Concentration

Colin McDiarmid
University of Oxford

30 May 1998

Abstract

Upper bounds on probabilities of large deviations for sums of bounded independent random variables may be extended to handle functions which depend in a limited way on a number of independent random variables. This 'method of bounded differences' has over the last dozen or so years had a great impact in probabilistic methods in discrete mathematics and in the mathematics of operational research and theoretical computer science. Recently Talagrand introduced an exciting new method for bounding probabilities of large deviations, which often proves superior to the bounded differences approach. In this paper we introduce and survey these two approaches and some of their applications.

Contents

1 Introduction

What do we mean by 'concentration' here and why should we be concerned with it?

Suppose that a random variable X has expected value $\mathbf{E}(X) = \mu$ and variance $\mathbf{E}((X-\mu)^2) = \sigma^2$. Then Chebychev's inequality states that

$$\mathbf{P}(|X - \mu| \geq t) \leq \sigma^2/t^2$$

for any $t > 0$. Thus for $t >> \sigma$ the probability of deviating by more than t from μ is small. However, we shall often want or need the probability of large deviations to be *very* small, that is, we want to know that X is strongly *concentrated* around μ. The archetypical concentration result is Chernoff's bound on the tails of the binomial distribution [14], in other words on the tails of the sums of independent identically distributed binary (that is, $\{0,1\}$-valued) random variables.

Theorem *Let $X_1, X_2, \ldots, X_n$ be independent binary random variables, with $\mathbf{P}(X_k = 1) = p$ and $\mathbf{P}(X_k = 0) = 1 - p$ for each k, and let $S_n = \sum X_k$. Then for any $t \geq 0$,*

$$\mathbf{P}(|S_n - np| \geq nt) \leq 2e^{-2nt^2}.$$

Typically we shall be interested in a random variable like S_n, and not in the corresponding 'bounded differences' X_k that make it up. The variance of S_n here is $np(1-p) = n/4$ when $p = 1/2$, and then Chebyshev's inequality yields only that $\mathbf{P}(|S_n - np| \geq nt) \leq 1/(4nt^2)$, which will often not be a small enough bound for us. In some cases we shall want good bounds for their own interest, and sometimes as tools within some larger endeavour.

As an example of the former case, consider quicksort. Quicksort is one of the most important sorting algorithms, and its value rests entirely on its good typical behaviour. It is well known that it has good average time complexity. Further, the variance of the time taken is not too large, and so large deviations from the average are not very likely – see for example [36, 59]. However, one would hope that large positive deviations are *very* unlikely, and the bounds that can be obtained from the variance and Chebychev's inequality are weak. It turns out [49] that the method of bounded differences shows that indeed large deviations are exceedingly unlikely (and the method yields essentially best possible bounds). We shall meet several further examples below, including the study of isoperimetric inequalities.

There are also many cases when we need to know concentration results as a step towards something else. One example concerns the behaviour of the chromatic number of a random graph – see Section 3.1 below. Concentration inequalities have become essential tools in the probabilistic analysis of algorithms [16, 25, 63] and the study of randomised algorithms [51], and in probabilistic methods in discrete mathematics (in particular when we wish to use the Lovász Local Lemma) [3]. Some have reached standard undergraduate text books in probability – see for example [28] section 12.2, or [57] section 6.3.

We shall introduce the two main approaches for proving concentration results, namely the bounded differences or martingale method and the recent method of Talagrand, and give several applications of each. We shall also mention briefly how some such results can be proved using ideas from information theory.

The natural starting point is to consider sums of independent random variables, starting with the classical Chernoff bound, introduced above. We do this in Section 2, where we give full proofs in a form which is intended to be widely accessible, and to generalise for the next section.

Section 3 is devoted to the martingale method. We shall not use *any* results about martingales beyond understanding the definition, and indeed the first two subsections do not even mention the word martingale. We first present the 'independent bounded differences inequality'. This is a special case of various more powerful inequalities which we develop later, but it is easy to grasp and has proved to be very useful. We give applications to bin packing, colouring random graphs, and isoperimetric inequalities involving Hamming distances. After that we present closely related extensions of the independent bounded differences inequality, namely Theorems 3.7, 3.8 and 3.9, and illustrate these extensions by describing an early application concerning permutations and a recent application to finding matchings in hypergraphs. These extensions include some results that have been presented very recently, though they can be traced back to earlier work.

In these first two subsections of section 3 which we have just discussed, the applications are proved but not the concentration inequalities, as it is most natural to prove the concentration results in the framework of martingales. The third subsection introduces martingales onto the scene. Following that, the next subsection starts by paralleling the earlier treatment of sums of independent random variables but now considering martingale difference sequences: we find that we can mainly re-use the earlier proofs. Then we give a pair of more general results, Theorems 3.14 and 3.15, which include (nearly) all the previous results, and prove them in the following subsection. Thus Theorems 3.14 and 3.15 could be regarded as the most important of all the results discussed so far, but often a more focussed special case, such as Theorem 3.1 or 3.9, is sufficient for an application, and is then the best tool to use. We end the section on the martingale method with a brief discussion on 'centering' sequences.

The final part, Section 4, introduces Talagrand's inequality (or rather, what seems to be the most useful of his many inequalities!). We give applications to increasing subsequences and common subsequences, to travelling salesman tours and Steiner trees, and to minimum spanning trees. While presenting these applications we deduce from Talagrand's inequality two useful 'packaged' results, Theorems 4.3 and 4.5, which in fact handle all the applications in this paper. These 'packaged' results, which are tailored to our applications, are in fact rather easy deductions from Talagrand's inequality, which itself is proved afterwards. Finally, we discuss briefly how results from information theory may be used to derive concentration results.

We shall stick throughout to bounded discrete 'time', typically $1, \ldots, n$. Thus there are two major related topics that we shall not discuss: for analogous martingale results in continuous time see for example [39], and for an introduction to the asymptotic theory of large deviations see for example [20, 19, 28]. Both these topics are harder work than the discrete case we consider, and seem to be of much less use in discrete mathematics and theoretical computer science.

2 Inequalities for sums of bounded independent random variables

We restate from above the 1952 Chernoff [14] bound on the tails of the binomial distribution.

Theorem 2.1 *Let $0 < p < 1$, let $X_1, X_2, \ldots, X_n$ be independent binary random variables, with $\mathbf{P}(X_k = 1) = p$ and $\mathbf{P}(X_k = 0) = 1 - p$ for each k, and let $S_n = \sum X_k$. Then for any $t \geq 0$,*

$$\mathbf{P}(|S_n - np| \geq nt) \leq 2e^{-2nt^2}.$$

The sum above is over k running from 1 to n. Throughout the paper, when we write an unadorned sum $\sum$ or product $\prod$ the index k runs from 1 to n. The above result will be proved below by bounding the moment generating function $M(h) = \mathbf{E}(e^{hS_n})$ and using Markov's inequality, following the method introduced by Bernstein. Indeed, all the results of this section and the next section use this method. (See [58] for a variant of this method which yields similar results, but assuming only limited independence, and see also [64].)

Recall that Markov's inequality states that for a non-negative random variable X, $\mathbf{P}(X \geq t) \leq \mathbf{E}(X)/t$ for each $t > 0$. To prove this, we use the indicator function $\mathbf{1}_A$ for an event A, and note that, since $X \geq t\mathbf{1}_{(X \geq t)}$, we have

$$\mathbf{E}(X) \geq t\, \mathbf{E}(\mathbf{1}_{(x \geq t)}) = t\, \mathbf{P}(X \geq t).$$

Proof of Theorem 2.1
Let $m = n(p + t)$. Let $h > 0$. Then

$$\mathbf{P}(S_n \geq m) = \mathbf{P}(e^{hS_n} \geq e^{hm}) \leq e^{-hm}\mathbf{E}(e^{hS_n}), \tag{1}$$

by Markov's (or Bernstein's) inequality. By the independence of the random variables X_k,

$$\mathbf{E}(e^{hS_n}) = \mathbf{E}\left(\prod e^{hX_k}\right) = \prod \mathbf{E}(e^{hX_k}) = (1 - p + pe^h)^n.$$

Hence, for any $h > 0$,

$$\mathbf{P}(S_n \geq m) \leq e^{-hm}(1 - p + pe^h)^n.$$

If $0 < t < 1 - p$ then we may set $e^h = \frac{(p+t)(1-p)}{p(1-p-t)}$ to minimise the above bound, and we obtain

$$\mathbf{P}(S_n - np \geq nt) \leq e^{-2nt^2}. \tag{2}$$

This implies by a continuity argument that the inequality holds also for $t = 1 - p$. But the inequality is trivial for $t = 0$ or $t > 1 - p$, and thus it holds for all $t \geq 0$.

Now let $Y_k = 1 - X_k$ for each k. Then by the above result (2),

$$\mathbf{P}(S_n - np \leq -nt) = \mathbf{P}(\sum Y_k - n(1 - p) \geq nt) \leq e^{-2nt^2}$$

for any $t \geq 0$. □

Hoeffding [29] presents extensions of the above theorem which can be based on the following lemma.

Lemma 2.2 *Let the random variables $X_1, X_2, \ldots, X_n$ be independent, with $0 \leq X_k \leq 1$ for each k. Let $S_n = \sum X_k$, let $\mu = \mathbf{E}S_n$, let $p = \mu/n$ and let $q = 1-p$. Then for any $0 \leq t < q$,*

$$\mathbf{P}(S_n - \mu \geq nt) \leq \left(\left(\frac{p}{p+t} \right)^{p+t} \left(\frac{q}{q-t} \right)^{q-t} \right)^n .$$

Proof We follow the lines of the proof of Theorem 2.1. Let $p_k = \mathbf{E}(X_k)$ for each k. Let $m = \mu + nt$, and let $h > 0$. Note that, by the convexity of the function e^{hx} for $0 \leq x \leq 1$, we have $e^{hx} \leq 1 - x + xe^h$, and so $\mathbf{E}(e^{hX_k}) \leq 1 - p_k + p_k e^h$. Thus, since S_n is the sum of the independent random variables S_{n-1} and X_n,

$$\begin{aligned} \mathbf{E}(e^{hS_n}) &= \mathbf{E}(e^{hS_{n-1}})\mathbf{E}(e^{hY_n}) \\ &\leq \mathbf{E}(e^{hS_{n-1}})(1 - p_n + p_n e^h) \\ &\leq \prod (1 - p_k + p_k e^h), \end{aligned}$$

on iterating. Hence,

$$\mathbf{E}(e^{hS_n}) \leq (1 - p + pe^h)^n,$$

by the arithmetic mean – geometric mean inequality. But by Markov's inequality,

$$\mathbf{P}(S_n \geq m) \leq e^{-hm}\mathbf{E}(e^{hS_n}) \leq e^{-hm}(1 - p + pe^h)^n .$$

Thus, for any $h \geq 0$,

$$\mathbf{P}(S_n - \mu \geq nt) \leq \left(e^{-(p+t)h}(1 - p + pe^h) \right)^n . \tag{3}$$

The desired inequality now follows on setting $e^h = \frac{(p+t)(1-p)}{p(1-p-t)}$, as in the proof of Theorem 2.1. □

Our interest is in large deviations and the above bound is good in this case, though inequalities closer to the normal approximation of DeMoivre-Laplace are naturally better for small deviations – see for example [9]. From the above result we may deduce weaker but more useful bounds, which generalise the bounds in Theorem 2.1 or improve on them when p is small.

Theorem 2.3 *Let the random variables $X_1, X_2, \ldots, X_n$ be independent, with $0 \leq X_k \leq 1$ for each k. Let $S_n = \sum X_k$, let $\mu = \mathbf{E}(S_n)$, let $p = \mu/n$ and let $q = 1-p$.*
(a) For any $t \geq 0$,

$$\mathbf{P}(|S_n - \mu| \geq nt) \leq 2e^{-2nt^2}.$$

(b) For any $\epsilon > 0$,

$$\mathbf{P}(S_n \geq (1+\epsilon)\mu) \leq e^{-((1+\epsilon)\ln(1+\epsilon)-\epsilon)\mu} \leq e^{-\frac{\epsilon^2\mu}{2(1+\epsilon/3)}}.$$

(c) For any $\epsilon > 0$,

$$\mathbf{P}(S_n \leq (1-\epsilon)\mu) \leq e^{-\frac{1}{2}\epsilon^2\mu}.$$

Part (a) is due to Hoeffding [29], who also discusses relationships between that result and other similar inequalities. Results similar to parts (b) and (c) appear in [4] (in the binomial case). For similar results in the binomial case based on Stirling's approximation to $n!$ see [9] chapter 1. In order to prove Theorem 2.3 we need one technical lemma.

Lemma 2.4 *For all* $x \geq 0$,

$$(1+x)\ln(1+x) - x \geq 3x^2/(6+2x).$$

Proof Let

$$f_1(x) = (6+8x+2x^2)\ln(1+x) - 6x - 5x^2.$$

We want to show that $f_1(x) \geq 0$ for all $x \geq 0$. Now $f_1(0) = 0$, and $f_1'(x) = 4f_2(x)$ where $f_2(x) = (2+x)\ln(1+x) - 2x$. It suffices to show that $f_2(x) \geq 0$ for all $x \geq 0$. Now $f_2(0) = 0$, and $f_2'(x) = (1+x)^{-1} + \ln(1+x) - 1$. Now $f_2''(0) = 0$, so it suffices to show that $f_2''(x) \geq 0$ for all $x \geq 0$. But $f_2''(x) = x(1+x)^{-2} \geq 0$, and so we are done. □

Proof of Theorem 2.3
(a) Consider p fixed, let $q = 1-p$, and for $0 \leq t < q$ let

$$f(t) = \ln\left(\left(\frac{p}{p+t}\right)^{p+t}\left(\frac{q}{q-t}\right)^{q-t}\right).$$

Then

$$f'(t) = \ln\left(\frac{p(q-t)}{(p+t)q}\right),$$

and

$$f''(t) = -\left((p+t)(1-(p+t))\right)^{-1} \leq -4.$$

Now $f(0) = f'(0) = 0$ and so it follows by Taylor's theorem that for $0 \leq t < q$, $f(t) = (t^2/2)f''(s)$ for some s with $0 \leq s \leq t$. Hence $f(t) \leq -2t^2$. Hence by Lemma 2.2,

$$\mathbf{P}(S_n - \mu \geq nt) \leq e^{-2nt^2}. \tag{4}$$

By applying this result to $n - S_n$ we obtain

$$\mathbf{P}(S_n - \mu \leq -nt) \leq e^{-2nt^2}. \tag{5}$$

(b) To prove part (b) it is simpler to use the inequality (3) in the proof of Lemma 2.2 rather than the lemma itself. If we set $t = \epsilon p$ and $e^h = (1+\epsilon)$ there, and use the inequality $1 + x \leq e^x$, we obtain

$$\mathbf{P}(S_n \geq (1+\epsilon)\mu) \leq \left((1+\epsilon)^{-(1+\epsilon)p}(1+\epsilon p)\right)^n \leq \left((1+\epsilon)^{-(1+\epsilon)}e^{\epsilon}\right)^{np},$$

and this gives the first inequality in (b) (see also Appendix A of [3]). The second inequality in (b) follows from Lemma 2.4.

(c) Let the function f be as in (a) above, and let $h(x) = f(-xp)$ for $0 \le x < 1$. Then $h'(x) = -pf'(-xp)$ and

$$h''(x) = p^2 f''(-xp) = -\frac{p}{(1-x)(q+xp)} \le -p.$$

Thus we may use Taylor's theorem as above to see that $h(x) \le -px^2/2$, and then Lemma 2.2 completes the proof. □

The first inequality in part (b) yields useful results for very large deviations. In particular,

$$\mathbf{P}(S_n \ge 2\mu) \le e^{-\mu}. \tag{6}$$

Also,

$$\mathbf{P}(S_n \ge \delta\mu) \le e^{-\delta(\ln\delta - \delta + 1)\mu} \le e^{-\delta\ln(\delta/e)\mu},$$

and so, if $\delta \ge 2e$, then

$$\mathbf{P}(S_n \ge \delta\mu) \le 2^{-\delta\mu}. \tag{7}$$

The second inequality in part (b) yields immediately that

$$\mathbf{P}(S_n \ge (1+\epsilon)\mu) \le e^{-\frac{1}{3}\epsilon^2\mu} \tag{8}$$

for $0 \le \epsilon \le 1$, which is often a sufficiently precise inequality in applications, see for example [4]. Hoeffding also gives the following extension of part (a) above to the case when the ranges of the summands may differ.

Theorem 2.5 *Let the random variables $X_1, \ldots, X_n$ be independent, with $a_k \le X_k \le b_k$ for each k, for suitable constants a_k, b_k. Let $S_n = \sum X_k$ and let $\mu = \mathbf{E}(S_n)$. Then for any $t \ge 0$,*

$$P(|S_n - \mu| \ge t) \le 2e^{-2t^2/\sum(b_k - a_k)^2}.$$

To prove this result we need one lemma, from [29].

