The Annals of Probability
1996, Vol. 24, No. 1, 1–34

SPECIAL INVITED PAPER

A NEW LOOK AT INDEPENDENCE[1]

BY MICHEL TALAGRAND

C.N.R.S. and Ohio State University

The concentration of measure phenomenon in product spaces is a far-reaching abstract generalization of the classical exponential inequalities for sums of independent random variables. We attempt to explain in the simplest possible terms the basic concepts underlying this phenomenon, the basic method to prove concentration inequalities and the meaning of several of the most useful inequalities.

1. Introduction. What is the most important theorem of probability? The following statement could be a reasonable candidate:

(1.1) In a long sequence of tossing a fair coin, it is likely that heads will come up nearly half of the time.

This rather imprecise statement could serve as an introduction to the study of laws of large numbers. These are limit theorems. A commonly heard piece of conventional wisdom (which certainly should not be hastily dismissed) asserts, however, that the "age of computing" coming upon us will shift much of the focus of mathematics from the infinite to the discrete. A precise discrete statement of (1.1) is as follows:

Consider an independent sequence of Bernoulli random variables $(\varepsilon_i)_{i\le N}$ [i.e., $P(\varepsilon_i = 1) = P(\varepsilon_i = -1) = \frac{1}{2}$]. Then for all $t \ge 0$ we have the following inequality [which will be proved in (4.7) below:

$$P\left(\left|\sum_{i\le N}\varepsilon_i\right| \ge t\right) \le 2\exp\left(-\frac{t^2}{2N}\right). \tag{1.2}$$

To relate (1.2) to (1.1), we simply observe that if B_N is the number of 1's in the sequence $(\varepsilon_i)_{i\le N}$, then $\sum_{i\le N}\varepsilon_i = 2B_N - N$, so that (1.2) is equivalent to

$$P\left(\left|B_N - \frac{N}{2}\right| \ge t\right) \le 2\exp\left(\frac{-2t^2}{N}\right). \tag{1.3}$$

Inequality (1.2) is possibly the simplest occurrence of the concentration of measure phenomenon that will be explored in the present paper. Upon evoking generalizations of (1.2), the words "exponential inequalities" and the names of Chernoff, Bennett, Prokhorov and Hoeffding (and more) come to mind. The generalizations of (1.2) we have in mind, however, require a change

Received October 1994.

[1]Based on the Rietz Lecture delivered by the author at the 58th Annual Meeting of the IMS in Montreal, Canada, in July 1995.

AMS 1991 *subject classifications*. Primary 60E15, 28A35.

Key words and phrases. Exponential inequalities, concentration of measure, product spaces.

of perspective: simply to think of the random variable $X = \sum_{i \le N} \varepsilon_i$ as a *function* of the individual variables ε_i and to state (1.2) [or rather (1.1)] as

(1.4) X is essentially constant $(= 0)$.

This statement seems pretty offensive, since the fluctuations of X are of order $\sqrt{N}$, which is hardly zero. This impression is misleading and is simply created by the fact that we do not look at X on the proper scale. As X can take values as large as N, this should be the scale at which one should measure X, in which case (1.4) is indeed true (i.e., X/N is essentially zero).

In words, the form of the concentration of measure phenomenon we will study could be stated as follows:

(1.5) A random variable that depends (in a "smooth" way) on the influence of many independent variables (but not too much on any of them) is essentially constant.

This statement will of course be quantified by inequalities such as (1.2). Most of these inequalities will be of exponential type, so another (shameless ...) way to advertise the results of the present paper is by the following:

(1.6) A random variable that smoothly depends on the influence of many independent random variables satisfies Chernoff-type bounds.

It should be self-evident why a statement such as (1.6) is important. Of special interest is the case where the random variable is defined in an indirect or a complicated way and where explicit computations are all but impossible. A typical situation is the case where the random variable is the solution of a (stochastic) optimization problem, in which case it is sometimes rather hard to say anything at all about it. The body of inequalities underlying the imprecise statement (1.6) has by now been applied to a variety of such optimization problems and has in each occurrence improved and streamlined previous results. These problems include, in particular, stochastic versions of famous questions such as bin packing, the traveling salesman problem and, not surprisingly, models for randomness in physics, such as percolation theory and models for disordered matter in statistical mechanics. (Many applications have also been given to more classical areas of probability such as probability in Banach spaces [9] and empirical processes theory [27].) While going through a large number of applications would have been a fair attempt at impressing upon the reader the importance of the present material, I have resisted the temptation. The main reason is that the abstract inequalities that form the core of the paper (and, in particular, the one presented in Section 6) are sufficiently powerful so that once the basic mechanism of their application is understood, this application becomes mostly a routine matter. The two examples presented in Section 6 should be a sufficient illustration. Numerous other applications are presented in [29], and I hope that the reader will be interested enough by the present essay to ask for more and will be immediately at ease while plunging into this considerably more detailed work.

While the topic of giving a meaning to (1.6) has now become almost a theory in itself, it is a rather pleasant fact that the proof of the main results is very simple. However, how can such simply obtained results have such drastic consequences? The answer lies, of course, in using a good point of view. This requires several layers of abstraction. While the key ideas are again very simple once understood, this is not necessarily the case beforehand. Therefore, these ideas will be explained in considerable detail, and I must apologize should I insist too much on trivialities; triviality is apparently in the eye of the beholder [32]. The true motivation for insisting upon the abstract ideas is that it is while pursuing abstract principles that the main discoveries have been made, and thereby this appears to be the best way of fostering further advances.

The idea of concentration of measure (which was discovered by V. Milman) is arguably one of the great ideas of analysis in our times. While its impact on Probability is only a small part of the whole picture, this impact should not be ignored. The present paper represents my best attempt to explain in the simplest way I can achieve what this is all about, without *ever* doing anything technical. Due to this exacting requirement of simplicity (and even more to space limitation), the present work is very far from being a complete account of what is known. (We refer for this to [28–30]). I hope, however, that it will be informative for the casual reader and will even possibly induce him/her to learn more about this ever-fascinating topic.

2. The Gromov–Milman formulation. The Gromov–Milman [4, 17] formulation is rather simple and very effective. It is also our first step toward increased abstraction and the opportunity to stress a number of key features.

First of all, to examine (1.5) it will be convenient, in contrast with a long-standing tradition, to specify the underlying probability space. The probabilistic notion of independence is intimately related to the notion of product measure, and product measures will be the focus of our interest.

Consider a probability space (Ω, Σ, μ) and a power (Ω^N, P), where $P = \mu^{\otimes N}$. One could consider different factors, but it would not truly increase the generality. The coordinate functions are probabilistically independent, and any sequence of probabilistically independent functions can be realized as above. Thus to study (1.5), we will study functions defined on a product of probability spaces provided with a product measure.

How should we define the fact that a function depends smoothly on the argument? A reasonable answer seems to be that a small variation of the argument produces a small change in the value of the function. The most natural way to define a small variation of the argument is to assume that the underlying space is provided with a distance. Fortunately, a product space Ω^N is provided with a natural distance: the Hamming distance given by

$$d(x, y) = \mathrm{card}\{i \leq N;\ x_i \neq y_i\}, \tag{2.1}$$

where $x = (x_i)_{i \leq N}$ and $y = (y_i)_{i \leq N}$. This initial success should not hide a basic limitation: unless the factor Ω is provided with some kind of structure,

it seems difficult to define a genuinely different distance than (2.1). Much of Sections 6–8 will be devoted to showing how to bypass this limitation.

The basic object in the Gromov–Milman formulation of the concentration of measure phenomenon is a (Polish) metric space (X,d) provided with a (Borel) probability P. It is not required here that X be a product space, so that this formulation is considerably more general than the special case of product spaces. Quite naturally, in view of the preceding discussion, the class of well-behaved functions will be the class of 1-Lipschitz functions, that is, functions f from X to $\mathbb{R}$ that satisfy

$$\forall\, x,y \in X, \qquad |f(x)-f(y)| \le d(x,y), \tag{2.2}$$

and the object is to find situations where the Lipschitz functions are essentially constant. How do we identify the value of the constant? It turns out that the most convenient choice is through a median M_f of f, that is, a number such that

$$P(f \le M_f) \ge \tfrac{1}{2}, \qquad P(f \ge M_f) \ge \tfrac{1}{2}.$$

The statement that f is essentially constant is then quantified by a bound for

$$P(|f-M_f| \ge t),$$

for $t>0$.

Consider the set $A=\{f \le M_f\}$. Thus $P(A) \ge \frac{1}{2}$. Consider the set

$$A_t = \{x \in X;\ \inf\{d(x,y);\ y \in A\} \le t\} = \{x;\ d(x,A) \le t\}. \tag{2.3}$$

It follows from (2.2) that

$$x \in A_t \quad \Rightarrow \quad f(x) \le t + M_f,$$

so that

$$P(f > M_f + t) \le 1 - P(A_t).$$

This simple observation has accomplished a key fact: it has reduced the study of functions f to the study of sets, which are genuinely simpler objects, and the central concern now is to show that when $P(A) \ge \frac{1}{2}$, the "enlargement" of A defined by (2.3) has probability close to 1. This question (in the setting of product spaces) will be the central objective of the paper. To quantify this phenomenon in their setting, Gromov and Milman introduce the *concentration function* $\alpha(P,t)$ (depending also on X,d) as the smallest number such that

$$P(A) \ge \tfrac{1}{2} \quad \Rightarrow \quad 1 - P(A_t) \le \alpha(P,t).$$

The above discussion should then make clear that, for any 1-Lipschitz function f,

$$P(|f - M_f| > t) \le 2\alpha(P,t). \tag{2.4}$$

If we define the Lipschitz constant $\|f\|_{\mathrm{Lip}}$ of any function f on X as the smallest number such that

$$\forall\, x,y \in X, \qquad |f(x)-f(y)| \le \|f\|_{\mathrm{Lip}}\, d(x,y),$$

homogeneity and (2.4) imply

$$P(|f - M_f| > t) \leq 2\alpha\left(P, \frac{t}{\|f\|_{\mathrm{Lip}}}\right). \tag{2.5}$$

The point of these definitions is that, in a variety of situations, the function $\alpha(P,t)$ decreases very fast as t increases. We will summarize this in a somewhat imprecise manner by the statement that concentration of measure holds in that case. The origin of this terminology is that, whenever one considers a set A with $P(A) \geq \frac{1}{2}$, most of the points of X are close to A; thus P "concentrates" around each such set A.

In Section 5, we will prove somewhat more than the following proposition.

PROPOSITION 2.1. *If X is the product of N probability spaces, P is a product measure and X is provided with the Hamming distance d, the concentration function satisfies*

$$\alpha(P,t) \leq 2\exp\left(-\frac{t^2}{N}\right). \tag{2.6}$$

In order to compare this with (1.2), on the space $X = \{-1,1\}^N$, we consider the function f that is the sum of the coordinates and we observe that [when X is provided with the Hamming distance given by (2.1)] $\|f\|_{\mathrm{Lip}} = 2$. Since $M_f = 0$ by symmetry, combining (2.4) and (2.6) yields

$$P\left(\left|\sum_{i\leq N} \varepsilon_i\right| \geq t\right) \leq 4\exp\left(-\frac{t^2}{4N}\right). \tag{2.7}$$

This is not quite as good as (1.2), but still captures its main features.

