(X_1^n)}{n}} + 6\sqrt{\frac{\log\frac{2}{\delta} + 2\log k}{n}},$$

where $T_k(X_1^n)$ is the random shatter coefficient of class $\mathcal{A}_k$. This suggests a model selection rule based on minimizing the empirical quantity on the right-hand side. Note that the right-hand side is the empirical loss penalized by a data-dependent penalty, involving the random VC entropy of the $k$-th class. In particular, the following result is an immediate consequence.

**Theorem 5** *Let $\widehat{f}_1, \widehat{f}_2, \ldots$ be a sequence of classifiers defined as above. Assume that the classifier $f_n$ is selected among these by minimizing the quantity*

$$\widehat{L}(\widehat{f}_k) + \sqrt{\frac{6\log T_k(X_1^n)}{n}} + 6\sqrt{\frac{\log\frac{2}{\delta} + 2\log k}{n}}$$

*over all $k = 1, 2, \ldots$. Then with probability greater than $1-\delta$,*

$$L(f_n) \le \inf_k\left(\widehat{L}(\widehat{f}_k) + \sqrt{\frac{6\log T_k(X_1^n)}{n}} + 6\sqrt{\frac{\log\frac{2}{\delta} + 2\log k}{n}}\right).$$

Similar results, but with significantly more involved proofs were shown by Shawe-Taylor, Bartlett, Willamson, and Anthony [24] and Lugosi and Nobel [16]. For discussion on the significance of these results and related work we refer to these papers.

# Appendix: Proof of Theorem 1

Here we prove the following, stronger, version of Theorem 1:

**Theorem 6** . *Let $\{X_i, i \in \mathcal{I}\}$ be some finite family of independent random variables. Define $X = (X_j)_{j\in\mathcal{I}}$ and for every $i \in \mathcal{I}$, let $X^{(i)} = (X_j)_{j\in\mathcal{I}\setminus\{i\}}$. Let $Z = \zeta(X)$, be some nonnegative and bounded measurable function of $X$. Assume that for every $i \in \mathcal{I}$ there exists some measurable function $Z^{(i)}$ of $X^{(i)}$ such that*

$$0 \leq Z - Z^{(i)} \leq 1. \tag{12}$$

*Assume furthermore that*

$$\sum_{i\in\mathcal{I}} \left(Z - Z^{(i)}\right) \leq Z. \tag{13}$$

*Defining $h$ as $h(u) = (1+u)\log(1+u) - u$, for $u \geq -1$, the following inequalities hold. For every positive number $t$*

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

*and for every positive number $t \leq \mathbb{E}[Z]$*

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

To prove Theorem 6, we need a modification of an information inequality for functionals of independent variables presented in Massart [20] (see Lemma 2.3 therein).

**Lemma 2** *Let $\mathcal{I}$ be some finite set. Let, for every $i \in \mathcal{I}$, $X_i$ be some random variable with values in some measurable space $\Omega_i$ and define $\Omega = \prod_{j\in\mathcal{I}} \Omega_j$, $\Omega^{(i)} = \prod_{j\in\mathcal{I}\setminus\{i\}} \Omega_j$. Let $\zeta$ be some real valued measurable function on $\Omega$ and for every $i \in \mathcal{I}$, $\zeta^{(i)}$ be some real valued measurable function on $\Omega^{(i)}$. Define $X = (X_j)_{j\in\mathcal{I}}$ , $Z = \zeta(X)$ and for every $i \in \mathcal{I}$ $X^{(i)} = (X_j)_{j\in\mathcal{I}\setminus\{i\}}$, $Z^{(i)} = \zeta^{(i)}\left(X^{(i)}\right)$. If we assume the variables $\{X_i, i \in \mathcal{I}\}$ to be independent and the Laplace transform $\lambda \to \mathbb{E}\left[e^{\lambda Z}\right]$ to be finite on some non empty open interval $I$ then, for any $\lambda \in I$*

$$\lambda \mathbb{E}\left[Ze^{\lambda Z}\right] - \mathbb{E}\left[e^{\lambda Z}\right] \log \mathbb{E}\left[e^{\lambda Z}\right] \leq \sum_{i\in\mathcal{I}} \mathbb{E}\left[e^{\lambda Z} \phi\left(-\lambda(Z - Z^{(i)})\right)\right], \tag{14}$$