Lemma 2.6 *Let the random variable X satisfy $\mathbf{E}(X) = 0$ and $a \le X \le b$, where a and b are constants. Then for any $h > 0$*

$$\mathbf{E}(e^{hX}) \le e^{\frac{1}{8}h^2(b-a)^2}.$$

Proof Since e^{hx} gives a convex function of x, for $a \le x \le b$

$$e^{hx} \le \frac{x-a}{b-a}e^{hb} + \frac{b-x}{b-a}e^{ha},$$

and so

$$\begin{aligned}\mathbf{E}(e^{hX}) &\le \frac{b}{b-a}e^{ha} - \frac{a}{b-a}e^{hb} \\ &= (1-p)e^{-py} + pe^{(1-p)y} \\ &= e^{-py}(1-p+pe^y) \;=\; e^{f(y)}\end{aligned}$$

where $p = -a/(b-a)$, $y = (b-a)h$ and $f(x) = -px + \ln(1-p+pe^x)$. But

$$f'(x) = -p + \frac{pe^x}{(1-p)+pe^x} = -p + \frac{p}{p+(1-p)e^{-x}},$$

and so

$$f''(x) = \frac{p(1-p)e^{-x}}{(p+(1-p)e^{-x})^2} \leq \frac{1}{4}$$

(since the geometric mean is at most the arithmetic mean). Also $f(0) = f'(0) = 0$, and hence by Taylor's theorem

$$f(y) \leq \frac{1}{8}y^2 = \frac{1}{8}(b-a)^2h^2,$$

which gives the desired inequality. □

Proof of Theorem 2.5 By Lemma 2.6, for $h > 0$

$$\begin{aligned} \mathbf{E}(e^{h(S_n-\mu)}) &= \mathbf{E}\left(\prod e^{h(X_k-E(X_k))}\right) \\ &= \prod \mathbf{E}\left(e^{h(X_k-E(X_k))}\right) \\ &\leq e^{\frac{1}{8}h^2\sum(b_k-a_k)^2}. \end{aligned}$$

Hence by Markov's inequality,

$$\begin{aligned} \mathbf{P}(S_n-\mu \geq t) &\leq e^{-ht}\mathbf{E}(e^{h(S_n-\mu)}) \\ &\leq e^{-ht+\frac{1}{8}h^2\sum(b_k-a_k)^2}. \end{aligned}$$

Now set $h = 4t/\sum(b_k-a_k)^2$ to obtain

$$\mathbf{P}(S_n-\mu \geq t) \leq e^{-2t^2/\sum(b_k-a_k)^2}.$$

Finally, replace $\mathbf{X}$ by $-\mathbf{X}$ to obtain

$$\mathbf{P}(S_n-\mu \leq -t) \leq e^{-2t^2/\sum(b_k-a_k)^2},$$

and thus complete the proof. □

Much work has also been done on tail bounds for the sum S_n when, as well as knowing bounds on the ranges of the summands X_k, we know bounds on their variances $var(X_k)$ – see for example [7, 29]. The following result builds on work of Bernstein (see [7] and [29] equation (2.13)). We shall develop more general results along these lines later. The reader may notice the similarity to part (b) of Theorem 2.3.

Theorem 2.7 *Let the random variables $X_1, \ldots, X_n$ be independent, with $X_k - \mathbf{E}(X_k) \leq b$ for each k. Let $S_n = \sum X_k$, and let S_n have expected value μ and variance V (the sum of the variances of the X_k). Then for any $t \geq 0$,*

$$\begin{aligned} & \mathbf{P}(S_n-\mu \geq t) \\ \leq\ & e^{-(V/b^2)((1+\epsilon)\ln(1+\epsilon)-\epsilon)} \quad \textit{where } \epsilon = bt/V \qquad (9) \\ \leq\ & e^{-\frac{t^2}{2V(1+(bt/3V))}}. \qquad (10) \end{aligned}$$

In typical applications of the inequality (10), the 'error' term $bt/3V$ will be negligible. Suppose for example that the random variables X_k have the same bounded distribution, with positive variance σ^2, and so $V = n\sigma^2$. Then for $t = o(n)$, the bound in (10) is $e^{-(1+o(1))\frac{t^2}{2V}}$. This is the natural 'target', since by the Central Limit Theorem $S_n - \mu$ is asymptotically normal with mean 0 and variance V.

In the proof of Theorem 2.5 above we used Lemma 2.6 to give a bound on the moment generating function for a bounded random variable with expected value 0. In order to prove Theorem 2.7, we now need a related result, see [65].

Lemma 2.8 *Let*

$$g(x) = \frac{1}{2} + \frac{x}{3!} + \frac{x^2}{4!} + \cdots = (e^x - 1 - x)/x^2$$

if $x \neq 0$. Then the function g is increasing; and, if the random variable X satisfies $\mathbf{E}(X) = 0$ and $X \leq b$, then

$$\mathbf{E}(e^X) \leq e^{g(b)var(X)}.$$

Proof To show that g is increasing, note that for $x \neq 0$,

$$g'(x) = x^{-3}((x-2)e^x + 2 + x),$$

and so it suffices to show that $h(x) = (x-2)e^x + 2 + x$ satisfies $h(x) \geq 0$ for all x. Now $h(0) = 0$ and $h'(x) = (x-2)e^x + 1$. Then $h'(0) = 0$ and $h''(x) = xe^x$, so $h'(x) < 0$ for $x < 0$ and $h'(x) > 0$ for $x > 0$, and thus indeed $h(x) \geq 0$ for all x as required.

For the second part of the lemma, note that

$$e^x = 1 + x + x^2 g(x) \leq 1 + x + x^2 g(b)$$

for $x \leq b$. Hence, if $\mathbf{E}(X) = 0$ and $X \leq b$, then

$$\mathbf{E}(e^X) \leq 1 + g(b)var(X) \leq e^{g(b)var(X)},$$

as required. □

Proof of Theorem 2.7 The proof follows the lines of the proof of Theorem 2.5 above. By Lemma 2.8, for any h

$$\mathbf{E}(e^{h(S_n - \mu)}) = \prod \mathbf{E}\left(e^{h(X_k - E(X_k))}\right) \leq e^{g(hb)h^2 V}.$$

Hence by Markov's inequality, for any $h \geq 0$

$$\mathbf{P}(S_n - \mu \geq t) \leq e^{-ht}\mathbf{E}(e^{h(S_n - \mu)}) \leq e^{-ht + g(hb)h^2 V}. \tag{11}$$

To minimise this bound we set $h = \frac{1}{b}\ln(1 + \frac{bt}{V})$, and then we obtain (9), and finally Lemma 2.4 yields (10).

Inequalities for maxima

All the theorems above on sums of independent random variables can be strengthened to refer to maxima. Since we have no natural applications in the present context

for these strengthenings, we restrict ourselves to a comment here and then say a little more at the end of subsection 3.5.

Each of the theorems is based on the elementary Bernstein inequality

$$\mathbf{P}(Z \geq t) \leq e^{-ht}\mathbf{E}(e^{hZ}) \quad \text{for each } h \geq 0.$$

Consider for example the Chernoff Theorem, Theorem 2.1, where $S_n = \sum X_k$ and $\mu_n = \mathbf{E}(S_n)$: to prove this result we may apply the above inequality with $Z = S_n - \mu_n$ where $\mu_n = \mathbf{E}(S_n) = np$, that is we use the inequality

$$\mathbf{P}(S_n - \mu_n \geq t) \leq e^{-ht}\mathbf{E}(e^{h(S_n-\mu_n)}) \quad \text{for each } h \geq 0.$$

However, a stronger inequality holds. Let $S_k = \sum_{i=1}^{k} X_i$ and $\mu_k = \mathbf{E}(S_k)$: then

$$\mathbf{P}(\max(S_k - \mu_k) \geq t) \leq e^{-ht}\mathbf{E}(e^{h(S_n-\mu_n)}) \quad \text{for each } h \geq 0.$$

Here the maximum is over $k = 1, \ldots, n$. Thus the same proof as before shows that, for any $t \geq 0$,

$$\mathbf{P}(\max(|S_k - kp|) \geq nt) \leq 2e^{-2nt^2}.$$

However, in typical applications of concentration inequalities in discrete mathematics or theoretical computer science, we do not start with the X_k and then wish to investigate the sums $S_1, S_2, \ldots$: we start with a random quantity Z of interest and then define further random variables X_k such that $Z = \sum X_k$ in order to investigate Z, so that we are not interested for example in S_{n-1}.

Not only may the theorems above on sums of independent random variables be strengthened to refer to maxima, but also this holds for many of the more general results in the next section, as they are also based on the Bernstein inequality – see the comment at the end of subsection 3.5.

3 Martingale methods

We shall make some introductory comments about martingales in subsection 3.3 below. No knowledge of martingales will be required in the first two subsections below! Indeed, they will not be mentioned, though we shall see later that the inequalities presented in these subsections are most naturally understood in the context of martingales, and indeed they could be called closet martingale results.

3.1 The independent bounded differences inequality

In this subsection, we introduce and give several applications for the 'independent bounded differences inequality', Theorem 3.1 below, from [45]. This result is a special case of Theorem 3.7 below (and thus also of Theorem 3.14), but it has proved very useful and is immediately accessible and so we discuss it first. (We should insist below that the function f be appropriately integrable: we ignore such details here and throughout the paper.)

Theorem 3.1 *Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be a family of independent random variables with X_k taking values in a set A_k for each k. Suppose that the real-valued function f defined on $\prod A_k$ satisfies*

$$|f(\mathbf{x}) - f(\mathbf{x}')| \leq c_k \tag{12}$$

whenever the vectors $\mathbf{x}$ *and* $\mathbf{x}'$ *differ only in the* k*th co-ordinate. Let* μ *be the expected value of the random variable* $f(\mathbf{X})$*. Then for any* $t \geq 0$*,*

$$\mathbf{P}(f(\mathbf{X}) - \mu \geq t) \leq e^{-2t^2/\sum c_k^2}. \tag{13}$$

The inequality (13) is 'one-sided'. If we apply it to $-f$ we obtain

$$\mathbf{P}(f(\mathbf{X}) - \mu \leq -t) \leq e^{-2t^2/\sum c_k^2}, \tag{14}$$

and so we have deduced the 'two-sided' inequality

$$\mathbf{P}(|f(\mathbf{X}) - \mu| \geq t) \leq 2e^{-2t^2/\sum c_k^2}. \tag{15}$$

A similar comment holds for most of the one-sided results we present.

If we let each set $A_k = \{0, 1\}$ and let $f(\mathbf{x}) = \sum x_k$ we obtain Theorem 2.1 above; and if each set A_k is a bounded set of numbers we obtain Theorem 2.5. We consider a variety of applications below. We do not prove Theorem 3.1 at this point, as the proof is most naturally set in the framework of martingales and we shall shortly develop more general results – see in particular Theorem 3.7 below.

3.1.1 Bin packing

Our first application is quick and easy. Given an n-vector $\mathbf{x} = (x_1, \ldots, x_n)$ where $0 \leq x_k \leq 1$ for each k, let $B(\mathbf{x})$ be the least number of unit size bins needed to store items with these sizes. We assume that the items have independent random sizes. Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a family of independent random variables each taking values in $[0, 1]$. Then the bounded differences condition (12) holds with each $c_k = 1$, and so (as noted in [45, 54]) it follows from Theorem 3.1 that

$$\mathbf{P}(|B(\mathbf{X}) - \mu| \geq t) \leq 2e^{-2t^2/n}, \tag{16}$$

where μ is the expected value of $B(\mathbf{X})$. Thus if $\omega(n) \to \infty$ as $n \to \infty$, then the probability that $B(\mathbf{X})$ deviates from its mean by more than $\omega(n)\sqrt{n}$ tends to 0 as $n \to \infty$. We may say that $B(\mathbf{X})$ is concentrated within width $O(\sqrt{n})$. For a similar result on random knapsacks see [45]. (For finer concentration results on bin packing that use also the variance of the random variables X_k see [68, 42].)

3.1.2 Random graphs

In Theorem 3.1 we may take A_k as a set of edges in a graph, as in the results below – see for example [10, 12]. Recall that the random graph $G_{n,p}$ has vertices $1, \ldots, n$ and the possible edges appear independently with probability p.

Lemma 3.2 *Let* $(A_1, \ldots, A_m)$ *be a partition of the edge set of the complete graph* K_n *into* m *blocks; and suppose that the graph function* f *satisfies* $|f(G) - f(G')| \leq 1$ *whenever the symmetric difference* $E(G) \triangle E(G')$ *of the edge-sets is contained in a single block* A_k*. Then the random variable* $Y = f(G_{n,p})$ *satisfies*

$$\mathbf{P}(Y - \mathbf{E}(Y) \geq t) \leq e^{-2t^2/m} \quad \textit{for } t \geq 0.$$

This result follows directly from Theorem 3.1 with each $c_k = 1$. The next two results are immediate consequences of Lemma 3.2: for the former let A_k be the set of edges $\{j, k\}$ where $j < k$, and for the latter let the blocks A_k be singletons. We may think of 'exposing' the random graph step-by-step: at step k we expose which edges in the set A_k are present.

Lemma 3.3 *Suppose that the graph function f satisfies $|f(G) - f(G')| \leq 1$ whenever G' can be obtained from G by changing edges incident with a single vertex. Then the corresponding random variable $Y = f(G_{n,p})$ satisfies*

$$\mathbf{P}(Y - \mathbf{E}(Y) \geq t) \leq e^{-2t^2/n} \text{ for } t \geq 0.$$

When we consider the chromatic number $\chi(G)$ and let $Y = \chi(G_{n,p})$ (and use the two-sided version of the last lemma), we find that

$$\mathbf{P}(|Y - \mathbf{E}(Y)| \geq t) \leq 2e^{-2t^2/n}, \tag{17}$$

which is (a slight sharpening of) the early result of Shamir and Spencer [60] which was important in introducing martingale methods into this area.

Lemma 3.4 *Suppose that the graph function f satisfies $|f(G) - f(G')| \leq 1$ whenever G and G' differ in only one edge. Then the corresponding random variable $Y = f(G_{n,p})$ satisfies*

$$\mathbf{P}(Y - \mathbf{E}(Y) \geq t) \leq e^{-4t^2/n^2} \text{ for } t \geq 0.$$

Perhaps the most exciting application of the bounded differences method uses this lemma. It is the proof by Bollobás [11] of what was a long-standing conjecture about the chromatic number $\chi(G_{n,p})$ of random graphs. Consider a constant edge probability p with $0 < p < 1$ and let $q = 1 - p$. Then for any $\epsilon > 0$,

$$\mathbf{P}\left((1-\epsilon)\frac{n}{2\log_q n} \leq \chi(G_{n,p}) \leq (1+\epsilon)\frac{n}{2\log_q n}\right) \to 1 \text{ as } n \to \infty.$$

(For a more precise result see [46].)

The lower bound part of the proof is easy: the interest is in establishing the upper bound for $\chi(G_{n,p})$. The key step in the proof is to show that the probability $\tilde{p}(n)$ that $G_{n,p}$ fails to contain a stable (independent) set with $s(n) = \lceil (2-\epsilon)\log_q n \rceil$ vertices is very small, say

$$\tilde{p}(n) = O(e^{-n^{\frac{4}{3}}}). \tag{18}$$

To see how this will yield the upper bound on $\chi(G_{n,p})$, let $\tilde{n} = \lceil n/\log^2 n \rceil$ and call a set W of at least $\tilde{n}$ vertices in $G_{n,p}$ *bad* if it contains no stable set of size at least $s(\tilde{n})$. The probability that there is a bad set is at most $2^n \tilde{p}(\tilde{n}) = o(1)$. But if there is no bad set W, then we can repeatedly colour a stable set of size at least $s(\tilde{n})$ and delete it, until there remain fewer than $\tilde{n}$ vertices, which may each get a new colour. The total number of colours used by this procedure is then at most

$$n/s(\tilde{n}) + \tilde{n} = (\frac{1}{2-\epsilon} + o(1))n/\log_q n.$$

Thus we wish to see that (18) is true. The clever idea is to consider not just big stable sets but packings of such sets. Given a graph G on n vertices, define $f(G)$ to

be the maximum number of stable sets of size $s(n)$ which pairwise contain at most one common vertex. If graphs G and G' differ in only one edge then $f(G)$ and $f(G')$ differ by at most 1. Let $X_n = f(G_{n,p})$. It is not hard to check that $\mu = \mathbf{E}(X_n)$ is large, say at least $n^{\frac{5}{3}}$ for n sufficiently large. Hence by (the other one-sided version of) Lemma 3.4, the probability $\tilde{p}(n)$ that $G_{n,p}$ has no stable set of size $s(n)$ equals

$$\mathbf{P}(X_n = 0) = \mathbf{P}(X_n - \mu_n \leq -\mu_n) \leq e^{-4\mu_n^2/n^2} \leq e^{-4n^{\frac{4}{3}}},$$

for n sufficiently large.

3.1.3 Hamming distances and isoperimetric inequalities

Next let us consider an application of the independent bounded differences inequality Theorem 3.1 involving Hamming distances in product spaces, and corresponding isoperimetric inequalities. This application will link in with our discussion later on Talagrand's inequality and on the use of ideas from information theory to prove concentration results.

Let $\Omega_1, \ldots, \Omega_n$ be probability spaces, and let Ω denote the product space $\prod \Omega_k$. Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a family of independent random variables with X_k taking values in Ω_k. Recall that for points $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ in Ω, the *Hamming distance* $d_H(\mathbf{x}, \mathbf{y})$ is the number of indices i such that $x_i \neq y_i$. We can use the independent bounded differences inequality to show that for any subset A of Ω such that $\mathbf{P}(\mathbf{X} \in A)$ is not too small, the probability that a random point $\mathbf{X}$ is 'close' to A is near 1. Recall that the Hamming distance from a point $\mathbf{x}$ to a set A is defined by setting $d_H(\mathbf{x}, A)$ to be $\inf\{d_H(\mathbf{x}, \mathbf{y}) : \mathbf{y} \in A\}$.

Theorem 3.5 *Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a family of independent random variables and let A be a subset of the product space. Then for any $t \geq 0$,*

$$\mathbf{P}(\mathbf{X} \in A)\, \mathbf{P}(d_H(\mathbf{X}, A) \geq t) \leq e^{-t^2/2n}. \tag{19}$$

Let us rephrase this result before we prove it. Define the *t-fattening* of a subset A of Ω to be the set of points $\mathbf{x} \in \Omega$ such that $d_H(\mathbf{x}, A) < t$, and let the *measure* $\nu(A)$ be $\mathbf{P}(\mathbf{X} \in A)$. Then (19) says that

$$\nu(A)(1 - \nu(A_t)) \leq e^{-t^2/4}.$$

Thus if $\nu(A) \geq \frac{1}{2}$ then $\nu(A_t) \geq 1 - 2e^{-t^2/4}$. In particular, when each random variable X_k is uniformly distributed on the set $\Omega_k = \{0, 1\}$ we obtain an isoperimetric inequality for the n-cube – see for example [37, 45, 63].