A prime example of a space where concentration of measure holds is the sphere S_N of $\mathbb{R}^{N+1}$ equipped with its geodesic distance d and normalized Haar measure Q_N. In that case, P. Lévy proved in 1919 that for any (regular) set A of S_N we have

$$Q_N(A_t) \geq Q_N(C_t), \tag{2.8}$$

where C is a cap of the same measure as A. (This is a true isoperimetric inequality.) It follows, in particular, through a simple computation that

$$\alpha(P_N,t) \leq \left(\frac{\pi}{8}\right)^{1/2} \exp\left(-\frac{(N-1)}{2}t^2\right).$$

Keeping in mind that in (2.6) the diameter of X is N, while in (2.8) it is 1, one sees a great similarity between these inequalities. Around 1970, Milman understood that (2.8) was the key to the famous theorem of Dvoretzky on almost Euclidean sections of convex bodies [15, 16]. Subsequently, Milman most vigorously promoted the concept of concentration of measure (see, e.g., [1] for early thoughts about product spaces), and his ideas had a considerable influence. This concept now plays an important role in the local theory of Banach spaces and the dominant role in probability in Banach space. (This

author is, in particular, pleased to acknowledge that his contributions in this direction have their ultimate source in Milman's philosophy.)

More in line with the topic of the present paper is the case where $X = \mathbb{R}^N$ is provided with the Euclidean distance and where $P = \gamma_N$ is the canonical Gaussian measure. Thus γ_N is a product measure when each factor is provided with the canonical Gaussian measure γ_1 on $\mathbb{R}$, of density $(2\pi)^{-1/2}\exp(-t^2/2)$. The importance of this situation stems from the fact that all Gaussian measures (such as Wiener measure) can be suitably approximated by γ_N and that inequalities proved for γ_N can rather trivially be transferred to them.

The Gaussian measure on $\mathbb{R}^N$ can be seen as the limit of the projection of the dilatation of Q_M by a factor $\sqrt{M}$ on $\mathbb{R}^N$ as $M \to \infty$, a fact known as Poincaré's lemma. It can then be deduced from (2.8) that

$$\alpha(\gamma_N, t) \le \int_t^\infty \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-u^2}{2}\right) du \le \frac{1}{2}\exp\left(\frac{-t^2}{2}\right), \tag{2.9}$$

a fact of considerable importance [8, 9].

3. Classical isoperimetry and rearrangements. Inequalities such as (2.6) will be called concentration inequalities, and it is instructive to discuss the relationship of such an inequality with classical isoperimetry. The most recognized isoperimetric inequality is likely to be the following statement.

(3.1) Of the bodies of a given volume in $\mathbb{R}^N$, the Euclidean ball is the one with the smallest surface area.

This formulation needs the notion of surface area, which in the present case can be defined (when ∂A is smooth enough) as

$$\mathrm{Vol}_{N-1}(\partial A) = \lim_{t\to 0} \frac{\mathrm{Vol}_N(A_t \setminus A)}{t}, \tag{3.2}$$

where A_t is the set of points within Euclidean distance t of A.

As it turns out, (3.1) is equivalent to a lesser-known formulation that does not require the notion of surface area.

(3.3) Among the bodies A of a given volume in $\mathbb{R}^N$, the ones for which A_t has minimum volume are the Euclidean balls.

It should be clear through (3.2) that (3.3) implies (3.1) as $t \to 0$. Conversely, bounding below $d\,\mathrm{Vol}_N(A_t)/dt$ through (3.1) and (3.2) and integrating yield (3.3). The topic of Section 2 connects with (3.3) for the *large* values of t. This is uninteresting when $N = 3$, but it would be disastrous to stop there because our intuition does *not* function beyond the case $N \le 3$.

In the Gaussian case, the statement corresponding to (3.3) is as follows:

(3.4) Among the sets A of given measure (for γ_N), the ones for which $\gamma_N(A_t)$ are minimal are the half-spaces

(cf. [8, 9] and the references therein).

Using this when the half-space is orthogonal to a basis vector yields

$$\gamma_N(A) = \gamma_1((-\infty, a]) \quad \Rightarrow \quad \gamma_N(A_t) \geq \gamma_1((-\infty, a+t]), \tag{3.5}$$

from which (2.9) follows in the case $a = 0$.

An inequality such as (3.5) is extremely satisfactory. It is optimal and points to the so-called extremal sets on which equality occurs (here the half-spaces). It apparently is impossible to obtain a very simple proof of (3.5), or indeed of any inequality with the same quality of precision. The only known approach is based on rearrangements. Starting with A, one constructs a set $T(A)$ which is somewhat more regular than A, such that $\gamma_N(A) = \gamma_N(T(A))$, while $\gamma_N(T(A)_t) \leq \gamma_N(A_t)$. One then iterates the construction in such a way that the iterates "converge" to an extremal set. [See [3] for a proof of (3.5) in this spirit.] This is somewhat delicate. More important, it seems that the method is bound to failure unless the extremal sets have a reasonably simple structure. This does not appear to be the case in a number of situations of crucial interest. Therefore, it is of primary importance to find other methods.

To finish this section, we will describe a result which, while not in the main line of the paper, is connected by several key features. This result is of the same nature as (3.1), but in a setting where it was not obvious how to define "surface area." The space is $\Omega = \{-1, 1\}^N$ provided with the uniform measure P_N.

Given $x \in \Omega$ and $i \leq N$, we define the point $T_i x$ obtained by changing the sign of the ith component of x. Given a subset A of Ω and $x \in A$, we define

$$h_A(x) = \text{card}\{i \leq N;\ T_i(x) \notin A\}.$$

Thus $h_A(x)$ counts "the number of directions along which one can leave A from x." The following theorem was motivated by a result of Margulis [10].

THEOREM 3.1 [25]. *For some universal constant K and all subsets A of Ω, we have*

$$\int_A \sqrt{h_A(x)}\, dP_N(x) \geq \frac{1}{K} P_N(A)(1 - P_N(A)) \sqrt{\log \frac{1}{P_N(A)(1 - P_N(A))}}. \tag{3.6}$$

The philosophy of the result is that the left-hand side is a measure of the "surface area" of A.

Thus, (3.6) provides a lower bound for the surface area of A, given the "volume" $P_N(A)$ of A. To understand better the nature of this lower bound, we first state the Gaussian version of (3.1), which follows from (3.5) the way (3.1) follows from (3.3). We have

$$\gamma_N(A) = \gamma_1((-\infty, a]) \quad \Rightarrow \quad s_{N-1}(A) \geq \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-a^2}{2}\right), \tag{3.7}$$

where the "Gaussian surface area" $s_{N-1}(A)$ is defined as

$$s_{N-1}(A) = \lim_{t \to 0} t^{-1} \gamma_N(A_t \backslash A).$$

If we remember that, for $a \leq -1$, $\gamma_1([-\infty, a])$ is of order $(1/|a|)\exp(-a^2/2)$, we see that (3.7) implies

$$s_{N-1}(A) \geq \frac{1}{K}\gamma_N(A)(1-\gamma_N(A))\sqrt{\log\frac{1}{\gamma_N(A)(1-\gamma_N(A))}}. \tag{3.8}$$

The similarity between (3.6) and (3.8) is no accident. It arises from the fact that $(\mathbb{R}^N, \gamma_N)$ is essentially a "quotient" of $(\{-1,1\}^{N'}, P_{N'})$ when $N' \gg N$ (so that isoperimetry in the latter cannot be better than in the former). To see this, we simply observe that when M is large, γ_1 is close to the image of P_M under the map $(x_i)_{i\leq M} \to M^{-1/2}\sum_{i\leq M} x_i$, by the central limit theorem, so that γ_N is close to an image of P_{NM}. Thus, (3.6) can be seen as extending some aspects of (3.7).

One important feature of (3.6) (proved by induction over N) is that while it provides a bound of the correct order, it avoids the considerably more difficult "extremal" problem of determining the infimum of the left-hand side given $P_N(A)$. As already mentioned, this feature is shared by many of the inequalities we will present.

As should be expected from the discussion relating (3.6) and (3.8) and as is easy to see, both sides of (3.6) are of the same order when A is a set of the type

$$A_{n,k} = \left\{(x_i)_{i\leq N};\ \sum_{i\leq n} x_i \leq k\right\}.$$

An important feature of (3.6) is that it is "dimension independent," that is, does not depend on N (a feature already present in the original result of Margulis). Combinatorialists have considered the problem of finding which subsets of $\{-1,1\}^N$ have the smallest boundary ∂A [defined, e.g., as the set of points of A for which $h_A(x) > 0$], but their measure of the size of ∂A is simply $P_N(\partial A)$. This formulation, however, is *not* dimension independent. In particular, the sets $A_{n,n/2}$, for $n \leq N$, play essentially the same role with respect to (3.6), and for each of them both sides of (3.6) are of the same order. However, the size of their boundaries, when measured with the "dimension dependent" quantity $P_N(\partial A)$ is very different, and only $A_{N,N/2}$ has a boundary of the smallest possible order among all sets of measure about $\frac{1}{2}$. This matter of independence of dimension will be a crucial feature of the result of Section 5, where it will be discussed again.

4. Martingales. The martingale method has been important in exploring concentration of measure in cases that are not accessible to the rearrangement methods described in the previous section. It is elegant, robust and simple. Even when rearrangement could be used, the martingale method sometimes gives comparable results in a much simpler fashion.

In contrast with the approach of the previous sections, which concentrate on "enlargements" of large sets, the martingale method deals directly with functions. The basic idea is that if $(\Sigma_i)_{i\leq n}$ is an increasing sequence of

σ-algebras, such that Σ_0 is trivial, and if f is Σ_n measurable, then

$$f - Ef = \sum_{1\le i\le n} d_i, \tag{4.1}$$

where $d_i = E(f|\Sigma_i) - E(f|\Sigma_{i-1})$. Thus (d_i) is a martingale difference sequence, that is, d_i is Σ_i-measurable, and $E(d_i|\Sigma_{i-1}) = 0$. The next step is to get bounds on

$$P\left(\left|\sum_{i\le n} d_i\right| \ge t\right). \tag{4.2}$$

This could be the time to observe that the martingale will give bounds for

$$P(|f - Ef| \ge t),$$

in contrast with (2.4), which involves M_f. In practice, it is easier to deal with Ef rather than M_f. However, it must be pointed out that, under (2.4),

$$|M_f - Ef| \le E|f - M_f| \le 2\int_0^\infty \alpha(P,u)\,du,$$

so that (2.4) implies

$$P\left(|f - Ef| \ge t + 2\int_0^\infty \alpha(P,u)\,du\right) \le 2\alpha(P,t)$$

and the occurrence of M_f in (2.4) is only a secondary nuisance when $\alpha(P,t)$ is very small.