*where $\phi$ denotes the function $z \to \exp(z) - z - 1$.*



**Proof.** As the proof is exactly the same as that of Lemma 2.3 in Massart [20], we just give a sketch. For every $i \in \mathcal{I}$, we denote by $\mathbb{E}^{(i)}$ the expectation operator conditionally on $X^{(i)}$. Then, introducing $\Phi(t) = t \log t$, the tensorization inequality for entropy (see Ledoux [14]), yields, for any nonnegative function $g$ on $\Omega$ such that $G = g(X)$ satisfies $\mathbb{E}[G\,|\log G|] < \infty$,

$$\mathbb{E}[\Phi(G)] - \Phi(\mathbb{E}[G]) \leq \mathbb{E}\left[\sum_{i\in\mathcal{I}} \mathbb{E}^{(i)}[\Phi(G)] - \Phi\left(\mathbb{E}^{(i)}[G]\right)\right]. \tag{15}$$

Then, for every every $i \in \mathcal{I}$, the variational definition of entropy [11] asserts that for every positive measurable function $G^{(i)}$ of $X^{(i)}$, one has

$$\mathbb{E}^{(i)}[\Phi(G)] - \Phi\left(\mathbb{E}^{(i)}[G]\right) \leq \mathbb{E}^{(i)}\left[G\left(\log G - \log G^{(i)}\right) - \left(G - G^{(i)}\right)\right].$$

Applying the above inequality to the variables $G = e^{\lambda Z}$ and $G^{(i)} = e^{\lambda Z^{(i)}}$, one gets

$$\mathbb{E}^{(i)}[\Phi(G)] - \Phi\left(\mathbb{E}^{(i)}[G]\right) \leq \mathbb{E}^{(i)}\left[e^{\lambda Z}\phi(-\lambda(Z - Z^{(i)}))\right]$$

which, via (15), leads to (14). □

**Proof of Theorem 6.** We apply Lemma 2 so that inequality (14) holds for any $\lambda$. Since the function $\phi$ is convex with $\phi(0) = 0$, for any $\lambda$ and any $u \in [0,1]$ , $\phi(-\lambda u) \leq u\phi(-\lambda)$. Hence it follows from (12) that for every $\lambda$, $\phi(-\lambda\left(Z - Z^{(i)}\right)) \leq \left(Z - Z^{(i)}\right)\phi(-\lambda)$ and therefore we derive from (14) and (13) that

$$\begin{aligned} \lambda \mathbb{E}\left[Ze^{\lambda Z}\right] - \mathbb{E}\left[e^{\lambda Z}\right] \log \mathbb{E}\left[e^{\lambda Z}\right] &\leq \mathbb{E}\left[\phi(-\lambda)e^{\lambda Z} \sum_{i\in\mathcal{I}} \left(Z - Z^{(i)}\right)\right] \\ &\leq \phi(-\lambda)\mathbb{E}\left[Ze^{\lambda Z}\right]. \end{aligned}$$

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

$$[\lambda - \phi(-\lambda)] \frac{F'(\lambda)}{F(\lambda)} - \log F(\lambda) \leq v\phi(-\lambda), \tag{16}$$

which in turn implies

$$\left(1 - e^{-\lambda}\right) \Psi'(\lambda) - \Psi(\lambda) \leq v\phi(-\lambda) \text{ with } \Psi(\lambda) = \log F(\lambda).$$

Now observe that the function $\Psi_0 = v\phi$ is a solution of the ordinary differential equation $\left(1 - e^{-\lambda}\right) \Psi'(\lambda) - \Psi(\lambda) = v\phi(-\lambda)$. We want to show that $\Psi \leq \Psi_0$. In fact, if $\Psi_1 = \Psi - \Psi_0$, then

$$\left(1 - e^{-\lambda}\right) \Psi_1'(\lambda) - \Psi_1(\lambda) \leq 0. \tag{17}$$