Proof of Theorem 3.5 Let $\rho = \mathbf{P}(\mathbf{X} \in A)$ and let $\mu = \mathbf{E}(d_H(\mathbf{X}, A))$. We may assume that $\rho > 0$. By the independent bounded differences inequality, for $t \geq 0$

$$\mathbf{P}(d_H(\mathbf{X}, A) - \mu \geq t) \leq e^{-2t^2/n}, \tag{20}$$

and

$$\mathbf{P}(d_H(\mathbf{X}, A) - \mu \leq -t) \leq e^{-2t^2/n}. \tag{21}$$

Now $d_H(\mathbf{x}, A) = 0$ if and only if $\mathbf{x} \in A$, so if we take $t = \mu$ in the inequality (21) above, we obtain

$$\rho = \mathbf{P}(\mathbf{X} \in A) = \mathbf{P}(d_H(\mathbf{X}, A) - \mu \leq -\mu) \leq e^{-2\mu^2/n},$$

and so

$$\mu \leq (\tfrac{1}{2} n \ln(1/\rho))^{\frac{1}{2}}, \ = t_0 \ \text{ say.}$$

Now use this bound in the inequality (20) above, to find

$$\mathbf{P}(d_H(\mathbf{X}, A) \geq t + t_0) \leq e^{-2t^2/n}.$$

Thus for $t \geq t_0$ we have

$$\mathbf{P}(d_H(\mathbf{X}, A) \geq t) \leq e^{-2(t-t_0)^2/n}. \tag{22}$$

Now $(t-a)^2 \geq t^2/4$ for $t \geq 2a$, so if we take $t \geq 2t_0$ in the inequality (22) we obtain

$$\mathbf{P}(d_H(\mathbf{X}, A) \geq t) \leq e^{-t^2/2n}.$$

But for $0 \leq t \leq 2t_0$, the right hand side above is at least $e^{-2t_0^2/n} = \rho = \mathbf{P}(A)$. Thus

$$\min(\mathbf{P}(\mathbf{X} \in A),\ \mathbf{P}(d_H(\mathbf{X}, A) \geq t)) \leq e^{-t^2/2n}$$

for any $t \geq 0$. □

We may generalise the above discussion. Let $\alpha = (\alpha_1, \ldots, \alpha_n) \geq \mathbf{0}$ be an n-vector of non-negative real numbers. Recall that the (L_2) norm is given by

$$||\alpha|| = (\sum \alpha_k^2)^{\frac{1}{2}},$$

and we call α a *unit vector* if it has norm $||\alpha|| = 1$. For points $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ in Ω, the *α-Hamming distance* $d_\alpha(\mathbf{x}, \mathbf{y})$ is the sum of the values α_i over those indices i such that $x_i \neq y_i$. Thus when α is the all 1's vector, it has norm $\sqrt{n}$ and α-Hamming distance is just the same as Hamming distance. Also, for a subset A of Ω, we define

$$d_\alpha(\mathbf{x}, A) = \inf\{d_\alpha(\mathbf{x}, \mathbf{y}) : \mathbf{y} \in A\}.$$

Exactly the same proof as for Theorem 3.5 yields the following extension of it.

Theorem 3.6 *Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a family of independent random variables, let α be a non-negative unit n-vector, and let A be a subset of the product space. Then for any $t \geq 0$,*

$$\mathbf{P}(\mathbf{X} \in A)\, \mathbf{P}(d_\alpha(\mathbf{X}, A) \geq t) \leq e^{-t^2/2}.$$

Similar results appear in [50, 68, 69]. The central result of Section 4, namely Talagrand's inequality Theorem 4.1, looks rather similar to Theorem 3.6 but is in fact far more powerful, since it refers not just to one unit vector α but simultaneously to all such vectors.

The above result will give us back a result like Theorem 3.1, centered around a median rather than the mean. Let us see how to do this. Consider a function f defined

on $\prod A_k$ as there, and let $\mathbf{c}$ be the vector $(c_1, \ldots, c_n)$. Then the bounded differences condition (12), that $|f(\mathbf{x}) - f(\mathbf{x}')| \leq c_k$ whenever the vectors $\mathbf{x}$ and $\mathbf{x}'$ differ only in the kth co-ordinate, is equivalent to the condition that $|f(\mathbf{x}) - f(\mathbf{x}')| \leq d_{\mathbf{c}}(\mathbf{x}, \mathbf{x}')$. Now assume that the condition (12) holds. Let

$$A_a = \{\mathbf{y} \in \prod A_k : f(\mathbf{y}) \leq a\}.$$

Consider an $\mathbf{x} \in \prod A_k$. For each $\mathbf{y} \in A_a$,

$$f(\mathbf{x}) \leq f(\mathbf{y}) + d_{\mathbf{c}}(\mathbf{x}, \mathbf{y}) \leq a + d_{\mathbf{c}}(\mathbf{x}, \mathbf{y}),$$

and so, minimising over such $\mathbf{y}$,

$$f(\mathbf{x}) \leq a + d_{\mathbf{c}}(\mathbf{x}, A_a).$$

Let $c = \|\mathbf{c}\|$, and let α be the unit vector $\mathbf{c}/c$ along $\mathbf{c}$. If $f(\mathbf{x}) \geq a + t$ then

$$d_\alpha(\mathbf{x}, A_a) = d_{\mathbf{c}}(\mathbf{x}, A_a)/c \geq (f(\mathbf{x}) - a)/c \geq t/c.$$

Hence by Theorem 3.6, for any $t \geq 0$,

$$\mathbf{P}(f(\mathbf{X}) \leq a)\, \mathbf{P}(f(\mathbf{X}) \geq a + t) \leq \mathbf{P}(\mathbf{X} \in A_a)\, \mathbf{P}(d_\alpha(\mathbf{X}, A_a) \geq t/c) \leq e^{-t^2/2c^2}.$$

Now let m be a median of $f(\mathbf{X})$, that is $\mathbf{P}(f(\mathbf{X}) \leq m) \geq \frac{1}{2}$ and $\mathbf{P}(f(\mathbf{X}) \geq m) \geq \frac{1}{2}$. Taking $a = m$ above gives

$$\mathbf{P}(f(\mathbf{X}) \geq m + t) \leq 2e^{-t^2/2c^2}, \tag{23}$$

and taking $a = m - t$ we have

$$\mathbf{P}(f(\mathbf{X}) \leq m - t) \leq 2e^{-t^2/2c^2}. \tag{24}$$

The above two inequalities are like the conclusion of Theorem 3.1, at least if we are not too bothered about constants. They refer to concentration about the median m rather than the mean $\mu = \mathbf{E}(f(\mathbf{X}))$, but that makes little difference since the concentration inequalities themselves imply that $|\mu - m|$ is small. Indeed, the inequalities (23) and (24) together with Lemma 4.6 in subsection 4.2 below show that

$$|\mu - m| \leq \sqrt{2\pi}c. \tag{25}$$

So it is not important whether we refer to median or mean, and Theorem 3.6 and Theorem 3.1 are quite similar.

3.2 Extensions

In this subsection we refine the independent bounded differences inequality, Theorem 3.1, and the Bernstein inequality, Theorem 2.7, to obtain more widely applicable results, namely Theorems 3.7, 3.8 and 3.9, but at the cost of some added complication. We shall deduce these theorems later as immediate consequences of martingale theorems (though they do not themselves mention martingales!). Theorems such as these have recently proved useful when the random variables X_k correspond to answering

questions such as whether two given vertices are adjacent in a random graph, and the question asked at time k may depend on the answers to previous questions – see for example [32, 2, 26]. We shall give part of an argument from [2] concerning hypergraph matchings at the end of this subsection.

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a family of random variables with X_k taking values in a set A_k, and let f be a real-valued function defined on $\prod A_k$. Typically the random variables X_k will be independent but we shall *not* assume this here. We define quantities which measure the variability of the random variable $f(\mathbf{X})$ when the random variables $X_1, \ldots, X_{k-1}$ are fixed. These quantities correspond to deviation, range and variance. It is convenient to note first an easy bound on variance. If the random variable X satisfies $\mathbf{E}(X) = 0$ and $a \leq X \leq b$, then

$$var(X) = \mathbf{E}(X(X-a)) \leq \mathbf{E}(b(X-a)) = |ab| \leq (b-a)^2/4. \tag{26}$$

Let $x_i \in A_i$ for each $i = 1, \ldots, k-1$, and let B denote the event that $X_i = x_i$ for each $i = 1, \ldots, k-1$. Let the random variable Y be distributed like X_k conditional on the event B (so if $k = 1$ then Y is distributed like X_1 with no conditioning, and if the random variables X_k are independent then for each k the random variable Y is distributed like X_k). For $x \in A_k$ let

$$g(x) = \mathbf{E}(f(\mathbf{X}) \mid B, X_k = x) - \mathbf{E}(f(\mathbf{X}) \mid B).$$

If the random variables X_k are independent then we may write $g(x)$ as

$$\mathbf{E}(f(x_1, \ldots, x_{k-1}, x, X_{k+1}, \ldots, X_n)) - \mathbf{E}(f(x_1, \ldots, x_{k-1}, X_k, X_{k+1}, \ldots, X_n)).$$

The function $g(x)$ measures how much the expected value of $f(\mathbf{X})$ changes if it is revealed that X_k takes the value x. Observe that $\mathbf{E}(g(Y)) = 0$.

Let $dev^+(x_1, \ldots, x_{k-1})$ be $\sup\{g(x) : x \in A_k\}$, the *positive deviation* of $g(Y)$, and similarly let $dev(x_1, \ldots, x_{k-1})$ be $\sup\{|g(x)| : x \in A_k\}$, the *deviation* of $g(Y)$. (If we denote $\mathbf{E}(f(\mathbf{X}))$ by μ, then for each $\mathbf{x} = (x_1, \ldots, x_n) \in \prod A_k$ we have

$$|f(\mathbf{x}) - \mu| \leq \sum dev(x_1, \ldots, x_{k-1}). \tag{27}$$

This inequality may be combined (or 'interpolated') with other inequalities like Theorem 3.1 – see [55, 38].) Let $ran(x_1, \ldots, x_{k-1})$ denote $\sup\{|g(x) - g(y)| : x, y \in A_k\}$, the *range* of $g(Y)$. Also, denote the variance of $g(Y)$ by $var(x_1, \ldots, x_{k-1})$.

For $\mathbf{x} \in \prod A_k$, let the *sum of squared ranges* be

$$R^2(\mathbf{x}) = \sum_{k=1}^{n} (ran(x_1, \ldots, x_{k-1}))^2,$$

and let the *maximum sum of squared ranges* $\hat{r}^2$ be the supremum of the values $R^2(\mathbf{x})$ over all $\mathbf{x} \in \prod A_k$. Similarly let the *sum of variances* be

$$V(\mathbf{x}) = \sum_{k=1}^{n} var(x_1, \ldots, x_{k-1}),$$

and let the *maximum sum of variances* $\hat{v}$ be the supremum of the values $V(\mathbf{x})$ over all $\mathbf{x} \in \prod A_k$. Observe that $V(\mathbf{x}) \leq R^2(\mathbf{x})/4$ for each $\mathbf{x}$ by (26), and so $\hat{v} \leq \hat{r}^2/4$. It is also of interest to note that

$$var(f(\mathbf{X})) = \mathbf{E}(V(\mathbf{X})) \leq \hat{v},$$

as is shown just before Theorem 3.14 below. Finally here, let $\max dev^+$ be the maximum of all the positive deviation values $dev(x_1, \ldots, x_{k-1})$, over all choices of k and the x_i, and similarly let $\max dev$ be the maximum of all the deviation values $dev(x_1, \ldots, x_{k-1})$.

Example Define the function $f : \{0,1\}^3 \to \{0,1\}$ by letting $f(\mathbf{x})$ be 0 on $(0,0,0), (0,1,0), (1,0,1)$ and be 1 otherwise. Let $\mathbf{X} = (X_1, X_2, X_3)$ be a family of independent random variables with $\mathbf{P}(X_k = 0) = \mathbf{P}(X_k = 1) = \frac{1}{2}$ for each k. Thus $\mathbf{E}(f(\mathbf{X})) = 5/8$, and $var(f(\mathbf{X})) = 5/8 - (5/8)^2 = 15/64$.

At the 'root', $g(0) = \mathbf{E}(f(0, X_2, X_3)) - \mathbf{E}(f(\mathbf{X})) = 1/2 - 5/8 = -1/8$, and similarly $g(1) = 3/4 - 5/8 = 1/8$. Thus $ran() = 1/4$, $dev^+() = dev() = 1/8$ and $var() = 1/64$.

What happens if $X_1 = 1$? We have $\mathbf{E}(f(\mathbf{X}) \mid X_1 = 1) = \mathbf{E}(f(1, X_2, X_3)) = 3/4$, and so $g(0) = \mathbf{E}(f(1,0,X_3)) - 3/4 = -1/4$ and $g(1) = \mathbf{E}(f(1,1,X_3)) - 3/4 = 1/4$. Thus $ran(1) = 1/2$, $dev^+(1) = dev(1) = 1/4$, and $var(1) = 1/16$. Similarly, $ran(1,0) = 1$ and $var(1,0) = 1/4$.

Now let $\mathbf{x} = (1,0,1)$ (or $(1,0,0)$). The corresponding sum of squared ranges $R^2(\mathbf{x})$ is $ran()^2 + ran(1)^2 + ran(1,0)^2 = 1/16 + 1/4 + 1 = 21/16$, which in fact equals $\hat{r}^2$. The corresponding sum of variances $V(\mathbf{x})$ is $var() + var(1) + var(1,0) = 15/64 + 1/64 + 1/4 = 1/2$, which in fact equals $\hat{v}$.

We are now ready to state the frist of our more general results, which extends the independent bounded differences inequality, Theorem 3.1.

Theorem 3.7 *Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a family of random variables with X_k taking values in a set A_k, and let f be a bounded real-valued function defined on $\prod A_k$. Let μ denote the mean of $f(\mathbf{X})$, and let $\hat{r}^2$ denote the maximum sum of squared ranges. Then for any $t \geq 0$,*

$$\mathbf{P}(f(\mathbf{X}) - \mu \geq t) \leq e^{-2t^2/\hat{r}^2}.$$

More generally, let B be any 'bad' subset of $\prod A_k$, such that $R^2(\mathbf{x}) \leq r^2$ for each $\mathbf{x} \notin B$. Then

$$\mathbf{P}(f(\mathbf{X}) - \mu \geq t) \leq e^{-2t^2/r^2} + \mathbf{P}(\mathbf{X} \in B).$$

The first inequality above of course yields

$$\mathbf{P}(f(\mathbf{X}) - \mu \leq -t) \leq e^{-2t^2/\hat{r}^2}$$

by considering $-f$ (as in the comment after Theorem 3.1), and thus

$$\mathbf{P}(|f(\mathbf{X}) - \mu| \geq t) \leq 2e^{-2t^2/\hat{r}^2}. \tag{28}$$

If for each $k = 1, \ldots, n$ we let $\hat{r}_k$ be the supremum of the values $ran(x_1, \ldots, x_{k-1})$ over all choices of the x_i, then of course $\hat{r}^2$ is at most $\sum \hat{r}_k^2$. This bound for $\hat{r}^2$ yields Corollary 6.10 of [45]. Further, it yields also the independent bounded differences inequality, Theorem 3.1. For suppose that f satisfies the bounded differences condition (12) in that theorem. Let $1 \leq k \leq n$ and let $x_i \in A_i$ for $i = 1, \ldots, k-1$. We shall see that $ran(x_1, \ldots, x_{k-1}) \leq c_k$, so $\hat{r}^2 \leq \sum \hat{r}_k^2 \leq \sum c_k^2$, and then Theorem 3.1 follows. To see this, for each $x \in A_k$ let Z_x be the random variable $f(x_1, \ldots, x_{k-1}, x, X_{k+1}, \ldots, X_n)$. Then $|Z_x - Z_y| \leq c_k$. Hence, in the notation introduced before the statement of the last theorem, for any $x, y \in A_k$

$$|g(x) - g(y)| = |\mathbf{E}(Z_x) - \mathbf{E}(Z_y)| \leq \mathbf{E}(|Z_x - Z_y|) \leq c_k.$$

Thus $ran(x_1, \ldots, x_{k-1}) \leq c_k$, as required.

Observe that the above argument will in fact yield a slightly stronger form of Theorem 3.1. Denote $\sum c_k^2$ by c^2. The theorem will still hold if we weaken the assumption on f to the condition that for each $\mathbf{x}$ there exists a $\tilde{\mathbf{c}}$ (possibly depending on $\mathbf{x}$) such that $\sum \tilde{c}_k^2 \leq c^2$, and $|f(\mathbf{x}) - f(\mathbf{x}')| \leq \tilde{c}_k$ whenever the vectors $\mathbf{x}$ and $\mathbf{x}'$ differ only in the kth co-ordinate. The inequality of Talagrand that we shall meet later has a similar flavour.

Let us give one application of the above result, Theorem 3.7, before we go on to give extensions of the Bernstein theorem, Theorem 2.7. This application is from Maury [44], and was, together with [1], one of the first uses of a concentration inequality outside probability theory.