While there is an extensive and deep theory of martingale inequalities, the inequalities required to bound (4.2) are simple martingale adaptations of classical exponential inequalities for sums of independent random variables. Namely, one bounds $E\exp(\lambda\sum_{i\le n} d_i)$ in order to use Chebyshev's exponential inequality

$$P(Z \ge t) \le \inf_\lambda \big(\exp(-\lambda t) E\exp\lambda Z\big). \tag{4.3}$$

To do this, we observe that

$$\begin{aligned} E\left(\exp\lambda\sum_{i\le n} d_i\right) &= E\left(\left(\exp\lambda\sum_{i\le n-1} d_i\right) E(\exp\lambda d_n|\Sigma_{n-1})\right) \\ &\le E\left(\exp\lambda\sum_{i\le n-1} d_i\right) \big\|E(\exp\lambda d_n|\Sigma_{n-1})\big\|_\infty, \end{aligned} \tag{4.4}$$

so that, by iteration,

$$E\exp\lambda\sum_{i\le n} d_i \le \prod_{1\le i\le n} \big\|E(\exp\lambda d_i|\Sigma_{i-1})\big\|_\infty. \tag{4.5}$$

The key of the success of this method lies in an efficient control of d_i. Probably the most important case is when one controls $\|d_i\|_\infty$. In that case, it is a simple matter to show that

$$\big\|E(\exp\lambda d_i|\Sigma_{i-1})\big\|_\infty \le \exp\frac{\lambda}{2}\|d_i\|_\infty^2,$$

which, when combined with (4.3) and (4.5), yields

$$P\left(\left|\sum_{i\le n} d_i\right| \ge t\right) \le 2\exp\left(-\frac{2t^2}{\sum_{i\le n}\|d_i\|_\infty^2}\right), \tag{4.6}$$

a result usually referred to as Azuma's inequality. In the case where $d_i = a_i\varepsilon_i$ [$(\varepsilon_i)_{i\le n}$ are independent Bernoulli random variables], (4.6) specializes to the so-called sub-Gaussian inequality

$$P\left(\left|\sum_{i\le n} a_i\varepsilon_i\right| \ge t\right) \le 2\exp\left(-\frac{t^2}{2\sum_{i\le n} a_i^2}\right), \tag{4.7}$$

a very important fact that contains (1.2) as a special case.

The use of martingales in the spirit above was apparently first made by Yurinskii [33] in the case $f = \|\sum_{i\le n} Y_i\|$, where Y_i are independent Banach space-valued r.v.'s. In this case, taking for Σ_i the σ-algebra generated by $Y_1,\dots,Y_i$, the key observation is that d_i is estimated by

$$d_i \le \|Y_i\| + E(\|Y_i\||\Sigma_{i-1}).$$

An important step was performed by Maurey [13], who discovered how to use (4.5) in a situation where neither the choice of Σ_i nor the control of d_i is obvious. The generality of the method was understood by Schechtman [21]. It yields concentration of measure in several important situations (cf. Chapter 1 of the beautiful book [17]).

In more applied fields, (4.6) was used independently by Shamir and Spencer [22] in studying the chromatic number of random graphs, by Rhee and Talagrand [20] in studying stochastic bin packing and the stochastic traveling salesman problem and later by Pastur and Shcherbina [19] in statistical mechanics. Since then, it has literally swept the world (see, e.g., [14]).

For all its qualities, the martingale method has a great drawback: it does not seem to yield results of optimal order in several key situations. In particular, it seems unable to obtain even a weak version of concentration of measure phenomenon in Gaussian space, as described in Section 3, and does not allow the main inequalities of the present paper to be obtained. For this reason, a new method needed to be invented. It will be explained and demonstrated in the rest of the paper.

5. Approximation by one point. In this section we will prove (2.6). The reason for the title of the section is that (2.6) means that when $P(A) \ge \frac{1}{2}$ most points of Ω^N belong to A_t for t not too large, which in turn means they can be well approximated by at least one point of A.

Inequality (2.6) can (essentially) be obtained using (4.6). A special case, with identical proof, is obtained in [17]. In fact, given a set A with $P(A) \ge \frac{1}{2}$, it suffices to apply (4.6) to the function $f(x) = d(x, A)$, where d denotes the Hamming distance and where Σ_i is generated by the first i coordinates. Then

one can show that $|d_i| \le 2$, so that by (4.6),

$$P(|f - Ef| \ge t) \le 2\exp\left(-\frac{t^2}{2N}\right).$$

Now, when $t = Ef$, the left-hand side is at least $\frac{1}{2}$, since $P(f \le 0) = P(A) \ge \frac{1}{2}$ and we get $t = Ef \le (2N \log 4)^{1/2}$ so that

$$P\left(f \ge t + (2N \log 4)^{1/2}\right) \le 2\exp\left(-\frac{t^2}{2N}\right),$$

a weak form of (2.6) that is of comparable strength.

Somewhat weaker statements than (2.6) were also discovered independently through a completely different approach in information theory; see, for example, [11]. The reason for which we choose (2.6) to explain our basic approach is simply that because the meaning of what we try to prove is easy to understand, the reader should be better able to concentrate on the mechanism of the proof.

The most natural way to prove (2.6) seems to be by induction over N. Thus, starting with $A \subset \Omega^N$, one should try to define sets in Ω^{N-1} to which the induction hypothesis can be applied. These sets will not necessarily be of measure greater than or equal to $\frac{1}{2}$, so that it is necessary to use as induction hypothesis a statement valid whatever the value of $P(A)$. In view of what is done for martingales, it is natural to try to bound $E\exp td(x, A)$, where $d(x, A)$ is the Hamming distance of x and A. [One might object that $d(x, A)$ need not be measurable, but measurability questions are irrelevant here and will be ignored.] It is remarkable that about the simplest bound one can expect for $E\exp td(x, A)$ turns out to be suitable.

PROPOSITION 5.1.

$$E\exp td(\cdot, A) \le \frac{1}{P(A)}\exp\frac{t^2 N}{4}. \tag{5.1}$$

In particular,

$$P(d(\cdot, A) \ge k) \le \frac{1}{P(A)}\exp\left(-\frac{k^2}{N}\right). \tag{5.2}$$

We observe that (5.2) follows from (5.1) by the Chebyshev exponential inequality, and that (5.2) implies (2.6).

The first key feature of the method of proof (which we will simply call the induction method) is that it will reduce the proof of a statement such as (5.1) concerning Ω^N to the proof of a statement concerning only functions on Ω. Most of the time the proof of this statement is easy; sometimes it is a bit harder, but its very elementary nature ensures success with sufficient effort.

The second key feature is that (as of today) the method of proof has turned out to be almost miraculously sharp in *every* situation. The reasons for this success are not entirely clear at present.

In the present case, the induction method reduces the proof of Proposition 5.1 to the following lemma.

LEMMA 5.2. *Consider a measurable function g on Ω. Then we have*

$$\int_\Omega \min\left(e^t, \frac{1}{g(\omega)}\right) d\mu(\omega) \int_\Omega g(\omega)\, d\mu(\omega) \le \exp\frac{t^2}{4}. \tag{5.3}$$

PROOF. We observe that

$$\min\left(e^t, \frac{1}{g(\omega)}\right) \le 1 + e^t(1 - g(\omega)),$$

so that the left-hand side of (5.3) is at most

$$a(1 + e^t(1 - a)),$$

where $a = \int g\, d\mu$. The maximum over a is

$$\left(\frac{e^{t/2} + e^{-t/2}}{2}\right)^2.$$

Now $(e^u + e^{-u})/2 \le e^{u^2/2}$, as is clear from a power series expansion. □

The proof of Proposition 5.1 goes by induction over N. The case $N = 1$ follows from the application of (5.3) to $g = 1_A$.

Suppose now that the result has been proved for N, and let us prove it for $N + 1$. Consider $A \subset \Omega^{N+1} = \Omega^N \times \Omega$. For $\omega \in \Omega$, we set

$$A(\omega) = \{x \in \Omega^N;\ (x, \omega) \in A\} \tag{5.4}$$

and

$$B = \{x \in \Omega^N;\ \exists\, \omega \in \Omega,\ (x, \omega) \in A\}.$$

With obvious notation, we have

$$d((x, \omega), A) \le d(x, A(\omega)).$$

Indeed, if $y \in A(\omega)$, then $(y, \omega) \in A$, and the number of coordinates where (y, ω) and (x, ω) differ is the number of coordinates where x and y differ. Thus, by the induction hypothesis, we have

$$\int_{\Omega^N} \exp\bigl(td((x, \omega), A)\bigr)\, dP(x) \le \frac{\exp(t^2 N/4)}{P(A(\omega))}. \tag{5.5}$$

We also observe that

$$d((x, \omega), A) \le d(x, B) + 1.$$

Indeed, if $y \in B$, then for some $\omega' \in \Omega$ we have $(y, \omega') \in A$, and the numbers of coordinates at which (x, ω) and (y, ω') differ is at most one more than

the number of coordinates at which x and y differ. Thus, by the induction hypothesis, we have

$$\int_{\Omega^N} \exp\bigl(td((x,\omega),A)\bigr)\,dP(x) \le \frac{\exp(t^2N/4)}{P(B)},$$

and combining this with (5.5) we get

$$\int_{\Omega^N} \exp\bigl(td((x,\omega),A)\bigr)\,dP(x) \le \exp\left(\frac{t^2N}{4}\right)\min\left(\frac{\exp(t)}{P(B)},\frac{1}{P(A(\omega))}\right).$$

Integrating with respect to ω, we have

$$\int_{\Omega^{N+1}} \exp\bigl(td((x,\omega),A)\bigr)\,dP(x)\,d\mu(\omega)$$
$$\le \exp\left(\frac{t^2N}{4}\right)\int_\Omega \min\left(\frac{e^t}{P(B)},\frac{1}{P(A(\omega))}\right)d\mu(\omega).$$

To complete the induction, it suffices to show, by Fubini's theorem, that

$$\int_\Omega \min\left(\frac{e^t}{P(B)},\frac{1}{P(A(\omega))}\right)d\mu(\omega) \le \frac{\exp(t^4/4)}{P\otimes\mu(A)} = \frac{\exp(t^4/4)}{\int_\Omega P(A(\omega))\,d\mu(\omega)}.$$

This follows from Lemma 5.2 applied to the function $g(\omega) = P(A(\omega))/P(B)$.

One way to express the fundamental difference between the induction method of Proposition 5.1 and the martingale method is that the martingale method looks "forward" while the induction method conditions with respect to the last coordinate and looks "backward," taking full advantage of the fact that the measures obtained after conditioning are identical product measures.

An interesting extension of (5.2) is obtained by allowing a term $P(A)^\alpha$, $\alpha \ge 1$, rather than $P(A)$ in (5.2); that is,

$$P(d(\cdot,A)\ge k) \le \frac{1}{P(A)^\alpha}\exp\left(-\frac{2k^2}{N}\,\frac{\alpha}{1+\alpha}\right). \tag{5.6}$$

The proof is similar to that of (5.2), but requires more calculus. The point of (5.6) is that, as $\alpha\to\infty$, we obtain almost the best possible exponent $-2k^2/N$. [The claim that this is the best possible follows from the analysis of the situation of (1.2) that will be done in the next section.]

6. Approximation by many points. In order to evaluate the result of Section 5, let us analyze a situation equivalent to that of (1.2). We take $\Omega^N = \{0,1\}^N$, provided with the uniform measure, and

$$A = \left\{x = (x_i)_{i\le N};\ \sum_{i\le N} x_i \le \frac{N}{2}\right\}$$

and we assume for simplicity that N is even.

Consider $x\in\{0,1\}^N$, $m = m(x) = \sum_{i\le N} x_i$ and assume that $m > N/2$. We claim that $d(x,A) = m - N/2$. To prove that $d(x,A)\ge m - N/2$, we observe that the function $y\to\sum_{i\le N} y_i$ is 1-Lipschitz. On the other hand, if $y\in A$ is

such that for all i, $y_i \le x_i$ (which we summarize by the statement $y \le x$), we have $d(x,y) = m - N/2$. Thus, if $k > 0$,

$$\{d(\cdot, A) \ge k\} = \left\{x;\ \sum_{i \le N} x_i \ge k + \frac{N}{2}\right\}.$$

The central limit theorem shows that, for $k = t/2\sqrt{N}$,

$$P(d(\cdot, A) \ge k) \sim \gamma_1((t,\infty)) \sim \exp -\frac{t^2}{2} \sim \exp -\frac{2k^2}{N}$$

(neglecting polynomial terms in front of the exponential), so that (5.2) is sharp (except for the factor 2 in the exponent). The previous discussion could seem redundant since the derivation of (2.7) from (2.6) already shows that (2.6) (except for the numerical factor in the exponent) is sharp. There is, however, a detail of crucial importance. The definition of A_k only means that if $x \in A_k$, there is *one* y in A for which $d(x,y) \le k$. However, in the preceding example *every* y in A with $y \le x$ satisfies $d(x,y) = k$.