Hence, defining $f(\lambda) = \log\left(e^{\lambda} - 1\right)$ and $g(\lambda) = e^{-f(\lambda)}\Psi_1(\lambda)$, we have

$$\left(1 - e^{-\lambda}\right)\left[f'(\lambda)g(\lambda) + g'(\lambda)\right] - g(\lambda) \leq 0,$$

which yields since $f'(\lambda)(1 - e^{-\lambda}) = 1$

$$\left(1 - e^{-\lambda}\right) g'(\lambda) \leq 0.$$

Hence $g'$ is nonnegative on $(-\infty, 0)$ and nonpositive on $(0, \infty)$ and therefore $g$ is nondecreasing on $(-\infty, 0)$ and nonincreasing on $(0, \infty)$. Now, since $\widetilde{Z}$ is centered $\Psi_1'(0) = 0$. Using the fact that $\lambda e^{-f(\lambda)}$ tends to 1 as $\lambda$ goes to 0, we conclude that $g(\lambda)$ tends to 0 as $\lambda$ goes to 0. This shows that $g$ is nonpositive, therefore $\Psi \leq \Psi_0$ and we have proved that

$$\log \mathbb{E}\left[e^{\lambda(Z - \mathbb{E}[Z])}\right] \leq v\phi(\lambda) \text{ for every } \lambda \in \mathbb{R}. \tag{18}$$

Then by Markov's inequality

$$\mathbb{P}\left[Z - \mathbb{E}[Z] \geq t\right] \leq \exp\left[-\sup_{\lambda > 0}\left(t\lambda - v\phi(\lambda)\right)\right]$$

and

$$\mathbb{P}\left[Z - \mathbb{E}[Z] \leq -t\right] \leq \exp\left[-\sup_{\lambda < 0}\left(-t\lambda - v\phi(\lambda)\right)\right].$$

The proof can be completed by using the easy-to-check (and well-known) relations: $\sup_{\lambda > 0}\left[t\lambda - v\phi(\lambda)\right] = vh(t/v)$ for every $t > 0$ and $\sup_{\lambda < 0}\left[-t\lambda - v\phi(\lambda)\right] = vh(-t/v)$ for every $0 < t \leq v$. □



# 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] B. Bollobás. *Random Graphs*. Academic Press, Orlando, 1985.

[3] B. Bollobás. *Combinatorics*. Cambridge University Press, Cambridge, 1986.

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

[5] A. Dembo. Information inequalities and concentration of measure. *Annals of Probability*, 24, 1996.

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

[7] A.M. Frieze. On the independence number of random graphs. *Discrete Mathematics*, 81:171–175, 1990.

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

[9] T.S. Han. Non negative entropy measures of multivariate symmetric correlations. *Information and Control*, 36:133–156, 1978.

[10] D. Haussler, N. Littlestone, and M. Warmuth. Predicting $\{0, 1\}$-functions on randomly drawn points. *Information and Computation*, 115:248–292, 1994.

[11] R. Holley and D. Stroock. Logarithmic Sobolev Inequalities and stochastic Ising models. *J. Statist. Phys.*,46 (1987), 1159–1194.

[12] D.S. Johnson. A catalog of complexity classes. in *Handbook of Theoretical Computer Science*, A:67–162, MIT Press, Boston, 1990.

[13] 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.

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

[15] L.H. Loomis and H. Whitney. An inequality related to the isoperimetric inequality. *Bull. Amer. Math. Soc.* 55:961-962, 1949.

[16] G. Lugosi and A. Nobel. Adaptive model selection using empirical complexities. *under revision*, 1996.

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

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

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

[20] P. Massart. About the constants in Talagrand's concentration inequalities for empirical processes. *Annals of Probability*, page to appear, 1998.

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

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

[23] 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, 1997.

[24] J. Shawe-Taylor, P.L. Bartlett, R.C. Williamson, and M. Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE Transactions on Information Theory*, 44:1926–1940, 1998.

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

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

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

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

[29] D. Zuckerman. Every NP-complete problem has a hard version. *Proceedings of eighth IEEE Structure in Complexity Theory Conference*, 1993, 305–312.