Permutation graphs

Let S_n denote the set of all $n!$ permutations or linear orders on $\{1, \ldots, n\}$. The *permutation graph* G_n has vertex set S_n, and two vertices σ and τ are adjacent when $\sigma\tau^{-1}$ is a transposition, that is when τ can be obtained from σ by swapping the order of two elements. We are interested in isoperimetric inequalities for this graph. Given a set $A \subseteq S_n$ and $t > 0$, the t-fattening A_t of A consists of the vertices in G_n at graph distance less than t from some vertex in A. Thus, we want lower bounds on $|A_t|$ in terms of $|A|$, or upper bounds on $1 - |A_t|/n!$. We shall show that

$$(|A|/n!)\,(1 - |A_t|/n!) \leq e^{-t^2/2n}. \tag{29}$$

Think of a linear order in S_n as an n-tuple $\mathbf{x} = (x_1, \ldots, x_n)$ where the x_k are distinct. Let $a_1, \ldots, a_k$ be distinct and let B be the set of linear orders $\mathbf{x} \in S_n$ such that $x_1 = a_1, \ldots, x_k = a_k$. For x distinct from the a_i let B_x be the set of $\mathbf{x} \in B$ with $x_{k+1} = x$. Let f be any function on S_n satisfying the Lipschitz or unit change condition $|f(\mathbf{x}) - f(\mathbf{y})| \leq 1$ if $\mathbf{x}$ and $\mathbf{y}$ are adjacent in G_n.

Now let X be uniformly distributed over S_n. In the notation introduced before the last theorem above, consider

$$g(x) = \mathbf{E}(f(\mathbf{X}) \mid \mathbf{X} \in B_x) - \mathbf{E}(f(\mathbf{X}) \mid \mathbf{X} \in B).$$

For any relevant distinct x and y, there is a bijection ϕ between B_x and B_y such that $\mathbf{x}$ and $\phi(\mathbf{x})$ are adjacent in G_n. (We simply swap the positions of x and y.) Thus $\mathbf{E}(f(\mathbf{X}) \mid \mathbf{X} \in B_y) = \mathbf{E}(f(\phi(\mathbf{X})) \mid \mathbf{X} \in B_x)$. It follows that

$$\begin{aligned} |g(x) - g(y)| &= |\mathbf{E}\,(f(\mathbf{X}) - f(\phi(\mathbf{X})) \mid \mathbf{X} \in B_x)\,| \\ &\leq \mathbf{E}(|f(\mathbf{X}) - f(\phi(\mathbf{X}))| \mid \mathbf{X} \in B_x) \;\leq\; 1. \end{aligned}$$

Hence by Theorem 3.7,

$$\mathbf{P}(f(\mathbf{X}) - \mathbf{E}(f(\mathbf{X})) \geq t) \leq e^{-2t^2/n^2}.$$

Now let us specialise to the case when $f(\mathbf{x})$ is the graph distance between $\mathbf{x}$ and the set A. We may proceed exactly as in the proof of Theorem 3.5 above (after the first two inequalities) to show (29) as required. For related results and extensions see for example [50, 10, 45, 37, 68].

The next result extends the Bernstein theorem, Theorem 2.7.

Theorem 3.8 *Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a family of random variables with X_k taking values in a set A_k, and let f be a real-valued function defined on $\prod A_k$. Let μ denote the mean of $f(\mathbf{X})$. Let $b = maxdev^+$ and let $\hat{v}$ be the maximum sum of variances, both of which we assume to be finite. Then for any $t \geq 0$,*

$$\mathbf{P}(f(\mathbf{X}) - \mu \geq t) \leq e^{-\frac{t^2}{2\hat{v}(1+(bt/3\hat{v}))}}.$$

More generally, let B be any 'bad' subset of $\prod A_k$ such that $V(\mathbf{x}) \leq v$ for each $\mathbf{x} \notin B$. Then

$$\mathbf{P}(f(\mathbf{X}) - \mu \geq t) \leq e^{-\frac{t^2}{2v(1+(bt/3v))}} + \mathbf{P}(\mathbf{X} \in B).$$

As with Theorem 2.7 above, in typical applications of this result the 'error term' $bt/3v$ is negligible. Also, the 'bad set' B if present at all is such that $\mathbf{P}(\mathbf{X} \in B)$ is negligible. If we use the bounds $V(\mathbf{x}) \leq R^2(\mathbf{x})/4$ for each $\mathbf{x}$ and $\hat{v} \leq \hat{r}^2/4$, we can nearly obtain the bound in Theorem 3.7 for small t. If for each $k = 1, \ldots, n$ we let $\hat{v}_k$ be the maximum of the values $var(x_1, \ldots, x_{k-1})$ over all choices of the x_i, then $\hat{v}$ is at most $\sum \hat{v}_k$. If we use this bound for $\hat{v}$ together with the discussion below, we obtain a result related to inequalities used by Kim [35] in his marvellous $R(3,t)$ paper. However, the present more general result is needed for certain applications – see for example [32, 2, 26] and the example below.

Observe that if a random variable X has mean 0 and takes only two values, with probabilities p and $1-p$, then the two values are $-pr$ and $(1-p)r$ where r is the range of X, and $var(X) = p(1-p)r^2 \leq pr^2$ – see also (26) above. Thus if p is small so is $var(X)$ and we can get tight bounds on deviations. Let us state one corollary of Theorem 3.8, which is a tightening of the martingale inequality in [2].

Theorem 3.9 *Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a family of random variables with X_k taking values in a set A_k, and let f be a bounded real-valued function defined on $\prod A_k$. Let μ denote the mean of $f(\mathbf{X})$, let b denote the maximum deviation maxdev, and let $\hat{r}^2$ denote the maximum sum of squared ranges. Suppose that, for any given values taken by $X_1, \ldots, X_{k-1}$, the random variable X_k takes at most two values, and if it can take two values then the smaller of the probabilities is at most p, where $p \leq \frac{1}{2}$. Then for any $t \geq 0$,*

$$\mathbf{P}(|f(\mathbf{X}) - \mu| \geq t) \leq 2e^{-\frac{t^2}{2p\hat{r}^2(1+(bt/3p\hat{r}^2))}}.$$

As with Theorems 2.7 and 3.8 above, we hope to be able to ignore the 'error term' $bt/3p\hat{r}^2$. The important term in the bound is $e^{-\frac{t^2}{2p\hat{r}^2}}$, which is significantly better (smaller) than the corresponding term $e^{-\frac{2t^2}{\hat{r}^2}}$ from Theorem 3.7 when $p = o(1)$. In the next subsection we describe an application where this difference is crucial.

3.2.1 An application to hypergraph matchings

A *matching* in H is a set of pairwise disjoint edges. Let $k \geq 3$ be a fixed integer, and consider a k-uniform d-regular simple hypergraph H on n vertices. (Thus each edge contains exactly k vertices, each vertex is contained in exactly d edges, and each pair of distinct edges meet in at most one vertex.) It is shown in [2] that such a hypergraph H contains a matching covering all but a vanishing proportion of the vertices as $n \to \infty$. (Earlier results showed that the proportion of vertices that could not be covered tended to zero, but perhaps slowly.)

The idea of the proof is to find such a matching by repeatedly taking random 'bites' (like large 'Rödl nibbles' – see for example [3]). We take such a bite as follows. Form a set X of edges by choosing the edges independently with probability $1/d$. Call an edge 'isolated' if it meets no other edge in X. Let M consist of the isolated edges in X – these will form part of the final matching. Now delete from H all the vertices in the edges in M and all the edges meeting these vertices, forming a hypergraph H^* on the vertex set V^*, and take the next bite from H^*. We must show that H^* is approximately regular of appropriately smaller degree. (Many details have been omitted, in particular a neat degree stabilisation technique, but they do not affect the idea that we wish to illustrate.) A key part of the proof is to check that each vertex degree in H^* is close to its expected value with high probability, and that is what we now proceed to do. (We need the probability of a significant deviation to be very small since the next step in the proof is to use the Lovász Local Lemma: when using a 'Rödl nibble' often a second moment bound suffices – see for example [3].)

For each vertex $v \in V$ let Z_v be the number of edges $E \in H$ containing v such that $E \setminus \{v\} \subseteq V^*$. Observe that if $v \in V^*$ then Z_v equals the degree of v in H^*. (By defining Z_v in this way we need not worry about whether or not the vertex v is in V^*.) It turns out that it suffices to consider a fixed vertex $v \in V$, and show that for $t = o(d^{\frac{1}{2}})$ we have

$$\mathbf{P}(|Z_v - \mathbf{E}(Z_v)| > td^{\frac{1}{2}}) \leq e^{-\Omega(t^2)}.$$

(See Claim 2 in [2].) Let us see how we can obtain this result from Theorem 3.9. Recall that Theorem 3.9 gives a bound of roughly $e^{-\frac{t^2}{2p\hat{r}^2}}$ as long as the deviation t is not too large.

For each edge $E \in H$, let the random variable $X_E = 1$ if E appears in the random set X and let $X_E = 0$ if not. Thus $\mathbf{P}(X_E = 1) = p = 1/d$, and we shall be in business as long as the maximum sum of squared ranges $\hat{r}^2 = \max_{\mathbf{x}} R^2(\mathbf{x})$ is $O(d^2)$. (In order to use Theorem 3.7 we could tolerate only $\hat{r}^2 = O(d)$, which is no use here.)

Call an edge in H *primary* if it contains the vertex v, *secondary* if it not primary but meets a primary edge, and *tertiary* if it is not primary or secondary but meets a secondary edge. Let $\mathcal{E}_1$, $\mathcal{E}_2$ and $\mathcal{E}_3$ denote the sets of primary, secondary and tertiary edges respectively, and note that $|\mathcal{E}_1| = d$, $|\mathcal{E}_2| \leq (k-1)d^2$ and $|\mathcal{E}_3| \leq (k-1)^2 d^3$. Let $\mathcal{E}$ be the union of the sets $\mathcal{E}_i$.

The random variable Z_v is determined by the values of the random variables X_E for $E \in \mathcal{E}$. Let Ω be the set of binary vectors $\mathbf{x}$ indexed by $\mathcal{E}$. For each $\mathbf{x} \in \Omega$ let $f(\mathbf{x})$ be the corresponding value of the degree Z_v. Let $\mathbf{x}, \mathbf{y} \in \Omega$ differ only in co-ordinate F, where $F \in \mathcal{E}$. If $F \in \mathcal{E}_1$ then $|f(\mathbf{x}) - f(\mathbf{y})| \leq 1$. If $F \in \mathcal{E}_2$ then $|f(\mathbf{x}) - f(\mathbf{y})| \leq k^2$. So far the contribution to the term $R^2(\mathbf{x})$ is at most

$$|\mathcal{E}_1| + |\mathcal{E}_2|k^4 \leq k^5 d^2 = O(d^2),$$

which as we saw above is small enough. Similarly, if $F \in \mathcal{E}_3$ then $|f(\mathbf{x}) - f(\mathbf{y})| \leq k^2$. However, we cannot tolerate a contribution to $R^2(\mathbf{x})$ of order d^3, so we must do better.

Let $\mathbf{x} \in \Omega$. Call an edge $F' \in \mathcal{E}_2$ *important* if $x_{F'} = 1$ and F' meets no other edges $F'' \in \mathcal{E}_2$ with $x_{F''} = 1$. There are at most $(k-1)d$ important edges, and so at most $k^2 d^2$ tertiary edges can meet an important edge. Further, if $\mathbf{y} \in \Omega$ differs from $\mathbf{x}$ only in co-ordinate F for some tertiary edge F which meets no important edge then $f(\mathbf{x}) = f(\mathbf{y})$. Thus we can bound $R^2(\mathbf{x})$ by $k^5 d^2 + (k^2 d^2)k^4 \leq 2k^6 d^2$, and so the

maximum sum of squared ranges $\hat{r}^2 \leq 2k^6d^2$. Since each $\mathbf{P}(X_F = 1) = 1/d$ we may now use Theorem 3.9 to show that

$$\begin{aligned}\mathbf{P}(|Z_v - \mathbf{E}(Z_v)| > td^{\frac{1}{2}}) &\leq 2\exp\left(-\frac{t^2 d}{2(2k^6 d)(1 + (k^2 t d^{\frac{1}{2}})/(6k^6 d))}\right) \\ &= 2\exp\left(-\frac{t^2}{4k^6(1 + t/(6k^4 d^{\frac{1}{2}}))}\right),\end{aligned}$$

and this bound is at most $e^{-\Omega(t^2)}$ for $t = O(d^{\frac{1}{2}})$.

3.3 Martingales

We give here a brief introduction to the theory of martingales, focussing on the case when the underlying probability space is finite. For much fuller introductions see for example [28] or [72].

The starting point is a probability space $(\Omega, \mathcal{F}, \mathbf{P})$. Thus Ω is the non-empty set of all 'elementary outcomes', $\mathcal{F}$ is the set of 'events', and $\mathbf{P}$ is the probability measure. The collection $\mathcal{F}$ of events must be suitably closed under unions, intersections and complements, and is assumed to be a σ-field. A *σ-field* on Ω is a collection $\mathcal{G}$ of subsets of Ω which contains the empty set, and is closed under complementation (if $A \in \mathcal{G}$ then $\Omega \setminus A \in \mathcal{G}$) and under countable unions (if $A_1, A_2, \ldots \in \mathcal{G}$ then their union is in $\mathcal{G}$). It follows that such a collection $\mathcal{G}$ is also closed under countable intersections. In many applications the underlying set Ω is finite, and the σ-field $\mathcal{F}$ of events is the collection of all subsets of Ω. Let us assume in the meantime that Ω is finite, though what we say is either true in general or at least tells the right story.

Corresponding to any σ-field $\mathcal{G}$ on Ω there is a partition of Ω into non-empty sets, the *blocks* of the partition, such that the σ-field $\mathcal{G}$ is the collection of all sets which are unions of blocks. Corresponding to the σ-field of all subsets of Ω is the partition of Ω into singleton blocks. Suppose that we have a σ-field $\mathcal{G}$ contained in $\mathcal{F}$. Any function on Ω which is constant on the blocks of $\mathcal{G}$ is called *$\mathcal{G}$-measurable*. A *random variable* is an $\mathcal{F}$-measurable real-valued function X defined on Ω, so that in the case when $\mathcal{F}$ consists of all subsets of Ω any real-valued function defined on Ω is a random variable.

The *expectation of X conditional on $\mathcal{G}$*, $\mathbf{E}(X \mid \mathcal{G})$, is the $\mathcal{G}$-measurable function where the constant value on each block of $\mathcal{G}$ is the average value of X on the block. This is a very important notion. We may see that $\mathbf{E}(X \mid \mathcal{F}) = X$ (that is, $\mathbf{E}(X \mid \mathcal{F})(\omega) = X(\omega)$ for each $\omega \in \Omega$) ; and if $\mathcal{G}$ is the trivial σ-field $\{\emptyset, \Omega\}$ corresponding to the trivial partition of Ω into one block, then $\mathbf{E}(X \mid \mathcal{G})$ is the constant function with constant value $\mathbf{E}(X)$. Key properties of conditional expectations that we shall need are that if $\mathcal{G}_1 \subseteq \mathcal{G}_2$ then

$$\mathbf{E}(\mathbf{E}(X \mid \mathcal{G}_2)) = \mathbf{E}(X \mid \mathcal{G}_1) \tag{30}$$

and so in particular

$$\mathbf{E}(\mathbf{E}(X \mid \mathcal{G})) = \mathbf{E}(X), \tag{31}$$

and

$$\mathbf{E}(XY \mid \mathcal{G}) = X\mathbf{E}(Y \mid \mathcal{G}) \quad \text{if } X \text{ is } \mathcal{G}\text{-measurable.} \tag{32}$$

The *supremum of* X in $\mathcal{G}$, $\sup(X \mid \mathcal{G})$, is the $\mathcal{G}$-measurable random variable which takes the value at ω equal to the maximum value of X over the block containing ω. Clearly

$$\mathbf{E}(X \mid \mathcal{G}) \leq \sup(X \mid \mathcal{G}), \tag{33}$$

and if $\mathcal{G}_1 \subseteq \mathcal{G}_2$ then

$$\sup(X \mid \mathcal{G}_2) \leq \sup(X \mid \mathcal{G}_1). \tag{34}$$

Note that each of the above results holds for each $\omega \in \Omega$. It is time for an example!

Example Let $\Omega = \{0,1\}^n$, let $\mathcal{F}$ be the collection of all subsets of Ω, let $0 < p < 1$, and for each $\omega = (\omega_1, \ldots, \omega_n)$ let $\mathbf{P}(\{\omega\}) = p^j(1-p)^{n-j}$ where $j = \sum \omega_k$. This defines our probability space. For each $k = 1, \ldots, n$ define $X_k(\omega) = \omega_k$ for each $\omega \in \Omega$. Then $X_1, \ldots, X_n$ are independent random variables with $\mathbf{P}(X_k = 1) = 1 - \mathbf{P}(X_k = 0) = p$ for each k. Also, let $S_k = X_1 + \cdots + X_k$. Let $\mathcal{F}_k$ be the σ-field corresponding to the partition of Ω into the 2^k blocks $\{\omega \in \Omega : \omega_1 = x_1, \ldots, \omega_k = x_k\}$ for each $(x_1, \ldots, x_k) \in \{0,1\}^k$. Then the random variable $\mathbf{E}(S_n \mid \mathcal{F}_k)$ satisfies (for each $\omega \in \Omega$)

$$\mathbf{E}(S_n \mid \mathcal{F}_k) = S_k + (n-k)p = \omega_1 + \cdots + \omega_k + (n-k)p,$$

and $\mathbf{E}(S_n \mid \mathcal{F}_n) = S_n$, $\mathbf{E}(S_n \mid \mathcal{F}_0) = \mathbf{E}(S_n) = np$ and $\mathbf{E}(\mathbf{E}(S_n \mid \mathcal{F}_k)) = \mathbf{E}(S_k) + (n-k)p = np$. Also for example

$$\mathbf{E}(S_k S_n \mid \mathcal{F}_k) = S_k \mathbf{E}(S_n \mid \mathcal{F}_k) = S_k^2 + (n-k)pS_k.$$

Further

$$\sup(S_n \mid \mathcal{F}_k) = S_k + (n-k) \leq S_{k-1} + (n-k+1) = \sup(S_n \mid \mathcal{F}_{k-1}).$$

Another important idea is that of a filter. A nested sequence $(\emptyset, \Omega) = \mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots$ of σ-fields contained in $\mathcal{F}$ is called a *filter*. This corresponds (in the finite case) to a sequence of increasingly refined partitions of Ω, starting with the trivial partition into one block. We may think of the filter as corresponding to acquiring information as time goes on: at time k, we know which block of the partition corresponding to $\mathcal{F}_k$ contains our random elementary outcome ω. Given a filter, a sequence $X_0, X_1, X_2, \ldots$ of random variables is called a *martingale* if $\mathbf{E}(X_{k+1} \mid \mathcal{F}_k) = X_k$ for each $k = 0, 1, \ldots$. This implies that X_k is $\mathcal{F}_k$-measurable ('at time k we know the value of X_k'). It also implies that $\mathbf{E}(X_k) = \mathbf{E}(X)$ for each k. A sequence $Y_1, Y_2, \ldots$ of random variables is called a *martingale difference sequence* if Y_k is $\mathcal{F}_k$-measurable and $\mathbf{E}(Y_k \mid \mathcal{F}_{k-1}) = 0$ for each $k = 1, 2, \ldots$.