Given $x \in \{0,1\}^N$, $y \in A$, it is rather natural to measure the "failure" in approximating x by y by the set $\{i \le N,\ x_i \ne y_i\}$ or, equivalently, by the element $h(x,y)$ of $\mathbb{R}^N$ such that

$$h(x,y)_i = \begin{cases} 0, & \text{if } x_i = y_i, \\ 1, & \text{if } x_i \ne y_i. \end{cases} \tag{6.1}$$

To take into account the fact that the elements y of A that approximate x well do not "miss" the same coordinates of x, it is natural to investigate how small an average of points $h(x,y)$ can be. In the present case it is natural to average over all $y \le x$, with equal weight. This average is clearly equal to

$$h(x) = \frac{m - N/2}{m} x.$$

We now observe that the Euclidean norm $\|h(x)\|_2$ of $h(x)$ satisfies

$$\|h(x)\|_2 = \left(m - \frac{N}{2}\right)\frac{1}{\sqrt{m}} \approx \left(m - \frac{N}{2}\right)\sqrt{\frac{2}{N}}$$

since $m(x) \sim N/2$ (with overwhelming probability). Now (1.2) implies that

$$P\left(\left|m - \frac{N}{2}\right| \ge t\right) \le 2\exp(-2t^2),$$

so we get that (essentially)

$$P(\|h(x)\|_2 \ge t) \le \exp(-t^2).$$

Quite remarkably, the dimension N has disappeared from this formula. Well, maybe there is some kind of coincidence there, so let us now investigate a more general example, where Ω^N is provided with the probability P such that the law of the coordinates is independent and has expectation $p, 0 < p < 1$. In jargon,

$$P = \big((1-p)\delta_0 + p\delta_1\big)^{\otimes N}.$$

Assume again for simplicity that pN is an integer and that $p \leq \frac{1}{2}$, and define

$$A = \left\{ x = (x_i)_{i \leq N};\ \sum_{i \leq N} x_i \leq pN \right\}.$$

For $x \in \{0,1\}^N$, $m = \sum_{i \leq N} x_i$, we again have $d(y, A) = m - pN$. We should observe that (1.2) is now very inaccurate. Indeed, by the classical bounds on the tails of the binomial law [5] we have something like

$$P(m - pN \geq t) \leq \exp\left(- \frac{t^2}{2p(1-p)N + \text{smaller term}} \right) \tag{6.2}$$

(for $t \leq 2pN$), which is much better than (5.2) as $p \to 0$.

On the other hand, proceeding as in the case $p = \frac{1}{2}$, we get

$$h(x) = \frac{m - Np}{m} x,$$

so that

$$\|h(x)\|_2 = (m - Np)\sqrt{m} \simeq (m - Np)\sqrt{\frac{p}{N}},$$

and combining with (6.2) yields

$$P(\|h(x)\|_2 \geq t) \leq \exp\left(- \frac{t^2}{2(1-p)} \right) \leq \exp\left(- \frac{t^2}{2} \right). \tag{6.3}$$

Quite remarkably, not only N, but also p has vanished from this inequality: it can no longer be an accident, but only a special case of a general fact.

Consider now a probability space Ω. For $x, y \in \Omega^N$, we define $h(x, y) \in \mathbb{R}^N$ by

$$h(x, y)_i = \begin{cases} 1, & \text{if } x_i \neq y_i, \\ 0, & \text{if } x_i = y_i. \end{cases}$$

For a subset A of Ω^N, we define

$$U_A'(x) = \{ h(x, y);\ y \in A \} \subset \mathbb{R}^N.$$

Define $V_A'(x)$ as the convex hull of $U_A'(x)$, and define $f(A, x)$ as the Euclidean distance from zero to $V_A'(x)$. Thus $f(A, x)$ measures "how far x is from A."

Consider a product measure P on Ω^N.

THEOREM 6.1. *We have*

$$\int \exp \frac{1}{4} f^2(A, x)\, dP(x) \leq \frac{1}{P(A)}. \tag{6.4}$$

In particular,

$$P\big(f(A, \cdot) \geq t\big) \leq \frac{1}{P(A)} \exp\left(\frac{-t^2}{4} \right). \tag{6.5}$$

Compared with (6.3), we observe a loss of a factor of 2 in the exponent. This loss can, however, be almost recovered when one replaces in (6.4) the term $P(A)^{-1}$ by $P(A)^{-\alpha}$ [as in (5.6)].

Theorem 6.1 shares several important features with Theorem 3.1. [Somehow I feel that when $\Omega^N = \{0,1\}^N$, provided with the uniform probability, Theorem 6.1 is to Theorem 3.1 what (3.3) is to (3.1), although I do not know how to make this idea precise.] The most important feature is that it is dimension independent so that (in contrast to Proposition 5.1) it is useful to study (e.g.) infinite series.

The key to the proof of Theorem 6.1 is the following lemma. The proof is elementary calculus and need not be reproduced here.

LEMMA 6.2. *Consider* $0 \le r \le 1$. *Then*

$$\inf_{0\le\lambda\le 1} r^{-\lambda} \exp\frac{(1-\lambda)^2}{4} \le 2 - r. \tag{6.6}$$

Before we start the proof of Theorem 6.1, we need an equivalent way to define $f(A,x)$. This way is less transparent, but technically more convenient. We set

$$\begin{aligned} U_A(x) &= \big\{(s_i)_{i\le N} \in \{0,1\}^N;\ \exists\, y \in A,\ s_i = 0 \ \Rightarrow\ x_i = y_i\big\} \\ &= \big\{(s_i)_{i\le N} \in \{0,1\}^N;\ \exists\, y \in A,\ \forall\, i \le N,\ s_i \ge h(x,y)_i\big\}. \end{aligned}$$

For convenience, if $s_i \ge h(x,y)_i$, for each $i \le N$, we say that y *witnesses* that $s \in U_A(x)$. Thus, viewing $U_A(x)$ as a subset of $\mathbb{R}^N$, we have $U_A(x) \supset U_A'(x)$. We denote by $V_A(x)$ the convex hull of $U_A(x)$. It should be clear that

$$\forall\, z \in V_A(x), \qquad \exists\, z' \in V_A'(x), \qquad \forall\, i \le N, \qquad z_i \ge z_i',$$

so that $f(A,x)$ is also the distance from 0 to $V_A(x)$.

We now prove Theorem 6.1 by induction upon N. We leave to the reader the easy case $N = 1$. For the induction step from N to $N+1$, consider a subset A of Ω^{N+1} and its projection B on Ω^N. For $\omega \in \Omega$, we set as usual

$$A(\omega) = \{x \in \Omega^N;\ (x,\omega) \in A\}.$$

Consider $x \in \Omega^N$, $\omega \in \Omega$ and $z = (x,\omega)$. The basic observation is that

$$\begin{aligned} s \in U_{A(\omega)}(x) \quad &\Rightarrow \quad (s,0) \in U_A(z), \\ t \in U_B(x) \quad &\Rightarrow \quad (t,1) \in U_A(x). \end{aligned}$$

For the first claim, if $y \in A(\omega)$ witnesses that $s \in U_{A(\omega)}(x)$, then $(y,\omega) \in A$ and witnesses that $(s,0) \in U_A(z)$. For the second claim, if $y \in B$ witnesses that $t \in U_B(x)$, then for some ω' we have $(y,\omega') \in A$, and this point witnesses that $(t,1) \in U_A(x)$. Thus, for $s \in V_{A(\omega)}(x)$ and $t \in V_B(x)$, $0 \le \lambda \le 1$, we have $(\lambda s + (1-\lambda)t, 1-\lambda) \in V_A(z)$. The convexity of the function $u \to u^2$ shows that

$$f^2(A,z) \le (1-\lambda)^2 + \lambda f^2(A(\omega),x) + (1-\lambda) f^2(B,x). \tag{6.7}$$

The main trick of the proof is to resist the temptation to optimize now over λ. By Hölder's inequality and the induction hypothesis, we have

$$\begin{aligned}
&\int \exp \frac{1}{4} f^2(A,(x,\omega))\, dP(x) \\
&\quad \le \exp \frac{1}{4}(1-\lambda)^2 \left(\int_{\Omega^N} \exp \frac{1}{4} f^2(A(\omega),x)\, dP(x) \right)^{\lambda} \\
&\qquad \times \left(\int_{\Omega^N} \exp \frac{1}{4} f^2(B,x)\, dP(x) \right)^{1-\lambda} \\
&\quad \le \exp \frac{1}{4}(1-\lambda)^2 \left(\frac{1}{P(A(\omega))} \right)^{\lambda} \left(\frac{1}{P(B)} \right)^{1-\lambda} \\
&\quad = \frac{1}{P(B)} \exp \frac{1}{4}(1-\lambda)^2 \left(\frac{P(A(\omega))}{P(B)} \right)^{-\lambda}.
\end{aligned}$$

This inequality holds for all $0 \le \lambda \le 1$. Using (6.6) with $r = P(A(\omega))/P(B) \le 1$, we get

$$\int_{\Omega^N} \exp \frac{1}{4} f^2(A,(x,\omega))\, dP(x) \le \frac{1}{P(B)} \left(2 - \frac{P(A(\omega))}{P(B)} \right).$$

Integrating with respect to ω and using Fubini's theorem yields

$$\begin{aligned}
\int \exp \frac{1}{4} f^2(A,\cdot)\, d(P \otimes \mu) &\le \frac{1}{P(B)} \left(2 - \frac{P \otimes \mu(A)}{P(B)} \right) \\
&\le \frac{1}{P \otimes \mu(A)},
\end{aligned}$$

since $x(2-x) \le 1$ for all x real.

While Theorem 6.1 turns out to be a principle of considerable power, it takes some effort to realize this. One efficient way to use Theorem 6.1 is through the following observation.

LEMMA 6.3. *Consider $x \in \Omega^N$. Then given* any *sequence $(\alpha_i)_{i \le N}$, we can find y in A such that*

$$\sum_{i \le N} \{\alpha_i;\ x_i \ne y_i\} \le f(A,x) \sqrt{\sum_{i \le N} \alpha_i^2}. \tag{6.8}$$

PROOF. The linear functional $\overline{\alpha}$: $s \to \sum_{i \le N} \alpha_i s_i$ on $\mathbb{R}^N$, provided with the Euclidean norm, has a norm $\sqrt{\sum_{i \le N} \alpha_i^2}$. Since $V_A'(x)$ contains a point at distance $f(A,x)$ to the origin, the infimum of $\overline{\alpha}$ on $V_A'(x)$ is at most $f(A,x)\sqrt{\sum_{i \le N} \alpha_i^2}$, but since $V_A'(x)$ is the convex hull of $U_A'(x)$, the infimum of $\overline{\alpha}$ on $U_A'(x)$ is also at most $f(A,x)\sqrt{\sum_{i \le N} \alpha_i^2}$, which is the statement of the lemma. □

Theorem 6.1 has proved efficient in stochastic combinatorial optimization, so we describe a typical application. Consider a sequence $X_1, \dots, X_N$ of independent r.v.'s, uniformly distributed over $[0,1]$. We are interested in $L_N = L_N(X_1, \dots, X_N)$, where $L_N(X_1, \dots, X_N)$ denotes the longest increasing subsequence of the sequence $X_1, \dots, X_N$ of numbers. To reduce to the setting of sets in product spaces, we consider $\Omega = [0,1]$ and, for $x = (x_i)_{i \le N} \in \Omega^N$, we set $L(x) = L_N(x_1, \dots, x_N)$.