From a martingale $X_0, X_1, X_2, \ldots$ we obtain a martingale difference sequence by setting $Y_k = X_k - X_{k-1}$; and conversely from X_0 and a martingale difference sequence we obtain a martingale $X_0, X_1, X_2, \ldots$ by setting $X_k = X_0 + \sum_{i=1}^{k} Y_i$. Thus we may focus on either form.

We shall be interested here only in finite filters $(\emptyset, \Omega) = \mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}_n$ where $\mathcal{F}_n \subseteq \mathcal{F}$. Let X be a random variable and define $X_k = \mathbf{E}(X \mid \mathcal{F}_k)$ for $k = 0, 1, \ldots, n$. Then $X_0, X_1, \ldots, X_n$ is a martingale, with $X_0 = \mathbf{E}(X)$ and $X_n = X$ if X is $\mathcal{F}_n$-measurable. This is called Doob's martingale process, and for finite filters all corresponding martingales may be obtained in this way. If $Y_1, \ldots, Y_n$ is the corresponding martingale difference sequence then we have $X - \mathbf{E}(X) = \sum Y_k$.

Example (continued) There is a natural filter here, namely

$$\{\Omega, \emptyset\} = \mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}_n = \mathcal{F},$$

which corresponds to learning the values of the co-ordinates of ω one by one. The σ-field $\mathcal{F}_k$ is the σ-field *generated by* the random variables $X_1, \ldots, X_k$; that is, the smallest σ-field $\mathcal{G}$ such that each of $X_1, \ldots, X_k$ is $\mathcal{G}$-measurable. For each $k = 1, \ldots, n$ let T_k be the random variable $S_k - kp = (X_1 - p) + \cdots + (X_k - p)$. Then $\mathbf{E}(T_k \mid \mathcal{F}_{k-1}) = T_{k-1}$, and so the random variables T_k form a martingale, with corresponding martingale difference sequence $X_k - p$.

When the underlying set Ω is infinite we need to be a little more careful. In particular, the results discussed above hold with probability 1 (also called 'almost surely') rather than for *every* $\omega \in \Omega$; and we need to assume that various expectations are finite. However, the sketch introduction above should give the right ideas.

The most basic inequality for a bounded martingale difference sequence is the following lemma of Hoeffding (1963) [29], Azuma (1967) [6], which we shall refer to as the 'Hoeffding-Azuma inequality'.

Theorem 3.10 *Let $c_1, \ldots, c_n$ be constants, and let $Y_1, \ldots, Y_n$ be a martingale difference sequence with $|Y_k| \leq c_k$ for each k. Then for any $t \geq 0$,*

$$\mathbf{P}(|\sum Y_k| \geq t) \leq 2e^{-t^2/2\sum c_k^2}.$$

Suppose that $X_1, \ldots, X_n$ are independent, with $\mathbf{P}(X_k = 1) = p$ and $\mathbf{P}(X_k = 0) = 1 - p$. Set $Y_k = X_k - p$ and $c_k = \max(p, 1-p)$. We may then apply the above lemma to obtain the bound in Theorem 2.1, except that the bound is weakened if $p \neq \frac{1}{2}$. All our applications will be based on less symmetrical forms of the above result, and will thus avoid gratuitous factors less than 1 in the exponent in the bounds. In particular, Theorem 3.10 is a special case of Theorem 3.13 below.

3.4 Martingale results

The results in this subsection extend all the earlier results. In particular, the next result extends Lemma 2.2 on independent random variables.

Lemma 3.11 *Let $Y_1, Y_2, \ldots, Y_n$ be a martingale difference sequence with $-a_k \leq Y_k \leq 1 - a_k$ for each k, for suitable constants a_k. Let $a = \frac{1}{n}\sum a_k$ and let $\bar{a} = 1 - a$. Then for any $0 \leq t < \bar{a}$,*

$$\mathbf{P}(\sum Y_k \geq nt) \leq \left(\left(\frac{a}{a+t} \right)^{a+t} \left(\frac{\bar{a}}{\bar{a}-t} \right)^{\bar{a}-t} \right)^n. \tag{35}$$

Proof Since $S_n = S_{n-1} + Y_n$ and S_{n-1} is $\mathcal{F}_{n-1}$-measurable (and hence so is $e^{S_{n-1}}$), we may use (31) and (32) to show that for any h,

$$\mathbf{E}(e^{hS_n}) = \mathbf{E}\left(e^{hS_{n-1}} e^{hY_n}\right) = \mathbf{E}\left(e^{hS_{n-1}} \mathbf{E}(e^{hY_n} \mid \mathcal{F}_{n-1})\right).$$

Thus as in the proof of Lemma 2.2, for any $h > 0$,

$$\begin{aligned} \mathbf{E}(e^{hS_n}) &= \mathbf{E}\left(e^{hS_{n-1}} \mathbf{E}(e^{hY_n} \mid \mathcal{F}_{n-1})\right) \\ &\leq \mathbf{E}\left(e^{hS_{n-1}}\right) \left((1-a_n)e^{-ha_n} + a_n e^{h(1-a_n)}\right) \\ &\leq \prod \left((1-a_k)e^{-ha_k} + a_k e^{h(1-a_k)}\right), \end{aligned}$$

on iterating, and we may complete the proof exactly as for Lemma 2.2. □

We may deduce more useful inequalities from this lemma, just as we obtained Theorem 2.3 from Lemma 2.2.

Theorem 3.12 *Let $Y_1, Y_2, \ldots, Y_n$ be a martingale difference sequence with $-a_k \leq Y_k \leq 1 - a_k$ for each k, for suitable constants a_k; and let $a = \frac{1}{n}\sum a_k$.*
(a) For any $t \geq 0$,

$$\mathbf{P}(|\sum Y_k| \geq t) \leq 2e^{-2t^2/n}.$$

(b) For any $\epsilon > 0$,

$$\mathbf{P}(\sum Y_k \geq \epsilon an) \leq e^{-((1+\epsilon)\ln(1+\epsilon)-\epsilon)an} \leq e^{-\frac{\epsilon^2 an}{2(1+\epsilon/3)}}.$$

(c) For any $\epsilon > 0$,

$$\mathbf{P}(\sum Y_k \leq -\epsilon an) \leq e^{-\frac{1}{2}\epsilon^2 an}.$$

To deduce Theorem 2.3 from Theorem 3.12, let $a_k = \mathbf{E}(X_k)$ and $Y_k = X_k - a_k$, so that $-a_k \leq Y_k \leq 1 - a_k$: then $\mu = \sum a_k = na$, $p = a$ and $\sum Y_k = S_n - \mu$. The next result extends both the independent bounded differences inequality, Theorem 3.1, and the Hoeffding-Azuma inequality, Theorem 3.10.

Theorem 3.13 *Let $Y_1, \ldots, Y_n$ be a martingale difference sequence with $a_k \leq Y_k \leq b_k$ for each k, for suitable constants a_k, b_k. Then for any $t \geq 0$,*

$$\mathbf{P}(|\sum Y_k| \geq t) \leq 2e^{-2t^2/\sum (b_k - a_k)^2}. \tag{36}$$

The next pair of results, Theorems 3.14 and 3.15, are the most powerful of the martingale results we present, and include all the previous theorems (except for the first inequality in part (b) of Theorem 2.3 and of Theorem 3.12). In particular, Theorem 3.13 will follow immediately from Theorem 3.14. In order to state the two results we need some more definitions and notation. We postpone their proofs to the next subsection.

Let X be a bounded random variable and let $\mathcal{G}$ be a σ-field contained in the σ-field $\mathcal{F}$ of all events. The *conditional range of* X in $\mathcal{G}$, $ran(X \mid \mathcal{G})$, is the $\mathcal{G}$-measurable function $\sup(X \mid \mathcal{G}) + \sup(-X \mid \mathcal{G})$. The *conditional variance of* X in $\mathcal{G}$, $var(X \mid \mathcal{G})$, is $\mathbf{E}((X - Y)^2 \mid \mathcal{G})$, where $Y = \mathbf{E}(X \mid \mathcal{G})$. In the example in the last subsection, the conditional range of S_n in $\mathcal{F}_k$, $ran(S_n \mid \mathcal{F}_k)$, is the constant function $n - k$, and the conditional variance $var(S_n \mid \mathcal{F}_k)$ is the constant function $(n - k)p(1 - p)$.

Now let $(\emptyset, \Omega) = \mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}_n$ be a filter in $\mathcal{F}$. Let the bounded random variable X be $\mathcal{F}_n$-measurable, and let $X_0, X_1, \ldots, X_n$ be the martingale obtained by setting $X_k = \mathbf{E}(X \mid \mathcal{F}_k)$. Further, let $Y_1, \ldots, Y_n$ be the corresponding martingale difference sequence obtained by setting $Y_k = X_k - X_{k-1}$. For $1 \leq k \leq n$, we define four $\mathcal{F}_{k-1}$-measurable functions ran_k, dev_k^+, dev_k and var_k as follows. We let ran_k denote $ran(Y_k \mid \mathcal{F}_{k-1})$ $(= ran(X_k \mid \mathcal{F}_{k-1}))$; let dev_k^+ denote $\sup(Y_k \mid \mathcal{F}_{k-1})$, let dev_k denote $\sup(|Y_k| \mid \mathcal{F}_{k-1})$, and finally we let var_k denote $var(Y_k \mid \mathcal{F}_{k-1})$, $(= var(X_k \mid \mathcal{F}_{k-1}))$. Note that $dev_k^+ \leq dev_k \leq ran_k \leq 2dev_k$, and $var_k \leq (1/4)ran_k^2$ by (26).

Finally we define two random variables R^2 and V and four constants $\hat{r}^2$, $\hat{v}$, $\max dev^+$ and $\max dev$. Let the *sum of squared conditional ranges* R^2 be the random variable

$\sum ran_k^2$, and let the *maximum sum of squared conditional ranges* $\hat{r}^2$ be the (essential) supremum of the random variable R^2. Let the *sum of conditional variances* V be the random variable $\sum var_k$, and let the *maximum sum of conditional variances* $\hat{v}$ be the supremum of the random variable V. Finally let the *maximum conditional positive deviation* $\max dev^+$ be the supremum over all k of the random variables dev_k^+, and let the *maximum conditional deviation* $\max dev$ be the supremum over all k of the random variables dev_k.

The random variable V is also called the 'predictable quadratic variation' of the martingale (X_k), see for example [61], or the 'increasing sequence' associated with (X_k), see for example [20]. Note that

$$\begin{aligned}
\mathbf{E}(V) &= \mathbf{E}\left(\sum_{k=1}^{n} \mathbf{E}\left((X_k - X_{k-1})^2 \mid \mathcal{F}_{k-1}\right)\right) \\
&= \mathbf{E}\left(\sum_{k=1}^{n} (\mathbf{E}(X_k^2 \mid \mathcal{F}_{k-1}) - X_{k-1}^2)\right) \\
&= \sum_{k=1}^{n} (\mathbf{E}(X_k^2) - \mathbf{E}(X_{k-1}^2)) \\
&= \mathbf{E}(X_n^2) - \mathbf{E}(X_0^2) \;=\; var(X).
\end{aligned}$$

Theorem 3.14 *Let X be a bounded random variable with $\mathbf{E}(X) = \mu$, and let $(\emptyset, \Omega) = \mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}_n$ be a filter in $\mathcal{F}$. Then for any $t \geq 0$,*

$$\mathbf{P}(X - \mu \geq t) \leq e^{-2t^2/\hat{r}^2}, \tag{37}$$

where $\hat{r}^2$ is the maximum sum of squared conditional ranges. More generally, for any $t \geq 0$ and any value r^2,

$$\mathbf{P}((X - \mu \geq t) \wedge (R^2 \leq r^2)) \leq e^{-2t^2/r^2}, \tag{38}$$

where the random variable R^2 is the sum of squared conditional ranges.

The earlier result Theorem 3.7 is essentially this result when the σ-field $\mathcal{F}_k$ is the σ-field generated by $X_1, \ldots, X_k$. Suppose that for each $k = 1, \ldots, n$, we let $\hat{r}_k$ be the supremum of the values $ran(x_1, \ldots, x_{k-1})$ over all choices of the x_i. (This corresponds to our earlier use of the notation $\hat{r}_k$ immediately after Theorem 3.7.) Then $\hat{r}^2$ is at most $\sum \hat{r}_k^2$. If we use this bound for $\hat{r}^2$ in Theorem 3.14 above we obtain Theorem (6.7) of [45], which extends Theorem 3.13 above. The next result extends the earlier results that use bounds on the variance, namely Theorems 2.7 and Theorem 3.8 (and thus Theorem 3.9), and is close to Theorem 4.1 in [21] – see also [32, 2, 26].

Theorem 3.15 *Let X be a random variable with $\mathbf{E}(X) = \mu$, and let $(\emptyset, \Omega) = \mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}_n$ be a filter in $\mathcal{F}$. Let $b = \max dev^+$, the maximum conditional positive deviation (and assume that b is finite). Then for any $t \geq 0$,*

$$\mathbf{P}(X - \mu \geq t) \leq e^{-\frac{t^2}{2\hat{v}(1+(bt/3\hat{v}))}}, \tag{39}$$

where $\hat{v}$ is the maximum sum of conditional variances (which is assumed to be finite). More generally, for any $t \geq 0$ and any value $v \geq 0$,

$$\mathbf{P}((X - \mu \geq t) \wedge (V \leq v)) \leq e^{-\frac{t^2}{2v(1+(bt/3v))}}, \tag{40}$$

where the random variable V is the sum of conditional variances.

As with the earlier results of this form, we think of the term $(bt/3v)$ as a negligible error term. To complete the proofs of all the results given above it suffices to prove the last two results. We do this in the next subsection.

3.5 Remaining proofs for martingale results

The following lemma is partly based on Lemma 3.4 of Kahn [32]. The lemma itself (in a special case) is used, rather than one of the theorems derived from it, in the proofs in [49] concerning the concentration of the number of comparisons used by quicksort. We shall always take $\mathcal{F}_0$ as the trivial σ-field $(\emptyset, \Omega)$ when we use the lemma, but we allow any $\mathcal{F}_0$ to give an easy induction.

Lemma 3.16 *Let $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}_n$ be a filter in $\mathcal{F}$, and let $Y_1, \ldots, Y_n$ be a corresponding martingale difference sequence, where each Y_k is bounded above. Let the random variable Z be the indicator of some event. Then for any h,*

$$\mathbf{E}(Ze^{h\sum Y_k} \mid \mathcal{F}_0) \le \sup\left(Z \prod \mathbf{E}(e^{hY_k} \mid \mathcal{F}_{k-1}) \mid \mathcal{F}_0\right).$$

Proof We use induction on n. The case $n = 0$ is trivial, since it asserts that $\mathbf{E}(Z \mid \mathcal{F}_0) \le \sup(Z \mid \mathcal{F}_0)$ as in (33). Now let $n \ge 1$ and suppose that the result holds for $n-1$. Let

$$A = Ze^{h\sum_{k=2}^{n} Y_k}$$

and

$$B = Z \prod_{k=2}^{n} \mathbf{E}(e^{hY_k} \mid \mathcal{F}_{k-1}).$$

Then by the induction hypothesis, $\mathbf{E}(A \mid \mathcal{F}_1) \le \sup(B \mid \mathcal{F}_1)$; and $\sup(B \mid \mathcal{F}_1) \le \sup(B \mid \mathcal{F}_0)$ as in (34). Hence

$$\begin{aligned}
\mathbf{E}(Ze^{h\sum_{k=1}^{n} Y_k} \mid \mathcal{F}_0) &= \mathbf{E}\left(e^{hY_1}\mathbf{E}(A \mid \mathcal{F}_1) \mid \mathcal{F}_0\right) \\
&\le \mathbf{E}\left(e^{hY_1} \sup(B \mid \mathcal{F}_0) \mid \mathcal{F}_0\right) \\
&= \sup(B \mid \mathcal{F}_0)\mathbf{E}(e^{hY_1} \mid \mathcal{F}_0) \quad \text{as in (32)} \\
&= \sup\left(Z \prod_{k=1}^{n} \mathbf{E}(e^{hY_k} \mid \mathcal{F}_{k-1}) \mid \mathcal{F}_0\right),
\end{aligned}$$

which completes the induction step. □

Proof of Theorem 3.14 Let $Y_1, \ldots, Y_n$ be the corresponding martingale difference sequence. Let the random variable Z be the indicator of the event that $R^2 \le r^2$, so that $0 \le ZR^2 \le r^2$. For any h, by Lemma 2.6,

$$\mathbf{E}(e^{hY_k} \mid \mathcal{F}_{k-1}) \le e^{\frac{1}{8}h^2 r_k^2}.$$

Hence by Lemma 3.16,

$$\begin{aligned}\mathbf{E}(Ze^{h(X-\mu)}) &\leq \sup\left(Z\prod e^{\frac{1}{8}h^2r_k^2}\right)\\ &= \sup(Ze^{\frac{1}{8}h^2R^2})\\ &\leq e^{\frac{1}{8}h^2\sup(ZR^2)}\\ &\leq e^{\frac{1}{8}h^2r^2}.\end{aligned}$$

Thus for any $h > 0$, by Markov's inequality,

$$\begin{aligned}\mathbf{P}((X-\mu \geq t) \wedge (R^2 \leq r^2)) &= \mathbf{P}(Ze^{h(X-\mu)} \geq e^{ht})\\ &\leq e^{-ht}\mathbf{E}(Ze^{h(X-\mu)})\\ &\leq e^{-ht+\frac{1}{8}h^2r^2}\\ &= e^{-2t^2/r^2}\end{aligned}$$

when $h = 4t/r^2$. □

Proof of Theorem 3.15 Let $Y_1, \ldots, Y_n$ be the corresponding martingale difference sequence. Note that $Y_k \leq b$ for each k. Let the random variable Z be the indicator of the event that $V \leq v$, so that $0 \leq ZV \leq v$. Now as in the proof of Theorem 2.7 we use Lemma 2.8, and the function $g(x)$ defined there. We find that, for any $h > 0$,

$$\mathbf{E}(e^{hY_k} \mid \mathcal{F}_{k-1}) \leq e^{h^2g(hdev_k^+)var_k} \leq e^{h^2g(hb)var_k}.$$

Hence by Lemma 3.16,

$$\begin{aligned}\mathbf{E}(Ze^{h(X-\mu)}) &\leq \sup\left(Z\prod e^{h^2g(hb)var_k}\right)\\ &= \sup\left(Ze^{h^2g(hb)V}\right)\\ &\leq e^{h^2g(hb)\sup(ZV)}\\ &\leq e^{h^2g(hb)v}.\end{aligned}$$

But now as in the proof of the last theorem,

$$\begin{aligned}\mathbf{P}((X-\mu \geq t) \wedge (V \leq v)) &\leq e^{-ht}\mathbf{E}(Ze^{h(X-\mu)})\\ &\leq e^{-ht+h^2g(hb)v},\end{aligned}$$

and we may complete the proof as for Theorem 2.7. □

Inequalities for maxima

We now amplify the comment at the end of Section 2 on maxima. Let $Y_1, \ldots, Y_n$ be a martingale difference sequence and let $S_k = Y_1 + \cdots + Y_k$ as usual. Let $h > 0$ and let $T_k = e^{hS_k}$. Then $T_1, \ldots, T_n$ form a submartingale (as long as the T_k are integrable), so we may apply Doob's maximal inequality for submartingales – see for example [28] section 12.6 or [72] section 14.6. We find that for any $t \geq 0$,

$$\mathbf{P}(\max S_k \geq t) = \mathbf{P}(\max T_k \geq e^{ht}) \leq e^{-ht}\mathbf{E}(T_n) = e^{-ht}\mathbf{E}(e^{hS_n}).$$

Thus all the martingale results based directly on the Bernstein inequality may be strengthened immediately to refer to maxima, just like those in Section 2, as noted on [29] (see also [64, 65, 66]).

This comment applies to Lemma 3.11 and Theorems 3.12 and 3.13 (and thus also to Theorem 3.10), and to the inequalities (37) and (39) In particular for example, in Theorem 3.13 the inequality (36) may be strengthened to read that for any $t \geq 0$,

$$\mathbf{P}(|\max(\sum_{i=1}^{k} Y_i)| \geq t) \leq 2e^{-2t^2/\sum(b_k-a_k)^2}, \tag{41}$$

where the maximum is over $k = 1, \ldots, n$.

3.6 Centering sequences

Given a sequence $X_1, X_2, \ldots$ of random variables, the corresponding difference sequence is $Y_1, Y_2, \ldots$ where $Y_k = X_k - X_{k-1}$ (and where we set $X_0 \equiv 0$). Let $\mu_k(x) = \mathbf{E}(Y_k \mid X_{k-1} = x)$. We call the distribution of the sequence *centering* if for each $k = 2, 3, \ldots$ $\mu_k(x)$ is a non-increasing function of x – see [47]. Observe that a martingale is trivially centering, since $\mu_k(x) \equiv 0$.

The basic inequalities discussed above for a martingale difference sequence may be extended to centering sequences with bounded differences. The most fundamental example for the martingale inequalities involves the binomial distribution, as in Theorem 2.1. Now we can include the hypergeometric distribution naturally in the same inequalities – see also [29, 15].

Let $(x_1, \ldots, x_n) \in \{0,1\}^n$ with $\sum x_k = r$. Let $(Z_1, \ldots, Z_n)$ be a random linear order on the set $\{1, \ldots, n\}$, where all $n!$ such orders are equally likely. Let $Y_j = x_{Z_j}$ and $X_k = \sum_{j=1}^{k} Y_j$. Then X_k has the hypergeometric distribution, corresponding to counting the red elements in a random sample picked without replacement from the set $\{1, \ldots, n\}$ with r elements painted red. We are interested in the concentration of X_k. Note that $\mathbf{E}(X_k) = rk/n$. But the sequence $X_1, X_2, \ldots, X_n$ is centering, since

$$\mu_k(x) = \mathbf{E}(Y_k \mid X_{k-1} = x) = \frac{r-x}{n-k+1},$$

which is a decreasing function of x. From the centering version in [47] of Theorem 2.3(c) above, it follows for example that, if μ denotes $\mathbf{E}(X_k)$, then for any $\epsilon > 0$,

$$\mathbf{P}(X_k \leq (1-\epsilon)\mu) \leq e^{-\frac{1}{2}\epsilon^2\mu}.$$

If we try to apply here the inequalities for martingales with bounded differences in the natural way (that is, with $\mathcal{F}_k$ as the σ-field generated by revealing the first k elements picked), we obtain an unwanted factor < 1 in the exponent in the bound. Centering sequences also arise naturally in occupancy or 'balls in boxes' problems – see [33, 47].

4 Talagrand's inequality

4.1 The inequality

Let $\Omega_1, \ldots, \Omega_n$ be probability spaces, and let Ω denote the product space. Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a family of independent random variables with X_k taking values in

Ω_k. We saw earlier that for any subset A of Ω such that $\mathbf{P}(\mathbf{X} \in A)$ is not too small, with high probability a random point $\mathbf{X}$ is close to A, when we consider Hamming distance or generalised Hamming distance. It turns out to be very fruitful to consider a related notion of distance.

Let $\alpha = (\alpha_1, \ldots, \alpha_n) \geq \mathbf{0}$ be an n-vector of non-negative real numbers. Recall that for points $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ in Ω, the *α-Hamming distance* $d_\alpha(\mathbf{x}, \mathbf{y})$ is the sum of the values α_i over those indices i such that $x_i \neq y_i$; and for a subset A of Ω, $d_\alpha(\mathbf{x}, A) = \inf\{d_\alpha(\mathbf{x}, \mathbf{y}) : \mathbf{y} \in A\}$. Talagrand's *convex distance* $d_T(\mathbf{x}, A)$ is defined to be $\sup(d_\alpha(\mathbf{x}, A))$ where the supremum is over all choices of non-negative unit n-vector α (that is, with $\|\alpha\| = 1$).

By considering the n-vector α with each co-ordinate $1/\sqrt{n}$, we see that $d_T(\mathbf{x}, A) \geq d_\alpha(\mathbf{x}, A) = (1/\sqrt{n})\, d_H(\mathbf{x}, A)$, so upper bounds on $d_T(\mathbf{x}, A)$ give us upper bounds on $d_H(\mathbf{x}, A)$, but we shall see that they will tell us much more. The reason for the name 'convex distance' will emerge later. Talagrand [68] in fact considers also other notions of distance (see also [70]), but we shall focus only on the convex distance. We call the following fundamental result 'Talagrand's inequality'.

Theorem 4.1 *Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a family of independent random variables and let A be a subset of the product space. Then for any $t \geq 0$,*

$$\mathbf{P}(\mathbf{X} \in A)\, \mathbf{P}(d_T(\mathbf{X}, A) \geq t) \leq e^{-t^2/4}. \tag{42}$$

If we consider a single non-negative unit vector α, then $d_T \geq d_\alpha$ and the above result yields a form of Theorem 3.6, but it is in fact far more powerful since it refers simultaneously to all possible generalised Hamming distances, as will be evident from the applications below. We shall see that this power is most evident when we consider the concentration of a function $f(\mathbf{X})$ where an inequality $f(\mathbf{x}) \geq b$ typically can be verified by examining only a few of the co-ordinate values x_i, and for different vectors $\mathbf{x}$ we may examine different co-ordinates. In some applications we profit greatly from the flexibility of choosing an appropriate unit vector α for each $\mathbf{x}$, rather than having to consider say Hamming distance. Often we shall choose α to put more weight on the 'important' or the 'awkward' co-ordinates of $\mathbf{x}$.

Note that we must assume that the random variables X_k are independent, in contrast to the situation with the martingale results (but see the recent paper of Marton [42], which gives an extension of Talagrand's inequality in which a limited dependence is allowed). Theorems 4.3 and 4.5 below are useful specialisations of Talagrand's inequality, on which we base all the applications here. We shall prove Theorem 4.1 later, but before that let us consider some applications.

4.2 Some applications

4.2.1 Subsequences and configuration functions

Given a sequence $\mathbf{x} = (x_1, \ldots, x_n)$ of real numbers, we let $inc(\mathbf{x})$ denote the length of a longest increasing subsequence. Thus $inc(\mathbf{x})$ is the maximum value of $|K|$ over all subsets K of $\{1, \ldots, n\}$ such that the corresponding subsequence $(x_i : i \in K)$ is increasing, that is $x_i < x_j$ whenever $i, j \in K$ with $i < j$.

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a family of independent random variables each taking real values. We are interested in the concentration of the random variable $inc(\mathbf{X})$. Let

μ be the mean of $inc(\mathbf{X})$. It follows directly from the independent bounded differences inequality, Theorem 3.1, that for any $t \geq 0$,

$$\mathbf{P}(|inc(\mathbf{X}) - \mu| \geq t) \leq e^{-2t^2/n}. \tag{43}$$

This shows that for large n, with high probability $inc(\mathbf{X})$ is confined within an interval of length $O(\sqrt{n})$. Using Talagrand's inequality we can deduce a much improved result. Let m be a median for $inc(\mathbf{X})$.

Theorem 4.2 *For any $t \geq 0$,*

$$\mathbf{P}(inc(\mathbf{X}) \geq m + t) \leq 2e^{-t^2/4(m+t)} \tag{44}$$

and

$$\mathbf{P}(inc(\mathbf{X}) \leq m - t) \leq 2e^{-t^2/4m}. \tag{45}$$

With ingenuity and endeavour, the bounded differences method will give nearly as good results – see [13]. It is known (see for example [63]) that, when the random variables X_k all have the same continuous distribution, the median $m \sim 2\sqrt{n}$ as $n \to \infty$. Thus the above result shows that with high probability $inc(\mathbf{X})$ is confined within an interval of length $O(n^{\frac{1}{4}})$, which is the best bound known. (In particular, the mean μ and the median m must be within $O(n^{\frac{1}{4}})$ of each other – see Lemma 4.6 below.)

It turns out that the approach based on Talagrand's inequality to the longest increasing subsequence problem will handle a general class of problems. Observe that the function $f(\mathbf{x}) = inc(\mathbf{x})$ has the following property. For each $\mathbf{x} \in \Omega$ there is a subset $K = K(\mathbf{x})$ of the index set $\{1, \ldots, n\}$ such that $f(\mathbf{x}) = |K|$, and for each $\mathbf{y} \in \Omega$ we have

$$f(\mathbf{y}) \geq |\{i \in K : y_i = x_i\}| = f(\mathbf{x}) - |\{i \in K : y_i \neq x_i\}|.$$

Thus for each $\mathbf{x} \in \Omega$ there is a non-negative unit n-vector α (namely the incidence vector of the set $K(\mathbf{x})$ scaled by dividing by $\sqrt{f(\mathbf{x})}$) such that, for each $\mathbf{y} \in \Omega$ we have

$$f(\mathbf{y}) \geq f(\mathbf{x}) - \sqrt{f(\mathbf{x})} d_\alpha(\mathbf{x}, \mathbf{y}).$$

This is the key property. We call a function f defined on a set Ω of n-vectors a *c-configuration function* if it has the following property: for each $\mathbf{x} \in \Omega$ there is a non-negative unit n-vector α such that, for each $\mathbf{y} \in \Omega$ we have

$$f(\mathbf{y}) \geq f(\mathbf{x}) - \sqrt{c\, f(\mathbf{x})} d_\alpha(\mathbf{x}, \mathbf{y}).$$

Thus $inc(\mathbf{x})$ gives a 1-configuration function, and so the next result extends the last one. (We shall give a related example below concerning common subsequences. Also we shall discuss concentration around the mean rather than the median in the next subsection – see Lemma 4.6.)

Theorem 4.3 *Let f be a c-configuration function, and let m be a median for $f(\mathbf{X})$. Then for any $t \geq 0$,*

$$\mathbf{P}(f(\mathbf{X}) \geq m + t) \leq 2e^{-t^2/4c(m+t)} \tag{46}$$

and

$$\mathbf{P}(f(\mathbf{X}) \leq m - t) \leq 2e^{-t^2/4cm}. \tag{47}$$

Proof

Let $\mathbf{x} \in \Omega$, and let α be a non-negative unit n-vector such that, for any $\mathbf{y} \in \Omega$,

$$f(\mathbf{x}) \leq f(\mathbf{y}) + \sqrt{cf(\mathbf{x})}d_\alpha(\mathbf{x}, \mathbf{y}).$$

Let $A_a = \{\mathbf{y} \in \Omega : f(\mathbf{y}) \leq a\}$. Then by the above

$$f(\mathbf{x}) \leq a + \sqrt{cf(\mathbf{x})}d_\alpha(\mathbf{x}, \mathbf{y})$$

for each $\mathbf{y} \in A_a$, and so by minimising over such $\mathbf{y}$ we have

$$f(\mathbf{x}) \leq a + \sqrt{cf(\mathbf{x})}d_\alpha(\mathbf{x}, A_a) \leq a + \sqrt{cf(\mathbf{x})}d_T(\mathbf{x}, A_a).$$

Thus if $f(\mathbf{x}) \geq a + t$ then

$$d_T(\mathbf{x}, A_a) \geq \frac{f(\mathbf{x}) - a}{\sqrt{c}\sqrt{f(\mathbf{x})}} \geq \frac{t}{\sqrt{c}\sqrt{a+t}},$$

since the function $g(t) = (t-a)/\sqrt{t}$ is increasing for $t \geq a$. Thus for any $t \geq 0$,

$$\mathbf{P}(f(\mathbf{X}) \geq a + t) \leq \mathbf{P}\left(d_T(\mathbf{X}, A_a) \geq \frac{t}{\sqrt{c(a+t)}}\right).$$

Hence by Talagrand's inequality, for any $t \geq 0$

$$\begin{aligned} & \mathbf{P}(f(\mathbf{X}) \leq a)\mathbf{P}(f(\mathbf{X}) \geq a + t) \\ \leq\ & \mathbf{P}(\mathbf{X} \in A_a)\mathbf{P}\left(d_T(\mathbf{X}, A_a) \geq \frac{t}{\sqrt{c(a+t)}}\right) \\ \leq\ & e^{-\frac{t^2}{4c(a+t)}}. \end{aligned}$$

Now we may complete the proof by appropriate choices of a in this last inequality. If we let $a = m$, then since $\mathbf{P}(f(\mathbf{X}) \leq m) \geq \frac{1}{2}$, we obtain (46); and if we let $a = m - t$ then since $\mathbf{P}(f(\mathbf{X}) \geq m) \geq \frac{1}{2}$, we obtain (47). □

Now let us consider a related problem concerning common subsequences of two sequences. Given two sequences $\mathbf{x} = (x_1, \ldots, x_{n_1})$ and $\mathbf{y} = (y_1, \ldots, y_{n_2})$, let $com(\mathbf{x}, \mathbf{y})$ denote the maximum length of a common subsequence of $\mathbf{x}$ and $\mathbf{y}$. Let $\mathbf{X} = (X_1, \ldots, X_{n_1})$ and $\mathbf{Y} = (Y_1, \ldots, Y_{n_2})$ be two independent families of independent random variables. We are interested in the concentration of the random variable $com(\mathbf{X}, \mathbf{Y})$. Let μ be the mean of $com(\mathbf{X}, \mathbf{Y})$.