For $a > 0$, we set

$$A(a) = \{x \in \Omega^N;\ L_N(x) \le a\}.$$

The basic observation follows:

LEMMA 6.4. *For all $x \in \Omega^N$, we have*

$$a \ge L_N(x) - f(A(a), x)\sqrt{L_N(x)}. \tag{6.9}$$

In particular,

$$L_N(x) \ge a + v \quad \Rightarrow \quad f(A(a), x) \ge \frac{v}{\sqrt{a+v}}. \tag{6.10}$$

PROOF. For simplicity, we write $b = L_N(x)$. By definition, we can find a subset I of $\{1, \dots, N\}$ of cardinality b such that if $i, j \in I$, $i < j$, then $x_i < x_j$. By Lemma 6.3 (taking $\alpha_i = 1$, if $i \in I$, and $\alpha_i = 0$, otherwise), there exists $y \in A(a)$ such that $\operatorname{card} J \le f(A(a), x)\sqrt{b}$, where $J = \{i \in I;\ y_i \ne x_i\}$. Thus $(x_i)_{i \in I \setminus J}$ is an increasing subsequence of y. Since $y \in A(a)$, we have $\operatorname{card}(I \setminus J) \le a$, which proves (6.9).

To prove (6.10), we observe that by (6.9) we have

$$f(A(a), x) \ge \frac{L_N(x) - a}{\sqrt{L_N(x)}}$$

and that the function $u \to (u-a)/\sqrt{u}$ increases for $u \ge a$. □

We denote by M $(= M_N)$ a median of L_N.

THEOREM 6.5. *For all $u > 0$, we have*

$$P(L_N \ge M + u) \le 2\exp\left(-\frac{u^2}{4(M+u)}\right), \tag{6.11}$$

$$P(L_N \le M - u) \le 2\exp\left(-\frac{u^2}{4M}\right). \tag{6.12}$$

PROOF. To prove (6.11), we combine (6.12) with $M = a$ and (6.5). To prove (6.12), we use (6.10) with $a = M - u, v = u$, to see that

$$L_N(x) \ge M \quad \Rightarrow \quad f(A(M-u), x) \ge \frac{u}{\sqrt{M}},$$

so that

$$P\left(f(A(M-u),x) \geq \frac{u}{\sqrt{M}}\right) \geq \frac{1}{2}. \tag{6.13}$$

On the other hand, by (6.5),

$$P\left(f(A(M-u),x) \geq \frac{u}{\sqrt{M}}\right) \leq \frac{1}{P(A(M-u))} \exp\left(-\frac{u^2}{4M}\right). \tag{6.14}$$

Comparing (6.13) and (6.14) gives the required bound on $P(A(M-u))$. □

It is known (and very easy to see) that M_N is of order $\sqrt{N}$, so that Theorem 6.4 proves that the fluctuations of L_N are not larger than $N^{1/4}$. Simulation [18] suggests, however, that the correct order of magnitude is smaller. Such a phenomenon cannot occur from a deficiency of Theorem 6.1, but rather from the specifics of the situation. We would like to suggest a plausible explanation of what happens.

We conjecture that (in most situations) a random sequence $(X_1, \ldots, X_N)$ has many subsequences of (nearly) maximal length. To see the relevance of this, let us go back to the proof of (6.9). Consider $b \leq L_N(x)$. Consider the family $\mathcal{J}$ of subsets I of $\{1, \ldots, N\}$ of cardinality b such that $i, j \in I$, $i < j$ implies $x_i < x_j$. Consider the family $\mathcal{H}$ of functions on $\{1, \ldots, N\}$ that consists of the indicators of sets of $\mathcal{J}$. Consider an element $(\alpha_i)_{i \leq N}$ in the convex hull of $\mathcal{H}$ and let $\sigma = \left(\sum_{i \leq N} \alpha_i^2\right)^{1/2}$. When the family $\mathcal{J}$ is "rich," we can expect that there is an averaging-out effect and that the sequence $(\alpha_i)_{i \leq N}$ can be chosen such that $\sigma^2 \ll b$. Using Lemma 6.3 we can find y in A with

$$\sum_{i \leq N} \{\alpha_i;\ x_i \neq y_i\} \leq \sigma f(A(a), x).$$

Thus, we can find I in $\mathcal{J}$ such that

$$\operatorname{card}\{i \in I;\ x_i \neq y_i\} \leq \sigma f(A(a), x).$$

As in the proof of (6.9), this shows that $b - \sigma f(A(a), x) \geq a$. Thus, if b is close to $L(x)$ and $\sigma^2 \ll b$, this allows us to improve upon (6.9). Deciding whether the phenomenon described above occurs or not is unrelated to the methods of the present paper and would certainly require a better understanding of the specifics of random sequences.

The reader must have observed that in Lemma 6.4 we do not use the full power of Lemma 6.3; rather, instead of using (6.8) for all sequences of numbers (α_i), we used it only for sequences of 0's and 1's. It seems reasonable to assert that Theorem 6.4 uses Theorem 6.1 at the very limit of its area of competence. This can also be seen by the fact that martingale methods can prove an inequality almost as good as (6.11) and (6.12) [2]. By contrast, martingale methods seem powerless to approach the applications where Theorem 6.1 is used at full power, such as in the following theorem.

THEOREM 6.6. *Consider a real-valued function f defined on $[-1,1]^N$. We assume that, for each real number a,*

$$\text{the set } \{f \le a\} \text{ is convex.} \tag{6.15}$$

Consider a convex set $B \subset [-1,1]^N$, consider $\sigma > 0$ and assume that the restriction of f to B has a Lipschitz constant at most σ; that is,

$$\forall\, x, y \in B, \qquad |f(x) - f(y)| \le \sigma \|x - y\|, \tag{6.16}$$

where $\|x\|$ denotes the Euclidean norm of x.

Consider independent random variables $(X_i)_{i \le N}$ valued in $[-1,1]$, and consider the random variable

$$h = f(X_1, \dots, X_N).$$

Then, if M is a median of h, we have, for all $t > 0$, that

$$P(|h - M| \ge t) \le 4b + \frac{4}{1-2b} \exp\left(-\frac{t^2}{16\sigma^2}\right), \tag{6.17}$$

where we assume

$$b = P\big((X_1, \dots, X_N) \notin B\big) < \tfrac{1}{2}.$$

Certainly the reader should first consider the special case $B = [-1,1]^N$, where $b = 0$ and where (6.17) reads

$$P(|h - M| \ge t) \le 4 \exp\left(-\frac{t^2}{16\sigma^2}\right). \tag{6.18}$$

To understand this inequality better, we will compare it with the Gaussian case. Let us now assume that f is defined on all $\mathbb{R}^N$ and has Lipshitz constant σ. Set $h' = f(Y_1, \dots, Y_N)$, where the sequence $Y_1, \dots, Y_N$ is independent standard normal. Combining (2.5) and (2.9), we have

$$P(|h' - M'| \ge t) \le \exp\left(-\frac{t^2}{2\sigma^2}\right), \tag{6.19}$$

where M' is of course a median of h'. Thus, what (6.18) does is to prove an inequality similar to (6.19) for random variables that need no longer be Gaussian (but rather are bounded) and this under only the pretty mild restriction (6.15).

PROOF OF THEOREM 6.6. Let us fix $a \in \mathbb{R}$, and consider the set $A(a) = \{f \le a\} \cap B$. The key observation is that, for any x in $[-1,1]^N$, we have

$$d(x, A(a)) \le 2f(A(a), x). \tag{6.20}$$

Indeed, if $y \in A(a)$, we have

$$\forall\, i \le N, \qquad |x_i - y_i| \le 2h(x,y)_i$$

[where $h(x, y)_i$ is defined in (6.1)] because the left-hand side is at most 2 and is zero when the right-hand side is not 2. Thus, for any points $(y^k)_{k\le M}$ of $A(a)$ and convex coefficients $(\alpha_k)_{k\le M}$, we have, for each $i \le N$, that

$$\left| x_i - \sum_k \alpha_k y_i^k \right| \le 2 \sum_k \alpha_k h(x, y^k)_i,$$

so that, since $A(a)$ is convex,

$$d(x, A(a)) \le \left\| x - \sum_k \alpha_k y^k \right\| \le 2 \left(\sum_{i \le N} \left(\sum_k \alpha_k h(x, y^k)_i \right)^2 \right)^{1/2},$$

from which (6.20) follows by definition of V_A'.

Now, if $x \in B$, it follows from (6.16) that

$$f(x) \le a + \sigma d(x, A(a)) \le a + 2\sigma f(A(a), x).$$

Thus, if we denote by P the law $(X_1, \dots, X_N)$ on $\Omega^N = [-1, 1]^N$, (6.5) implies

$$P(f \ge a + t) \le b + \frac{1}{P(A(a))} \exp\left(-\frac{t^2}{16\sigma^2} \right).$$

Taking $a = M$, we get $P(A(a)) \ge \frac{1}{2} - b$, so that

$$P(f \ge M + t) \le b + \frac{1}{\frac{1}{2} - b} \exp\left(-\frac{t^2}{16\sigma^2} \right).$$

Taking $a + t = M$, we get

$$\frac{1}{2} \le b + \frac{1}{P(A(M - t))} \exp\left(-\frac{t^2}{16\sigma^2} \right),$$

so that

$$P(A(M - t)) = P(f \le M - t) \le 2b + 2 \exp\left(-\frac{t^2}{16\sigma^2} \right). \qquad \square$$

COMMENTS. (i) Certainly the reader has observed the similarity of this proof with the proof of Theorem 6.5.

(ii) We have not been very cautious with the coefficients of b. This is not needed because in applications b is extremely small.

Here is an important corollary.

THEOREM 6.7. *Consider independent (real) random variables $(X_i)_{i\le N}$ valued in $[-1, 1]$ and vectors $(v_i)_{i\le N}$ in a Banach space Y. Define*

$$\sigma^2 = \sup \left\{ \sum_{i \le N} y^*(v_i)^2;\ y^* \in Y^*,\ \|y^*\| \le 1 \right\},$$

where Y^* is the dual of Y. Then, if M denotes a median of $\|\sum_{i\le N} X_i v_i\|$, we have

$$P\left(\left|\left\|\sum_{i\le N} X_i v_i\right\| - M\right| \ge t\right) \le 4\exp\left(-\frac{t^2}{16\sigma^2}\right). \tag{6.21}$$

REMARK. The most important case is where the r.v.'s X_i are Bernoulli, that is, $P(X_i = 1) = P(X_i = -1) = \frac{1}{2}$.

PROOF. We observe that $\|\sum_{i\le N} X_i v_i\| = f(X_1, \dots, X_N)$, where, for $x = (x_i)_{i\le N}$ in $\mathbb{R}^N$, we set

$$f(x) = \left\|\sum_{i\le N} x_i v_i\right\|.$$

By the Hahn–Banach theorem,

$$\left\|\sum_{i\le N} x_i v_i\right\| = \sup\left\{y^*\left(\sum_{i\le N} x_i v_i\right);\ y^* \in Y^*;\ \|y^*\| \le 1\right\}.$$

Now, by Cauchy–Schwarz,

$$y^*\left(\sum_{i\le N} x_i v_i\right) = \sum_{i\le N} x_i y^*(v_i) \le \left(\sum_{i\le N} x_i^2\right)^{1/2}\left(\sum_{i\le N} y^*(v_i)^2\right)^{1/2} \le \sigma\|x\|.$$

Thus, by the triangle inequality,

$$|f(x) - f(y)| \le f(x - y) \le \sigma\|x - y\|$$

and thus (6.21) is a specialization of (6.18). □

We now give another application of Theorem 6.6 to the Hopfield model of associative memory [6]. Consider two integers M, N. For $x = (x_{i,k})_{i\le N,\ k\le M} \in \mathbb{R}^{MN}$ and for $\varepsilon = (\varepsilon_i)_{i\le N} \in \{-1, 1\}^N$, we set

$$H(x, \varepsilon) = \frac{1}{2N}\sum_{k\le M}\left(\sum_{i\le N} x_{i,k}\varepsilon_i\right)^2$$

(the factor $1/2N$ is customary but unimportant).