As for the longest increasing subsequence problem, it follows directly from the independent bounded differences inequality, Theorem 3.1, that, for any $t \geq 0$,

$$\mathbf{P}(|com(\mathbf{X}, \mathbf{Y}) - \mu| \geq t) \leq 2e^{-2t^2/(n_1+n_2)}. \tag{48}$$

This shows that, when say $n_1 = n_2 = n$ and n is large, with high probability $com(\mathbf{X}, \mathbf{Y})$ is confined within an interval of length $O(n^{\frac{1}{2}})$. Using the above result on c-configuration functions we may obtain a similar result. For, if we regard $com(\mathbf{x}, \mathbf{y})$ as a function of $(n_1 + n_2)$ variables in the natural way, then it is a 2-configuration function. So, if we let m be a median for $com(\mathbf{X}, \mathbf{Y})$, we obtain

Theorem 4.4 *For any $t \geq 0$,*

$$\mathbf{P}(com(\mathbf{X}, \mathbf{Y}) \geq m + t) \leq 2e^{-t^2/8(m+t)} \tag{49}$$

and

$$\mathbf{P}(com(\mathbf{X}, \mathbf{Y}) \leq m - t) \leq 2e^{-t^2/8m}. \tag{50}$$

Consider the case when $n_1 = n_2 = n$ and n is large, and when the random variables X_i all have the same (fixed) discrete distribution F. It is easy to see (using superadditivity) that there is a constant $\delta_F > 0$ (depending on the distribution F) such that

$$\mathbf{E}(com((X_1, \ldots, X_n), (Y_1, \ldots, Y_n))/n \to \delta_F,$$

and the corresponding result holds for the median. But if say F is the uniform distribution on the set $\{1, \ldots, N\}$ where N is large, then the constant δ_F will be very small, and then the theorem above improves on (48).

4.2.2 Two geometric applications

We now consider applications to the lengths of travelling salesman tours and Steiner trees in the unit square. We shall use the following general result, which is derived from Talagrand's inequality, Theorem 4.1, and which is similar to Theorem 4.3.

Theorem 4.5 *Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a family of independent random variables with X_k taking values in a set Ω_k, and let $\Omega = \prod \Omega_k$. Let the real-valued function f on Ω satisfy the condition that, for each $\mathbf{x} \in \Omega$, there exists a non-negative unit n-vector α such that*

$$f(\mathbf{x}) \leq f(\mathbf{y}) + c\, d_\alpha(\mathbf{x}, \mathbf{y}) \quad \textit{for each } \mathbf{y} \in \Omega. \tag{51}$$

Then

$$\mathbf{P}(|f(\mathbf{X}) - m| \geq t) \leq 4e^{-t^2/4c^2},$$

where m is a median of $f(\mathbf{X})$. The same conclusion holds if the condition (51) is replaced by

$$f(\mathbf{y}) \leq f(\mathbf{x}) + c\, d_\alpha(\mathbf{x}, \mathbf{y}) \quad \textit{for each } \mathbf{y} \in \Omega. \tag{52}$$

Part of the power of this result arises from the asymmetry, that we do not require that *both* conditions (51) and (52) hold – either one will do. Observe that if both hold then we have a bound on $|f(\mathbf{x}) - f(\mathbf{y})|$, and thus on the sum of squared ranges R^2 when the random variables X_k are independent.

Proof For each real number a, let $A_a = \{\mathbf{y} \in \Omega : f(\mathbf{y}) \leq a\}$. Consider any point $\mathbf{x} \in \Omega$. There is a non-negative unit n-vector α such that for each $\mathbf{y} \in \Omega$

$$f(\mathbf{x}) \leq f(\mathbf{y}) + cd_\alpha(\mathbf{x}, \mathbf{y}),$$

and so

$$f(\mathbf{x}) \leq a + cd_\alpha(\mathbf{x}, \mathbf{y})$$

for each $\mathbf{y} \in A_a$. By minimising over such $\mathbf{y}$ we see that

$$f(\mathbf{x}) \leq a + cd_\alpha(\mathbf{x}, A_a) \leq a + cd_T(\mathbf{x}, A_a).$$

Thus if $f(\mathbf{x}) \geq a + t$ then $d_T(\mathbf{x}, A_a) \geq t/c$. Hence

$$\begin{aligned} \mathbf{P}(f(\mathbf{X}) \leq a)\mathbf{P}(f(\mathbf{X}) \geq a + t) &\leq \mathbf{P}(\mathbf{X} \in A_a)\mathbf{P}(d_T(\mathbf{X}, A_a) \geq t/c) \\ &\leq e^{-t^2/4c^2}, \end{aligned}$$

by Talagrand's inequality, Theorem 4.1. If we let $a = m$ we obtain

$$\mathbf{P}(f(\mathbf{X}) \geq m + t) \leq 2e^{-t^2/4c^2},$$

and similarly if we let $a = m - t$ we obtain

$$\mathbf{P}(f(\mathbf{X}) \leq m - t) \leq 2e^{-t^2/4c^2},$$

which completes the proof for the case when condition (51) holds.

Suppose now that condition (52) holds (but not necessarily condition (51)). Let $g(\mathbf{x}) = -f(\mathbf{x})$. Then g satisfies condition (51), and $(-m)$ is a median of $g(\mathbf{X})$, and so by the above

$$\mathbf{P}(|f(\mathbf{X}) - m| \geq t) = \mathbf{P}(|g(\mathbf{X}) - (-m)| \geq t) \leq 4e^{-t^2/4c^2},$$

as required. □

Before we consider the geometric applications, let us check that indeed, as we mentioned earlier, it does not much matter that Theorems 4.3 and 4.5 concern concentration around the median m rather than the mean μ, since the concentration inequalities themselves imply that $|\mu - m|$ is small.

Lemma 4.6 *Let the random variable Y have mean μ and median m, and let $a, b > 0$.*
(a) If $\mathbf{P}(Y - m \geq t) \leq ae^{-t^2/b}$ for any $t > 0$, then $\mu - m \leq (\sqrt{\pi}/2)a\sqrt{b}$; and so if also $\mathbf{P}(Y - m \leq -t) \leq ae^{-t^2/b}$ for any $t > 0$, then $|\mu - m| \leq (\sqrt{\pi}/2)a\sqrt{b}$.
(b) If $\mathbf{P}(Y - m \geq t) \leq ae^{-t^2/b(m+t)}$ for any $t > 0$, then $\mu - m \leq \sqrt{\pi/2}a\sqrt{bm} + 2abe^{-m/2b}$ (which is $O(\sqrt{m})$ as $m \to \infty$, assuming that a and b are constants).

Proof We have

$$\mu - m = \mathbf{E}(Y - m) \leq \mathbf{E}((Y - m)^+) = \int_0^\infty \mathbf{P}(Y - m > t)\, dt. \tag{53}$$

In case (a)

$$\int_0^\infty \mathbf{P}(Y - m > t)\, dt \leq a \int_0^\infty e^{-t^2/b} dt = (\sqrt{\pi}/2)a\sqrt{b},$$

and so the first part of (a) follows from (53). For the second part, note that $(-m)$ is a median for $(-Y)$ and $\mathbf{P}((-Y) - (-m) \geq t) = \mathbf{P}(Y - m \leq -t)$. So if $\mathbf{P}(Y - m \leq -t) \leq ae^{-t^2/b}$ for any $t > 0$, then by what we have just proved

$$m - \mu = \mathbf{E}(-Y) - (-m) \leq (\sqrt{\pi}/2)a\sqrt{b}.$$

In case (b), we again use (53). Now we have

$$\begin{aligned} \int_0^\infty \mathbf{P}(Y - m > t)\, dt &\leq \int_0^\infty ae^{-t^2/b(m+t)} dt \\ &\leq a \int_0^m e^{-t^2/2bm} dt + a \int_m^\infty e^{-t/2b} dt \\ &\leq \sqrt{\pi/2}a\sqrt{bm} + 2abe^{-m/2b}. \end{aligned}$$

□

We shall consider a family $\mathbf{X} = (X_1, \ldots, X_n)$ of independent random variables where each X_k takes values in the unit square $[0,1]^2$. Thus here $\Omega = ([0,1]^2)^n$.

Travelling salesman tours

Given a point $\mathbf{x} \in \Omega$, let $tsp(\mathbf{x})$ be the minimum length of a travelling salesman tour through these points. Much effort has been devoted to investigating the random variable $tsp(\mathbf{X})$, and to investigating its concentration in particular – see for example [56]. Talagrand's inequality effortlessly yields results which previously took great ingenuity.

We need to know one deterministic result, namely that there is a constant c such that the following holds. For every n and every $\mathbf{x} \in \Omega$, there is a tour $T(\mathbf{x})$ through the points in $\mathbf{x}$ such that the sum of the squares of the lengths of the edges in this tour is at most c. This may be proved for example by considering 'space-filling curves' – see [53, 63]. We shall use $T(\mathbf{x})$ to define an appropriate vector α, where the co-ordinate α_k corresponds to the 'awkwardness' of the point $\mathbf{x}_k$.

Given $\mathbf{x} \in \Omega$, we let β_k be the sum of the lengths of the two edges incident to the point x_k in the tour $T(\mathbf{x})$. Thus $\sum \beta_k^2 \leq 4c$ (using the fact that $(a+b)^2 \leq 2a^2 + 2b^2$). We shall see that for any $\mathbf{y} \in \Omega$,

$$tsp(\mathbf{x}) \leq tsp(\mathbf{y}) + d_\beta(\mathbf{x}, \mathbf{y}) \leq tsp(\mathbf{y}) + (2\sqrt{c}) d_\alpha(\mathbf{x}, \mathbf{y}), \tag{54}$$

where α is the unit vector $\beta / \|\beta\|$. Thus the function $tsp(\mathbf{x})$ satisfies the condition (51) in theorem 4.5 (with the value 'c' there being $2\sqrt{c}$). Hence, for any $t \geq 0$,

$$\mathbf{P}(|tsp(\mathbf{X}) - m| \geq t) \leq 4e^{-t^2/16c}, \tag{55}$$

where m is median for $tsp(\mathbf{X})$. A result of this form was first proved by Rhee and Talagrand [56], by a much more involved argument based on the martingale approach.

It remains then to prove (54). Let x, y denote the sets of points corresponding to $\mathbf{x}$, $\mathbf{y}$ respectively. If $x \cap y = \emptyset$ then $d_\alpha(\mathbf{x}, \mathbf{y})$ is twice the length of the tour $T(\mathbf{x})$, and so certainly the inequality (54) holds. Suppose then that $x \cap y \neq \emptyset$. We pick a multiset F of edges between the points of x as follows. For each segment in the tour $T(\mathbf{x})$ of the form $a, v_1, \ldots, v_j, b$ where $a, b \in x \cap y$ and $v_1, \ldots, v_j \in x \setminus y$ (note that $a = b$ if $|x \cap y| = 1$), we put into F each of the edges $v_i v_{i+1}$ doubled for $i = 1, \ldots, j-1$, and the shorter of the edges av_1 and bv_j, also doubled. Thus corresponding to each such segment we obtain a cycle, containing exactly one point in y, and with the sum of the lengths of the edges in it at most the sum of the co-ordinates of β corresponding to the points v_i. These cycles between them cover all the points in $x \setminus y$, and the sum of the lengths of all the edges in F is at most $d_\beta(\mathbf{x}, \mathbf{y})$.

Now let $T^*(\mathbf{y})$ be an optimal tour for $\mathbf{y}$. Consider the (multi)graph G with vertex set $x \cup y$ and with edge set consisting of the edges in $T^*(\mathbf{y})$ together with the edges in F. The graph G is connected and each vertex degree is even, and so G has an Eulerian tour. This tour can be shortcut to give a travelling salesman tour, which by the triangle inequality has length no more than the sum of the lengths of the edges in G, and this sum is at most $tsp(\mathbf{y}) + d_\beta(\mathbf{x}, \mathbf{y})$. This completes the proof of (54), as required.

Steiner trees

A *Steiner tree* for a set x of points in the unit square is a tree with vertex set some set of points in the plane containing x. Given $\mathbf{x} \in \Omega$, we let $st(\mathbf{x})$ denote the minimal length of a Steiner tree for the corresponding set x. We may use the tour $T(\mathbf{x})$ exactly as above to define a corresponding vector β.

Now let $\mathbf{y} \in \Omega$, and let $S^*(\mathbf{y})$ be an optimal Steiner tree for the corresponding set of points $\mathbf{y}$. Consider the set E of edges consisting of the edges in $S^*(\mathbf{y})$ together with those edges in $T(\mathbf{x})$ with at least one end in $x \setminus y$. The total length of these edges is at most $st(\mathbf{y}) + d_\beta(\mathbf{x}, \mathbf{y})$, and we have already seen that $\sum \beta_k^2 \leq 4c$. The key observation is that the graph G on $x \cup y$ with edge set E is connected; for, since $T(\mathbf{x})$ is connected each point in x is in the same component as some point in y, and since $S^*(\mathbf{y})$ is connected each point in y is in the same component. It follows that $st(\mathbf{x})$ is at most the sum of the lengths of the edges in E, and thus $st(\mathbf{x}) \leq st(\mathbf{y}) + d_\beta(\mathbf{x}, \mathbf{y})$. Hence by Theorem 4.5, for $t \geq 0$

$$\mathbf{P}(|st(\mathbf{X}) - m| \geq t) \leq 4e^{-t^2/16c}, \tag{56}$$

where m is a median for $st(\mathbf{X})$.

4.2.3 Random minimum spanning trees

Consider the complete graph K_n with random independent edge lengths X_e, each uniformly distributed on $(0, 1)$. Let L_n be the corresponding random length of a minimum spanning tree. It is known ([23]) that the expected value of L_n tends to $\zeta(3)$ as $n \to \infty$, where

$$\zeta(3) = \sum_{j=1}^{\infty} j^{-3} \sim 1.202.$$

It is shown in [24] that L_n is quite concentrated around $\zeta(3)$, using the method of bounded differences; and this result is improved in [8] using Talagrand's method. (Also, it is shown in [30] that $\sqrt{n}(L_n - \zeta(3))$ is asymptotically normally distributed.)

Both the bounded differences method and Talagrand's method can in fact be used to prove that L_n is very highly concentrated around the value $\zeta(3)$ – see [48], but the latter method is easier and will be described below. (In fact the bounded differences approach seems to yield a slightly stronger result.) Both approaches depend on the fact that long edges are not important. For $0 \leq b \leq 1$, let $L_n^{(b)}$ be the minimum length of a spanning tree when the edge lengths X_e are replaced by $\min(X_e, b)$. For simplicity we consider here the case of a fixed deviation $t > 0$. We need the following lemma.

Lemma 4.7 *[48] For any $t > 0$ there exist constants $c_1 > 0$ and $\nu > 0$ such that if we let $b = c_1/n$ then*

$$\mathbf{P}(L_n - L_n^{(b)} \geq t) \leq e^{-\nu n} \quad \textit{for all } n.$$

We shall prove the following concentration result for the minimum spanning tree length L_n.

Theorem 4.8 *For any $t > 0$ there exists $\delta > 0$ such that*

$$P(|L_n - \zeta(3)| \geq t) \leq e^{-\delta n} \quad \textit{for all } n.$$

It is easy to see that the bound above is of the right order. For example, for each $n \geq 5$ the probability that $L_n \geq 2$ is at least the probability that each edge incident with the first four vertices has length at least $1/2$, and this probability is at least $(1/16)^n$.

Proof Let $N = \binom{n}{2}$, and let $\mathbf{Y} = (Y_1, \ldots, Y_N)$ be a family of independent random variables with each Y_i uniformly distributed on $(0,1)$, corresponding to the edge lengths in the graph K_n. We may write the random variable L_n as $mst(\mathbf{Y})$.

Let $0 < b \leq 1$, and let $\Omega = (0,b)^N$. For each $i = 1, \ldots, N$ let $X_i = \min(Y_i, b)$. Then $\mathbf{X} = (X_1, \ldots, X_N)$ is a family of independent random variables each taking values in $(0,b)$, and $L_n^{(b)} = mst(\mathbf{X})$.

Now consider the random variable $mst(\mathbf{X})$. Let $\Omega = (0,b)^N$ and let $\mathbf{x} \in \Omega$. Denote the set of edges in a corresponding minimum spanning tree by $T = T(\mathbf{x})$. Let $\beta = \beta(\mathbf{x})$ be the N-vector with $\beta_i = b$ for $i \in T$ and $\beta_i = 0$ otherwise, and let $\alpha = \alpha(\mathbf{x})$ be the unit vector $\beta/(b\sqrt{n-1})$. Then for any $\mathbf{y} \in \Omega$,

$$\begin{aligned} mst(\mathbf{y}) &\leq \sum_{i \in T} y_i \\ &\leq \sum_{i \in T} x_i + \sum_{i \in T} (y_i - x_i)^+ \\ &\leq mst(\mathbf{x}) + d_\beta(\mathbf{x}, \mathbf{y}) \\ &\leq mst(\mathbf{x}) + b\sqrt{n}\; d_\alpha(\mathbf{x}, \mathbf{y}). \end{aligned}$$

Thus the function $mst(\mathbf{x})$ satisfies condition (52) in Theorem 4.5 with $c = b\sqrt{n}$, and so for any $t \geq 0$

$$\mathbf{P}(|mst(\mathbf{X}) - m| \geq t) \leq 4e^{-t^2/4b^2n},$$

where m is a median for $mst(\mathbf{X})$. We may use Lemma 4.7, together with this last inequality with $b = c_1/n$, to obtain

$$\begin{aligned} \mathbf{P}(|mst(\mathbf{Y}) - m| \geq 2t) &\leq \mathbf{P}(mst(\mathbf{Y}) - mst(\mathbf{X}) \geq t) + \mathbf{P}(|mst(\mathbf{X}) - m| \geq t) \\ &\leq e^{-\nu n} + 4e^{-t^2 n/4c_1^2}. \end{aligned}$$

It follows that for any $t > 0$ there exists $\delta_1 = \delta_1(t) > 0$ such that

$$\mathbf{P}(|L_n - m| \geq t) \leq e^{-\delta_1 n} \quad \text{for all } n.$$

It remains to tidy up, by replacing the m here by $\zeta(3)$ (in the spirit of Lemma 4.6). By the above

$$|\mathbf{E}(L_n) - m| \leq \mathbf{E}(|L_n - m|) \leq \frac{t}{4} + n\mathbf{P}(|L_n - m| > t/4) \leq t/3$$

for n sufficiently large. Also we saw earlier that for n sufficiently large, $|\mathbf{E}(L_n) - \zeta(3)| \leq t/3$, and so $|m - \zeta(3)| \leq 2t/3$ for n sufficiently large. Hence for n sufficiently large

$$\mathbf{P}(|L_n - \zeta(3)| \geq t) \leq \mathbf{P}(|L_n - m| \geq t/3) \leq e^{-\delta_1 n}$$

where $\delta_1 = \delta_1(t/3)$, and the theorem follows. □

4.3 Proof of Talagrand's inequality

In this subsection we shall prove an extended form of theorem 4.1.