Given a subset A of $\{-1, 1\}^N$, we set

$$f(x) = f_N(x) = \frac{1}{\beta}\log\left(\sum_{\varepsilon\in A}\exp\beta H(x, \varepsilon)\right).$$

The quantity of interest is the random variable $h_N = f_N(\eta)$, when $\eta = (\eta_{i,k})_{i\le N,\ k\le M}$ and when $(\eta_{i,k})_{i\le N,\ k\le M}$ are independent Bernoulli r.v.'s $[P(\eta_{i,k} = 1) = \frac{1}{2} = P(\eta_{i,k} = -1)]$. In the case $A = \{-1, 1\}^N$, h_N is the free energy of the Hopfield model (at temperature $T = 1/\beta$) and its study is extremely difficult, yet one has the following general result.

THEOREM 6.8. *Denoting by m_N a median of h_N, for some universal constant K and all $0 \le t \le (N+M)$, we have*

$$P(|h_N - m_N| \ge t) \le 12 \exp\left(-\frac{t^2}{K(N+M)}\right).$$

PROOF. The proof relies on Theorem 6.6, applied to the function f on $[-1,1]^{NM}$. It is not clear whether f is convex, but certainly $\exp \beta f$ is convex, and this implies (6.15). Consider a parameter L and set

$$B = \left\{x \in [-1,1]^{NM};\ \forall\ \varepsilon \in A,\ H(x,\varepsilon) \le L\right\}$$

so that B is convex. Consider now x and y in B. We try to prove (6.16). We observe that, given $\varepsilon \in A$,

$$2N\big(H(x,\varepsilon) - H(y,\varepsilon)\big) = \sum_{k\le M}\left(\sum_{i\le N}(x_{i,k} - y_{i,k})\varepsilon_i\right)\left(\sum_{i\le N} x_{i,k}\varepsilon_i + \sum_{i\le N} y_{i,k}\varepsilon_i\right).$$

Thus, by Cauchy–Schwarz,

$$|H(x,\varepsilon) - H(y,\varepsilon)| \le \frac{1}{2N} UV,$$

where

$$U^2 = \sum_{k\le M}\left(\sum_{i\le N}(x_{i,k} - y_{i,k})\varepsilon_i\right)^2,$$

$$V^2 = \sum_{k\le M}\left(\sum_{i\le N} x_{i,k}\varepsilon_i + \sum_{i\le N} y_{i,k}\varepsilon_i\right)^2.$$

Using the inequality $(a+b)^2 \le 2(a^2+b^2)$, we see that

$$V^2 \le 4N\big(H(x,\varepsilon) + H(y,\varepsilon)\big) \le 8NL.$$

Using Cauchy–Schwarz, we see that

$$U^2 \le N \sum_{\substack{k\le M\\ i\le N}} (x_{i,k} - y_{i,k})^2 = N\|x-y\|^2,$$

so that, finally, for each ε in A we have

$$|H(x,\varepsilon) - H(y,\varepsilon)| \le \|x-y\|\sqrt{2L}.$$

It is then very simple to see that this implies

$$|f(x) - f(y)| \le \sqrt{2L}\|x-y\|.$$

Thus, by (6.17) we have

$$P(|h_N - m_N| \ge t) \le 4b + \frac{4}{1-2b}\exp\left(-\frac{t^2}{32L}\right), \tag{6.22}$$

where $b = P(\eta \notin B)$. To choose L, we note that by (4.7) we have, for each k,

$$E \exp \frac{1}{4N} \Big(\sum_{i \le N} \eta_{i,k} \varepsilon_i \Big)^2 \le K_1,$$

where K_1 is a universal constant, so that, by independence,

$$E \exp \tfrac{1}{2} H(\eta, \varepsilon) \le K_1^M$$

and, by Chebyshev's inequality,

$$P(H(\eta, \varepsilon) > L) \le K_1^M e^{-L/2}.$$

Thus, if $L = 4N + 2M(1 + \log K_1)$, we have

$$P(H(\eta, \varepsilon) \ge L) \le e^{-2N-M},$$

so that

$$P\big(\exists\, \varepsilon \in \{-1,1\}^N;\ H(\eta, \varepsilon) \ge L\big) \le e^{-(N+M)}.$$

Thus $b = P(\eta \notin B) \le e^{-(N+M)}$. Since $b \le \frac{1}{4}$, we get from (6.22) that

$$P(|h_N - m_N| \ge t) \le 4e^{-(M+N)} + 8 \exp\Big(- \frac{t^2}{K(N+M)} \Big),$$

where K is a universal constant. The result follows. □

7. Approximation by very many points. Let us go back to the discussion at the beginning of Section 6. Given $x \in \{0,1\}^N$, what we have used is that the functions $h(x, y)$ ($y \in A$) have a small average $h(x) = (1 - N/2m)x$, where $m = \sum_{i \le N} x_i$. The existence of this average, however, does not fully reflect the multitude of points of A that approximate x. Indeed, to obtain such an average it would essentially suffice to have about $m/(m - N/2)$ elements $y \le x$ in A such that the sets $\{i;\ x_i \ne y_i\}$ are disjoint.

The result we will present in this section is a considerable strengthening of Theorem 6.1; it, however, requires a further leap into abstraction.

The basic idea is identical to that of Theorem 6.1. Given x, y in Ω^N, we associate an object $\nu_{x,y}$ and we express that x is close to A if there is a convex combination of the objects $\nu_{x,y}$, $y \in A$, that is "small." In Section 6, the object $\nu_{x,y}$ was the indicator of the set $\{i;\ x_i \ne y_i\}$. In the present section, we use a higher dimensional object so that it is harder to make averages of objects $\nu_{x,y}$ "small" and so that, in turn, the existence of such small averages yields stronger information.

Consider a number $0 < \theta < 1$ and the probability measure

$$\nu = ((1 - \theta)\delta_0 + \theta \delta_1)^{\otimes N}$$

on $\{0,1\}^N$. Thus ν is the law of an independent sequence $(\eta_i)_{i \le N}$ with $E\eta_i = \theta$, $\eta_i \in \{0,1\}$. Given x, y in Ω^N, we consider the measure $\nu_{x,y}$ on $\{0,1\}^N$ such

that $\nu_{x,y}$ is the image of ν by the map T of $\{0,1\}^n$ that "flips the coordinates" i for which $x_i \neq y_i$, that is,

$$T(u)_i = \begin{cases} u_i, & \text{if } x_i = y_i, \\ 1 - u_i, & \text{if } x_i \neq y_i. \end{cases}$$

In other words, $\nu_{x,y}$ is the law of an independent sequence $\eta_i \in \{0,1\}$ such that $E\eta_i = \theta$, if $x_i = y_i$, and $E\eta_i = 1 - \theta$, if $x_i \neq y_i$. Thus, if $\theta \neq \frac{1}{2}$, the more coordinates of x and y are different, the more different $\nu_{x,y}$ is from ν.

To measure how far a probability μ on $\{0,1\}^n$ is from ν, we will use the quantity

$$\int \left(\frac{d\mu}{d\nu}\right)^2 d\nu,$$

where the integral is over $\{0,1\}^N$. We observe that since $\int (d\mu/d\nu)\, d\nu = 1$, by Cauchy–Schwarz, we have $\int (d\mu/d\nu)^2\, d\nu \geq 1$.

For a subset A of Ω^N, we set

$$m(A,x) = \inf\left\{\int \left(\frac{d\mu}{d\nu}\right)^2 d\nu;\ \mu \in \operatorname{conv}\{\nu_{x,y};\ y \in A\}\right\}.$$

THEOREM 7.1 [29]. *Assume that* $\beta = |\theta - \frac{1}{2}| < \frac{1}{6}$ *and define*

$$\alpha = \frac{32\beta^2}{1 - 36\beta^2}.$$

Then for all subsets A *of* Ω^N *we have*

$$\int_{\Omega^N} m(A,x)\, dP(x) \leq \frac{1}{P(A)^\alpha}. \tag{7.1}$$

Certainly the condition $|\theta - \frac{1}{2}| < \frac{1}{6}$ looks strange. The previous result is, however, a good illustration of the wonders of the induction method. Analysis of the canonical example presented at the beginning of Section 6 allows one to show that the left-hand side of (7.1) can stay bounded when $P(A) \geq \frac{1}{2}$ independently of Ω, A, N only when $|\theta - \frac{1}{2}| \leq \frac{1}{6}$. We have no intuitive explanation to offer as the reason for this "phase transition."

As will be demonstrated later, Theorem 7.1 is in many respects a considerable strengthening of Theorem 6.1. However, it would have been hard to discover Theorem 7.1 this way, and the motivation came from the convolution problem of [24], that we recall now. Consider, on the group $G_N = \{-1,1\}^N$, the Haar measure λ and the measure

$$\nu = \{(1-\theta)\delta_{-1} + \theta\delta_1\}^{\otimes N}.$$

Consider the convolution operator T: $f \to f * \nu$ from $L^1(\lambda)$ to $L^1(\lambda)$. The conjecture means that T displays some regularization properties, as follows.

CONJECTURE 7.2. *Consider $f \in L^1(G)$, $f \geq 0$ and $\int f\, d\lambda = 1$. Then, for all $t \geq 0$, we have*

$$\lambda(\{Tf \geq t\}) \leq \frac{K}{t\sqrt{\log(e+t)}}, \tag{7.2}$$

where K is a universal constant.

The logarithmic factor in (7.2) is apparently related to the logarithmic factor in (3.6).

The idea of Theorem 7.1 was simply that $Tf(x) = \nu_x(f)$, where the probability ν_x is the translation of ν by x. Thus, if $A = \{x;\ \nu_x(f) \geq t\}$, it should help to know that for many y we have ν_y close to the set $\{\nu_x,\ x \in A\}$ (a fact whose formulation led to Theorem 7.1). We have, however, been unable to carry out the idea.

The progress that Theorem 7.1 represents over Theorem 6.1 is exemplified by the following result, where we set

$$I_k = \{\mathbf{i} = (i_1, \ldots, i_k);\ 1 \leq i_1 < i_2 < \cdots < i_k \leq N\}.$$

PROPOSITION 7.3. *Let us fix θ with $|\theta - \frac{1}{2}| < \frac{1}{6}$. Assume that $m(A,x) \leq e^t$. Then for each $k \geq 1$ and each family $(\alpha_{\mathbf{i}})_{\mathbf{i} \in I_k}$ there exists y in A such that if*

$$J_k = J_k(x,y) = \{\mathbf{i} \in I_k;\ \forall\, \ell \leq k,\ x_{i_\ell} \neq y_{i_\ell}\},$$

then

$$\sum_{\mathbf{i} \in J_k} \alpha_{\mathbf{i}} \leq C^k t^{k/2} \Big(\sum_{\mathbf{i} \in I_k} \alpha_{\mathbf{i}}^2 \Big)^{1/2},$$

where C depends on θ only.

To understand this result better, let us specialize to the case $k = 1$. Thus, given numbers $(\alpha_i)_{i \leq N}$, we can find $y \in A$ such that

$$\sum_{i \leq N} \{\alpha_i;\ x_i \neq y_i\} \leq C\sqrt{\log m(A,x)} \Big(\sum_{i \leq N} \alpha_i^2 \Big)^{1/2}. \tag{7.3}$$

To compare (7.3) with (6.8), we have to compare the set where $C\sqrt{\log m(A,\cdot)}$ is large with the set where $f(A,\cdot)$ is large. We note that from (7.1) we have

$$P\big(C\sqrt{\log m(A,\cdot)} \geq u\big) \leq \frac{1}{P(A)^\alpha} \exp\Big(-\frac{u^2}{C^2}\Big), \tag{7.4}$$

whereas from (6.4) we have

$$P\big(f(A,\cdot) \geq u\big) \leq \frac{1}{P(A)} \exp\Big(-\frac{u^2}{4}\Big). \tag{7.5}$$

Thus (6.8) and (7.3) are comparable [with the exception of worse constants in (7.4) versus (7.5)], but the conclusion of Proposition 7.3 holds for any $k \geq 1$.

8. Control by several points. Let us start by providing motivation. Suppose that we are given a sequence $(Y_i)_{i\le N}$ of nonnegative r.v.'s and that we know that

$$P\Big(\sum_{i\le N} Y_i \le M\Big) \ge \tfrac{1}{2}.$$

We attempt to find bounds for the tail probabilities $P(\sum_{i\le N} Y_i \ge t)$. The basic observation is as follows. Set $A = \{\sum_{i\le N} Y_i \le M\}$. Consider ω' and ω in A. Set

$$I = \{i \le N;\ Y_i(\omega) = Y_i(\omega')\}$$

so that

$$\sum_{i\in I} Y_i(\omega') = \sum_{i\in I} Y_i(\omega) \le M \tag{8.1}$$

by *positivity* of Y_i. Consider now $\omega^1, \dots, \omega^q$ in A and set

$$J = \{i \le N;\ \exists\, \ell \le q, Y_i(\omega') = Y_i(\omega^\ell)\}. \tag{8.2}$$

Thus, by (8.1) and the positivity of Y_i, we have

$$\sum_{i\le N} Y_i(\omega') \le qM + \sum_{i\notin J} Y_i. \tag{8.3}$$

One then hopes that if $\operatorname{card}\{i \notin J\}$ is small, we will be able to control the last term. This discussion should provide motivation for the following discussion. Consider an integer $q \ge 2$. For $x \in \Omega^N$ and subsets $A_1, \dots, A_q$ of Ω^N we define

$$\begin{aligned} &f(A_1, \dots, A_q, x) \\ &\quad = \inf\{\operatorname{card}\{i \le N;\ x_i \in \{y_i^1, \dots, y_i^q\}\}:\ y^1 \in A_1, \dots, y^q \in A_q\}. \end{aligned} \tag{8.4}$$

What we are really interested in is the case $A_1 = A_2 = \cdots = A_q$, but the proof by induction requires considering different sets. Later we will prove the following basic fact about $f(A_1, \dots, A_q, x)$.

THEOREM 8.1. *If P is a product measure, we have*

$$\int q^{f(A_1,\dots,A_q,x)}\, dP(x) \le \frac{1}{\Pi_{i\le q} P(A_i)} \tag{8.5}$$

and, in particular,

$$P\big(f(A, \dots, A, \cdot) \ge k\big) \le \frac{1}{q^k P(A)^q}. \tag{8.6}$$

Combining with (8.3) we see that if S_k denotes the sum of the largest k terms of the sequence $(Y_i)_{i\le N}$, we have

$$P\Big(\sum_{i\le N} Y_i \ge qM + t\Big) \le \frac{2^q}{q^k} + P(S_k \ge t). \tag{8.7}$$

Hopefully the last term can be controlled by classical methods, and it remains only to optimize (8.7) over the various parameters.

Certainly the previous method seems an overkill to study the tails of $\sum_{i\le N} Y_i$. Suppose, however, that we now have a function f on Ω^N and functions $(Y_i)_{i\le N}$ such that if $A = \{f \le M\}$, where M is the median of f, the following holds. Given $x \in \Omega^N$, $y^1, \dots, y^q \in A$ and

$$J = \{i \le N;\ \exists\, \ell \le q,\ x_i = y_i^\ell\},$$

then

$$f(x) \le qM + S_k, \tag{8.8}$$

where $k = \operatorname{card}\{i \le N;\ i \notin J\}$ and S_k is the sum of the k largest terms of the sequence $(Y_i(x_i))_{i\le N}$. Then

$$P(f \ge qM + t) \le \frac{2^q}{q^k} + P(S_k \ge t). \tag{8.9}$$

To give the most important example of this situation, let us consider the case where

$$f(x) = E_\varepsilon \left\| \sum_{i\le N} \varepsilon_i Z_i(x_i) \right\|,$$

for functions Z_i from Ω to a Banach space, and where $(\varepsilon_i)_{i\le N}$ are independent Bernoulli r.v.'s. The key observation is that the function

$$E_\varepsilon \left\| \sum_{i\in I} \varepsilon_i Z_i(x_i) \right\|$$

is an increasing function of I, as is seen by taking the expectation with respect to certain ε_i's inside rather than outside the norm. Thus, when $J = \bigcup_{\ell\le q} I_\ell$, by the triangle inequality we have

$$E_\varepsilon \left\| \sum_{i\in J} \varepsilon_i Z_i(x_i) \right\| \le \sum_{\ell\le q} E_\varepsilon \left\| \sum_{i\in J} \varepsilon_i Z_i(x_i) \right\|$$

and thus

$$f(x) \le \sum_{\ell\le q} E_\varepsilon \left\| \sum_{i\in J} \varepsilon_i Z_i(x_i) \right\| + \sum_{i\notin J} \|Z_i(x_i)\|.$$

This implies that (8.8) holds for $Y_i = \|Z_i\|$. An important contribution of Ledoux [7] made clear that controlling f is the main step in controlling $\|\sum_{i\le n} \varepsilon_i Z_i(x_i)\|$. This approach using inequality (8.9) has been very successful, as demonstrated in [9], and it is remarkable that its proof is now so simple.

The key fact of the proof of Theorem 8.1 is the following simple statement about functions.

LEMMA 8.2. *Consider a function g on Ω, such that $1/q \le g \le 1$. Then*

$$\int_\Omega \frac{1}{g}\, d\mu \left(\int_\Omega g\, d\mu \right)^q \le 1. \tag{8.10}$$

PROOF. Observing that $\log x \le x - 1$, to prove that $ab^q \le 1$ it suffices to show that $a + qb \le q + 1$. Thus, it suffices to show that

$$\int_\Omega \frac{1}{g}\, d\mu + q \int_\Omega g\, d\mu \le q + 1,$$

but this is obvious since $x^{-1} + qx \le q + 1$ for $q^{-1} \le x \le 1$. □

COROLLARY 8.3. *Consider functions g_i on Ω, $0 \le g_i \le 1$. Then*

$$\int_\Omega \min_{i \le q} \left(q, \frac{1}{g_i} \right) d\mu \prod_{i \le q} \int g_i\, d\mu \le 1. \tag{8.11}$$

PROOF. Set $g = (\min_{i \le q}(q, g_i^{-1}))^{-1}$, observe that $g_i \le g$ and use (8.10). □

We now prove Theorem 8.1 by induction over N. For $N = 1$, the result follows from (8.11) taking $g_i = 1_{A_i}$.

We assume now that Theorem 8.1 has been proved for N and we prove it for $N + 1$. Consider sets $A_1, \ldots, A_q$ of Ω^{N+1}. For $\omega \in \Omega$, we define the sets $A_i(\omega)$ as in (5.4) and we consider the projection B_i of A_i on Ω^N. The basic observation is that

$$f(A_1, \ldots, A_q, (x, \omega)) \le 1 + f(B_1, \ldots, B_q, x) \tag{8.12}$$

and that, whenever $j \le q$,

$$f(A_1, \ldots, A_q, (x, \omega)) \le f(C_1, \ldots, C_q, x), \tag{8.13}$$

where $C_i = B_i$ for $i \ne j$, $C_j = A_j(\omega)$.

To prove Theorem 8.1, we observe that, using (8.12) and induction hypothesis, we have

$$\int q^{f(A_1, \ldots, A_q, (x,\omega))}\, dP(x) \le q \frac{1}{\prod_{i \le q} P(B_i)}$$

while using (8.13) we get

$$\int q^{f(A_1, \ldots, A_q, (x,\omega))}\, dP(x) \le q \frac{1}{\prod_{i \le q} P(C_i)}.$$

Thus, setting $g_i(\omega) = P(A_i(\omega))/P(B_i)$, we have

$$\int q^{f(A_1, \ldots, A_q(x,\omega))}\, dP(x) \le \frac{1}{\prod_{i \le q} P(B_i)} \int \min \left(q, \min_{i \le q} \frac{1}{g_i(\omega)} \right) d\mu(\omega).$$

Using now the Fubini theorem and (8.11), we have

$$\int q^{f(A_1,\dots,A_q(x,\omega))}\,dP(x)\,d\mu(\omega) \le \frac{1}{\prod_{i\le q} P(B_i)\int g_i\,d\mu},$$

which finishes the proof since $\int g_i\,d\mu = P(A_i)/P(B_i)$.

9. Penalties. Roughly speaking, the Hamming distance measures how far x is from A by counting the smallest number of coordinates of x that cannot be captured by a point of A. Thus we get one penalty for each coordinate we miss. A natural extension of this idea is to consider a nonnegative function h on $\Omega\times\Omega$ and, for $x\in\Omega^N$, $A\subset\Omega^N$, to consider

$$f_h(A,x) = \inf\left\{\sum_{i\le N} h(x_i,y_i);\ y\in A\right\} \tag{9.1}$$

as a way to measure the "distance" from x to A.

It is reasonable to require

$$\forall\,\omega\in\Omega,\qquad h(\omega,\omega)=0. \tag{9.2}$$

Thus, the case of the Hamming distance is simply $h(x,y)=1_{\{x\neq y\}}$.

We observe that, since x, y do not play the same role, we will not require h to be symmetric. In constrast with the work of Sections 5–8 that requires *no* structure on Ω, Definition 9.1 does require a minimum of structure, namely, the existence of the function h. On the other hand, this opens the door to a theory whose complexity certainly would not have been suspected beforehand.

Certainly one needs some control on the size of h. The most obvious way to achieve this is through moment conditions on h. A typical result is as follows.

THEOREM 9.1. *Set*

$$\begin{aligned}\|h\|_\infty &= \sup\{h(x,y);\ x,y\in\Omega\},\\ \|h\|_2^2 &= \int\!\!\int_{\Omega^2} h^2(\omega,\omega')\,d\mu(\omega)\,d\mu(\omega').\end{aligned} \tag{9.3}$$

Then, for each subset A of Ω^N, we have

$$P(f_h(A,\cdot)\ge u) \le \frac{1}{P(A)}\exp\left(-\min\left(\frac{u^2}{8N\|h\|_2^2},\frac{u}{2\|h\|_\infty}\right)\right). \tag{9.4}$$

We do not know how to obtain sharp numerical constants in (9.4). Inequality (9.4) generalizes Bernstein's inequality the way Theorem 9.1 generalizes (1.3). If g is a function on Ω, setting $h(x,y)=|g(x)-g(y)|$, it is an interesting exercise to recover from (9.4) a qualitatively correct version of Bernstein's inequality (i.e., only the numerical constants are different).

It is arguable that Theorem 9.1 does not represent a truly new phenomenon. It turns out, however, that in Theorem 9.1 what matters is not really h, but rather the following functional, defined for all subsets B of Ω:

$$h(\omega,B) = \inf\{h(\omega,\omega');\ \omega'\in B\}. \tag{9.5}$$

THEOREM 9.2. *Assume that for each subset B of Ω we have*

$$\int_\Omega \exp 2h(x,B)\,d\mu(x) \le \frac{e}{\mu(B)}. \tag{9.6}$$

Then for $t \le 1$ and each subset A of Ω^N we have

$$\int_{\Omega^N} \exp t f_h(A,x)\,dP(x) \le \frac{\exp(t^2N)}{P(A)}. \tag{9.7}$$

In particular, if $u \le 2N$ we have

$$P\big(f_h(A,\cdot) \ge u\big) \le \frac{1}{P(A)} \exp\left(-\frac{u^2}{4N}\right). \tag{9.8}$$

The point of (9.6) is that taking the infimum in (9.5) has a dramatic effect and that condition (9.6) is less stringent than the control of

$$\int_{\Omega^2} \exp 2h(x,y)\,d\mu(x)\,d\mu(y)$$

one would expect would be required in order to obtain something like (9.8).

We illustrate this in the case where Ω is itself a product of m spaces and where $h(x,y) = ad(x,y)$, where d is the Hamming distance on Ω and a is a parameter. It follows from (5.1) that (9.6) holds for $a = 2m^{-1/2}$. On the other hand, if $\|h\|_2$ is given by (9.3), then for this value of a, $\|h\|_2$ is of order $\sqrt{m}$, so that there is a loss of a factor $\sqrt{m}$ in the exponent in (9.4) compared to (9.8).

To give a vivid illustration of what Theorem 9.2 can prove, consider a product space Ω^N. Consider a subset A of Ω^N, $P(A) \ge \frac{1}{2}$. Then for most elements x of Ω^N, we can find an element y of A such that the set $I = \{i \le N;\ x_i \ne y_i\}$ has a cardinal of order $\sqrt{N}$. This is the content of Proposition 5.1; but now if we view N as built from N_1 blocks of length N_2 ($N = N_1N_2$), we can moreover require that I meets only about $\sqrt{N_1}$ blocks.

One of the most interesting phenomena related to the theory of penalties occurs under a condition somewhat stronger than (9.6). However, rather than stating the most general theorem (it requires some effort to understand the hypothesis), we will only state the most important case. In that case, $\Omega = \mathbb{R}$, μ has a density $\frac{1}{2}e^{-|x|}$ with respect to Lebesgue measure, and the function h is given by

$$h(x,y) = \min(|x-y|, (x-y)^2).$$

THEOREM 9.3. *For some universal constant K and each subset A of $\mathbb{R}^N$, we have*

$$\int_{\Omega^N} \exp\left(\frac{1}{K} f_h(A,x)\right) dP(x) \le \frac{1}{P(A)}. \tag{9.9}$$

The most obviously remarkable feature of this theorem is that (9.9) does not depend on N. The depth of Theorem 9.3 can, however, better be measured

by the fact that it does constitute a kind of improvement upon what was previously known about Gaussian measure. To see this, consider the nondecreasing map φ from $\mathbb{R}$ to $\mathbb{R}$ that transforms μ into the one-dimensional Gaussian measure γ_1. It is a simple fact to see that, for some universal constant K, we have

$$(\varphi(x) - \varphi(y))^2 \le Kh(x,y). \tag{9.10}$$

On the other hand, the map ψ from $\mathbb{R}^N$ to $\mathbb{R}^N$ given by $\psi((x_i)_{i\le N}) = (\varphi(x_i))_{i\le N}$ transforms P into γ_N.

Consider now $B \subset \mathbb{R}^N$. Thus

$$\gamma_N(B) = P(\psi^{-1}(B))$$

Now, by (9.10), we have

$$d(\psi(A), \psi(x))^2 \le Kf_h(A,x),$$

where $d(B,y)$ is the Euclidean distance from B to y and thus from (9.9),

$$\int_{\Omega^N} \exp\left(\frac{1}{K}\, d\big(\psi(\psi^{-1}(B)), \psi(x)\big)^2\right) dP(x) \le \frac{1}{\gamma_N(B)},$$

so that (for a new constant K)

$$\int_{\mathbb{R}^N} \exp\left(\frac{1}{K}\, d(B,y)^2\right) d\gamma_N(y) \le \frac{1}{\gamma_N(B)}.$$

Therefore, for $t > 0$,

$$\gamma_N\big(d(B,\cdot) \ge t\big) \le \frac{1}{\gamma_N(B)} \exp\left(-\frac{t^2}{K}\right). \tag{9.11}$$

In the case $\gamma_N(B) = \frac{1}{2}$, this is a weak form of (2.9).

It turns out that, for many applications, (9.11) rather than (2.9) suffices. In particular, it is now clearly understood that (9.11) is one of the central facts that allows us to characterize continuity and boundedness of Gaussian processes [23]. The importance of Theorem 9.3 is that it allows extending these characterizations to more general processes [26].

One of the most intriguing further aspects of the theory of penalties is that the roles of x and y in the penalty function $h(x,y)$ are highly asymmetric. This is particularly apparent when the idea of penalty function is combined with the method of Section 6, a topic for which we must refer to [29].

In conclusion, we have tried to make the reader aware that there are unexpectedly subtle phenomena related to concentration of measure in product spaces. That such a rich theory should exist at all with such minimal structure is certainly remarkable, as is remarkable the width of its applications. It is not clear to me at present where the potential for future advances, if any, lies. A worthy project would be a systematic development of the "transportation method" that very recently arose from the work of Marton [12]. This method is a potentially serious competitor to the induction method presented here. It allows, in some cases, an easier computation of the best constants and an easier approach to Theorem 9.1 [30], but whether it can lead to genuinely new

results in the independent case is unclear at present. In a different direction, an obvious research question is whether there exists at all a usable theory beyond the case of product measures; see, for example, [29] for the case of the symmetric group (that resembles a product) and of [12] for certain Markov chains.

Acknowledgment. The author is indebted to Michel Ledoux and Gilles Godefroy for many useful comments.

Note added in proof. After this paper was written, further progress was made, in particular on the material of Section 7 [31].

REFERENCES

[1] AMIR, D. and MILMAN, V. D. (1980). Unconditional and symmetric sets in n-dimensional normed spaces. *Israel J. Math.* **37** 3–20.

[2] BOLLABÁS, B. and BRIGHTWELL, G. (1992). The height of a random partial order: concentration of measure. *Ann. Appl. Probab.* **2** 1009–1018.

[3] EHRHARD, A. (1983). Symétrisation dans l'espace de Gauss. *Math Scand.* **53** 281–301.

[4] GROMOV, M. and MILMAN, V. D. (1983). A topological application of the isoperimetric inequality. *Amer. J. Math.* **105** 843–854.

[5] HOEFFDING, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13–30.

[6] HOPFIELD, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proc. Nat. Acad. Sci. USA* **79** 2554–2558.

[7] LEDOUX, M. (1985). Gaussian randomization and the law of the iterated logarithm in Banach spaces. Unpublished manuscript.

[8] LEDOUX, M. (1993). Les inégalités isoperimétriques en analyse et probabilités. Seminaire Bourbaki, juin 1993. *Astérisque* **216** 343–375.

[9] LEDOUX, M. and TALAGRAND, M. (1991). *Probability in Banach Spaces*. Springer, Berlin.

[10] MARGULIS, G. A. (1977). Probabilistic characteristics of graphs with large connectivity. *Problems Inform. Transmission* **10** 174–179.

[11] MARTON, K. (1986). A simple proof of the blowing-up lemma. *IEEE Trans. Inform. Theory* **IT-32** 445–446.

[12] MARTON, K. (1994). A concentration of measure inequality for contracting Markov chains. Unpublished manuscript.

[13] MAUREY, B. (1979). Construction de suites symétriques. *C. R. Acad. Sci. Paris* **288** 679–681.

[14] MCDIARMID, C. (1989). On the method of bounded differences. In *Surveys in Combinatorics, 1989* (J. Simons, ed.) 148–188. Cambridge Univ. Press.

[15] MILMAN, V. D. (1971). A new proof of the theorem of A. Dvoretzky on sections of convex bodies. *Funct. Anal. Appl.* **5** 28–37.

[16] MILMAN, V. D. (1988). The heritage of P. Levy in geometrical functional analysis. *Astérisque* **157/158** 273–301.

[17] MILMAN, V. and SCHECHTMAN, G. (1986). *Asymptotic Theory of Finite Dimensional Normed Spaces. Lecture Notes in Math.* **1200**. Springer, Berlin.

[18] ODLYZKO, A. (1993). Private communication.

[19] PASTUR, L. and SHCHERBINA, M. (1991). Absence of self-averaging of the order parameter in the Sherrington–Kirkpatrick model. *J. Statist. Phys.* **62** 1–19.

[20] RHEE, W. and TALAGRAND, M. (1987). Martingales inequality and NP-complete problems. *Math. Oper. Res.* **12** 177–181.

[21] SCHECHTMAN, G. (1981). Levy type inequality for a class of metric spaces. In *Martingale Theory in Harmonic Analysis and Banach Spaces Lecture Notes in Math.* **939** 211–215. Springer, Berlin.

[22] SHAMIR, E. and SPENCER, J. (1987). Sharp concentration of the chromatic number of random graphs $G_{n,p}$. *Combinatorica* **7** 121–129.
[23] TALAGRAND, M. (1987). Regularity of Gaussian processes. *Acta Math.* **159** 99–149.
[24] TALAGRAND, M. (1989). A conjecture on convolution operators, and operators from L^1 to a Banach lattice. *Israel J. Math.* **68** 82–88.
[25] TALAGRAND, M. (1993). Isoperimetry, logarithmic Sobolev inequalities on the discrete cube and Margulis' graph connectivity theorem. *Geom. Funct. Anal.* **3** 295–314.
[26] TALAGRAND, M. (1994). Supremum of some canonical processes. *Amer. J. Math.* **116** 283–325.
[27] TALAGRAND, M. (1994). Sharper bounds for Gaussian and empirical processes. *Ann. Probab.* **22** 28–76.
[28] TALAGRAND, M. (1994). Isoperimetry in product spaces: higher level, large sets. Unpublished manuscript.
[29] TALAGRAND, M. (1995). Concentration of measure and isoperimetric inequalities in product spaces. *Publications Math. IHES* **81** 73–205.
[30] TALAGRAND, M. (1995). Transportation cost for Gaussian and other product measures. *Geom. Funct. Anal.* To appear.
[31] TALAGRAND, M. (1995). New concentration inequalities in product spaces. Unpublished manuscript.
[32] THURSTON, W. (1994). On proof and progress in mathematics. *Bull. Amer. Math. Soc.* **30** 161–177.
[33] YURINSKII, V. V. (1974). Exponential bounds for large deviations. *Theory Probab. Appl.* **19** 154–155.

EQUIPE D'ANALYSE—TOUR 56
E.R.A. AU C.N.R.S. NO. 754
UNIVERSITÉ PARIS VI
4 PLACE JUSSIEU
75230 PARIS CEDEX 05
FRANCE

DEPARTMENT OF MATHEMATICS
THE OHIO STATE UNIVERSITY
231 W. 18TH AVENUE
COLUMBUS, OHIO 43210-1174
E-mail: talagran@math.ohio-state.edu