Theorem 4.9 *Let* $\mathbf{X} = (X_1, \ldots, X_n)$ *be a family of independent random variables where* X_k *takes values in a set* Ω_k*, and let* A *be a subset of the product space* $\Omega = \prod \Omega_k$*. Then*

$$\mathbf{P}(\mathbf{X} \in A)\, \mathbf{E}\left(e^{\frac{1}{4} d_T(\mathbf{X},A)^2}\right) \leq 1, \tag{57}$$

and so, for any $t \geq 0$*,*

$$\mathbf{P}(\mathbf{X} \in A)\, \mathbf{P}(d_T(\mathbf{X}, A) \geq t) \leq e^{-t^2/4}. \tag{58}$$

The latter inequality (58) (which is Theorem 4.1) follows immediately from the former (57) by Markov's inequality. The scheme of the proof of (57) is as follows. We first develop an equivalent definition of Talagrand's distance d_T. Then after two technical lemmas we start the main proof by induction on n. We prove a claim relating the distance $d_T(\mathbf{x}, A)$ in dimension $n+1$ to certain distances involving only the first n co-ordinates. This claim involves a parameter λ. The induction hypothesis yields bounds for the distances in dimension n. We then optimise over λ and average over the last co-ordinate. The whole proof is neither long nor hard, but it is one of those proofs by induction from which it is not easy to get a good feel about why the result really is true. For a brief discussion of an alternative approach based on ideas from information theory see the next (final) subsection.

In order to prove (58) we first develop the alternative characterisation of Talagrand's convex distance $d_T(\mathbf{x}, A)$. Fix a point $\mathbf{x}$ and a set A in R^n. Let $U = U(\mathbf{x}, A)$ be the set of all binary vectors $\mathbf{u}$ such that starting from $\mathbf{x}$ we may reach a vector $\mathbf{y} \in A$ by changing only co-ordinates x_i such that $u_i = 1$ (and not necessarily changing all of them). Thus $\mathbf{0} \in U$ if and only if $\mathbf{x} \in A$. Further let $V = V(\mathbf{x}, A)$ be the convex hull of the set U. The following lemma explains the term 'convex distance'.

Lemma 4.10

$$d_T(\mathbf{x}, A) = \min\{\|\mathbf{v}\| \colon \mathbf{v} \in \mathbf{V}\}. \tag{59}$$

Proof If $\mathbf{x} \in A$ then both sides above equal 0. So we may assume that $\mathbf{x} \notin A$, and then both sides are positive. Denote the right hand side above by ρ. Let $\alpha = (\alpha_1, \ldots, \alpha_n) \geq \mathbf{0}$ be a unit vector. We write $\alpha.\mathbf{u}$ to denote the inner product $\sum \alpha_k u_k$. Then

$$d_\alpha(\mathbf{x}, A) = \min_{\mathbf{y} \in A} d_\alpha(\mathbf{y}, A) = \min_{\mathbf{u} \in U} \alpha.\mathbf{u} = \min_{\mathbf{v} \in V} \alpha.\mathbf{v}, \tag{60}$$

since the minimum of a linear functional over the convex hull V of the finite set U must be achieved at a point of U. But by the Cauchy-Schwarz inequality,

$$\alpha.\mathbf{v} \leq \|\alpha\|\, \|\mathbf{v}\| = \|\mathbf{v}\| .$$

Thus $d_\alpha(\mathbf{x}, A) \leq \rho$, and since this holds for every choice of α we deduce that $d_T(\mathbf{x}, A) \leq \rho$.

For the converse result, note that the minimum in (59) is achieved, that is there is a point $\tilde{\mathbf{v}} \in V$ with norm equal to ρ, since V is compact. Let α be the unit vector

$\tilde{\mathbf{v}}/\rho$. Consider any point $\mathbf{v} \in V$. Since V is convex, the point $\tilde{\mathbf{v}} + \lambda(\mathbf{v} - \tilde{\mathbf{v}})$ is in V for each $0 \leq \lambda \leq 1$; and so

$$(\tilde{\mathbf{v}} + \lambda(\mathbf{v} - \tilde{\mathbf{v}})).(\tilde{\mathbf{v}} + \lambda(\mathbf{v} - \tilde{\mathbf{v}})) \geq \tilde{\mathbf{v}}.\tilde{\mathbf{v}}.$$

This yields

$$2\lambda\, \tilde{\mathbf{v}}.(\mathbf{v} - \tilde{\mathbf{v}}) + \lambda^2\, (\mathbf{v} - \tilde{\mathbf{v}}).(\mathbf{v} - \tilde{\mathbf{v}}) \geq 0,$$

and by considering small λ we see that $\tilde{\mathbf{v}}.(\mathbf{v} - \tilde{\mathbf{v}}) \geq 0$. Thus $\alpha.\mathbf{v} \geq \alpha.\tilde{\mathbf{v}} = \rho$ for all $\mathbf{v} \in V$. Hence by (60),

$$d_T(\mathbf{x}, A) \geq d_\alpha(\mathbf{x}, A) = \min_{\mathbf{v} \in V} \alpha.\mathbf{v} = \rho,$$

and we are done. □

We need two further lemmas before we start the main proof of Talagrand's inequality. The first is from [31, 68].

Lemma 4.11 *For all* $0 < r \leq 1$,

$$\inf_{0 \leq \lambda \leq 1} r^{-\lambda} e^{\frac{1}{4}(1-\lambda)^2} \leq 2 - r.$$

Proof For the case $0 \leq r \leq e^{-\frac{1}{2}}$ we may consider $\lambda = 0$ and check that $e^{\frac{1}{4}} \leq 2 - e^{-\frac{1}{2}}$. So suppose that $e^{-\frac{1}{2}} \leq r \leq 1$. Let $\lambda = 1 + 2\ln r$ (so $0 \leq \lambda \leq 1$). We want to show that $f(r) \geq 0$, where $f(r)$ is the logarithm of the ratio of the right side of the inequality to the left side. Now

$$f(r) = \ln(2 - r) + \lambda \ln r - (1 - \lambda)^2/4 = \ln(2 - r) + \ln r + (\ln r)^2.$$

Since $f(1) = 0$, it suffices to show that $g(r) = rf'(r) \leq 0$. Note that

$$g(r) = r\left(-\frac{1}{2 - r} + \frac{1}{r} + \frac{2\ln r}{r}\right) = -\frac{r}{2 - r} + 1 + 2\ln r.$$

Since $g(1) = 0$, it suffices now to show that $g'(r) \geq 0$. But $g'(r) = 2\left(\frac{1}{r} - \frac{1}{(2-r)^2}\right)$, and $\frac{1}{r} \geq 1 \geq \frac{1}{(2-r)^2}$: thus indeed $g'(r) \geq 0$, which completes the proof. □

The last preliminary result we need is a form of Holder's inequality (see for example [20] page 465) which we state and prove here for completeness, in a form useful for us.

Lemma 4.12 *For any (appropriately integrable) functions* f *and* g, *and any* $0 \leq t \leq 1$,

$$\mathbf{E}\left(e^{tf(\mathbf{X})} e^{(1-t)g(\mathbf{X})}\right) \leq \left(\mathbf{E}(e^{f(\mathbf{X})})\right)^t \left(\mathbf{E}(e^{g(\mathbf{X})})\right)^{1-t}.$$

Proof Let $a, b > 0$, and for $0 < t < 1$ let $h(t) = a^t b^{1-t}$. Then $h'(t) = h(t)(\ln(a/b))^2 \geq 0$, so h is convex, and thus $a^t b^{1-t} \leq ta + (1 - t)b$. Now let $F = \mathbf{E}(e^{f(\mathbf{X})})$ and $G = \mathbf{E}(e^{g(\mathbf{X})})$. Then

$$(e^{f(\mathbf{x})}/F)^t (e^{g(\mathbf{x})}/G)^{1-t} \leq (t/F)e^{f(\mathbf{x})} + ((1 - t)/G)e^{g(\mathbf{x})}.$$

Taking expected values,

$$\begin{aligned}\mathbf{E}\left(e^{tf(\mathbf{X})}e^{(1-t)g(\mathbf{X})}\right)/(F^tG^{1-t}) &= \mathbf{E}\left((e^{f(\mathbf{X})}/F)^t(e^{g(\mathbf{X})}/G)^{1-t}\right)\\ &\leq (t/F)\mathbf{E}(e^{f(\mathbf{X})}) + ((1-t)/G)\mathbf{E}(e^{g(\mathbf{X})})\\ &= t+(1-t) \;=\; 1,\end{aligned}$$

which yields the required inequality. □

We may now start the main proof of the inequality (57). Let us write $\nu_n(A)$ for $\mathbf{P}(\mathbf{X} \in A)$. We use induction on n. Consider first the case $n = 1$. Now $d_T(\mathbf{x}, A)$ equals 0 if $\mathbf{x} \in A$ and otherwise equals 1. So

$$\mathbf{E}\left(e^{\frac{1}{4}d_T(\mathbf{X},A)^2}\right) = \nu_1(A) + e^{\frac{1}{4}}(1-\nu_1(A)).$$

But for $0 \leq p \leq 1$,

$$p(p + e^{\frac{1}{4}}(1-p)) \leq p(p+2(1-p)) = p(2-p) \leq 1,$$

which completes the proof of the case $n = 1$.

Now let $n \geq 1$, suppose that the inequality (57) holds for n, and consider the case $n+1$. Denote $\prod_{k=1}^{n}\Omega_k$ by $\Omega^{(n)}$. Write $\prod_{k=1}^{n+1}\Omega_k$ as $\Omega^{(n+1)} = \Omega^{(n)} \times \Omega_{n+1}$, with typical element written as $\mathbf{z} = (\mathbf{x}, \omega)$, where $\mathbf{x} \in \Omega^{(n)}$ and $\omega \in \Omega_{n+1}$. Let $A \subseteq \Omega^{(n+1)}$. For $\omega \in \Omega_{n+1}$, the *ω-section* A_ω of A is defined by

$$A_\omega = \{\mathbf{x} \in \Omega^{(n)} : (\mathbf{x}, \omega) \in A\}.$$

The *projection of* A is the set B defined by

$$B = \cup_\omega A_\omega = \{\mathbf{x} \in \Omega^{(n)} : (\mathbf{x}, \omega) \in A \text{ for some } \omega \in \Omega_{n+1}\}.$$

We next prove an inequality relating $d_T(\mathbf{z}, A)$ to corresponding distances between $\mathbf{x}$ and the ω-section and projection of A. The inequality involves a parameter λ which we shall later choose appropriately.

Claim Let $\mathbf{z} = (\mathbf{x}, \omega) \in \Omega^{(n)} \times \Omega_{n+1}$ and let $0 \leq \lambda \leq 1$. Then

$$d_T(\mathbf{z}, A)^2 \leq \lambda d_T(\mathbf{x}, A_\omega)^2 + (1-\lambda)d_T(\mathbf{x}, B)^2 + (1-\lambda)^2. \tag{61}$$

Proof of Claim By Lemma 4.10 above, there is a vector $\mathbf{v}_1 \in V(\mathbf{x}, A_\omega)$ with norm equal to $d_T(\mathbf{x}, A_\omega)$, and a vector $\mathbf{v}_2 \in V(\mathbf{x}, B)$ with norm equal to $d_T(\mathbf{x}, B)$. Now if $\mathbf{u} \in U(\mathbf{x}, A_\omega)$ then $(\mathbf{u}, 0) \in U(\mathbf{z}, A)$, and so if $\mathbf{v} \in V(\mathbf{x}, A_\omega)$ then $(\mathbf{v}, 0) \in V(\mathbf{z}, A)$. Similarly, if $\mathbf{u} \in U(\mathbf{x}, B)$ then $(\mathbf{u}, 1) \in U(\mathbf{z}, A)$, and so if $\mathbf{v} \in V(\mathbf{x}, B)$ then $(\mathbf{v}, 1) \in V(\mathbf{z}, A)$. Hence both $(\mathbf{v}_1, 0)$ and $(\mathbf{v}_2, 1)$ are in the convex set $V(\mathbf{z}, A)$, and so if we let

$$\mathbf{v}_3 = \lambda(\mathbf{v}_1, 0) + (1-\lambda)(\mathbf{v}_2, 1) = (\lambda\mathbf{v}_1 + (1-\lambda)\mathbf{v}_2, 1-\lambda),$$

then $\mathbf{v}_3 \in V(\mathbf{z}, A)$. By Lemma 4.10 again, $d_T(\mathbf{z}, A)$ is at most the norm of $\mathbf{v}_3$. Now the function $f(t) = t^2$ is convex, and so

$$(\lambda a + (1-\lambda)b)^2 \leq \lambda a^2 + (1-\lambda)b^2.$$

Hence

$$\begin{aligned} \|\mathbf{v}\|^2 &= \|(\lambda \mathbf{v}_1 + (1-\lambda)\mathbf{v}_2)\|^2 + (1-\lambda)^2 \\ &\leq \lambda \|\mathbf{v}_1\|^2 + (1-\lambda) \|\mathbf{v}_2\|^2 + (1-\lambda)^2 \\ &= \lambda d_T(\mathbf{x}, A_\omega)^2 + (1-\lambda) d_T(\mathbf{x}, B)^2 + (1-\lambda)^2. \end{aligned}$$

This completes the proof of the claim.

We are now ready to tackle the induction step. For each fixed ω, let $\mathbf{E}(\omega)$ denote

$$\mathbf{E}\left(e^{\frac{1}{4}d_T((\mathbf{X},\omega),A)^2}\right) = \mathbf{E}\left(e^{\frac{1}{4}d_T((\mathbf{X},X_{n+1}),A)^2} \mid X_{n+1} = \omega\right).$$

We shall first give an upper for $\mathbf{E}(\omega)$, and then average over ω. Fix ω, and note that the claim gives

$$e^{\frac{1}{4}d_T((\mathbf{X},\omega),A)^2} \leq e^{\frac{1}{4}(1-\lambda)^2} e^{\lambda(\frac{1}{4}d_T(\mathbf{X},A_\omega)^2)} e^{(1-\lambda)(\frac{1}{4}d_T(\mathbf{X},B)^2)}.$$

Hence by Lemma 4.12 (Holder's inequality), we obtain

$$\mathbf{E}(\omega) \leq e^{\frac{1}{4}(1-\lambda)^2} \mathbf{E}\left(e^{\frac{1}{4}d_T(\mathbf{X},A_\omega)^2}\right)^\lambda \mathbf{E}\left(e^{\frac{1}{4}d_T(\mathbf{X},B)^2}\right)^{1-\lambda}.$$

By the induction hypothesis applied to the two expectations above, we find that

$$\begin{aligned} \mathbf{E}(\omega) &\leq e^{\frac{1}{4}(1-\lambda)^2} (\nu_n(A_\omega))^{-\lambda} (\nu_n(B))^{-(1-\lambda)} \\ &= e^{\frac{1}{4}(1-\lambda)^2} (\nu_n(B))^{-1} \left(\frac{\nu_n(A_\omega)}{\nu_n(B)}\right)^{-\lambda}. \end{aligned}$$

Thus for all $0 \leq \lambda \leq 1$,

$$\mathbf{E}(\omega) \leq (\nu_n(B))^{-1} r^{-\lambda} e^{\frac{1}{4}(1-\lambda)^2},$$

where $r = \nu_n(A_\omega)/\nu_n(B)$, and so $0 \leq r \leq 1$. By Lemma 4.11, we find

$$\mathbf{E}(\omega) \leq (\nu_n(B))^{-1}(2 - \nu_n(A_\omega)/\nu_n(B)).$$

Now $\nu_n(A_\omega) = \mathbf{P}((\mathbf{X}, X_{n+1}) \in A \mid X_{n+1} = \omega)$. We can average over the values ω taken by X_{n+1} to obtain

$$\begin{aligned} \nu_{n+1}(A)\, \mathbf{E}\left(e^{\frac{1}{4}d_T((\mathbf{X},X_{n+1}),A)^2}\right) &\leq (\nu_{n+1}(A)/\nu_n(B))(2 - \nu_{n+1}(A)/\nu_n(B)) \\ &= x(2-x) \;\leq\; 1, \end{aligned}$$

where $x = \nu_{n+1}(A)/\nu_n(B)$. We have now completed the proof of the induction step, and thus of the theorem. □

4.4 Ideas from Information Theory

There is a third main approach to proving general concentration results, which uses ideas from information theory. Indeed, the first general concentration result seems to have been proved and used in this context, by Ahlswede, Gács and Körner [1] in 1976. Their concentration result, the 'blowing-up lemma', was sharpened by Csiszár and Körner [17], and then in 1986 Marton [40] gave a simple and elegant proof. This result resembled Theorem 3.5 above, though with a worse constant in the exponent. The optimal constant was obtained in 1996 by Marton [41], using the same elegant information-theoretic approach. Dembo [18] showed that the method is strong enough to recover all of the inequalities of Talagrand in [68] (including Theorem 4.9 above), where it is assumed that the random variables involved are independent. The method is extended in [42] to handle certain cases of weak dependence. For other recent work see [43, 71].

It is not clear if these ideas will lead to further new applications in discrete mathematics and theoretical computer science. However, they are very elegant and powerful, and so we try here to give a flavour of the method. We shall show how they give a very different proof of Theorem 3.5, following [40, 41].

Let $\Omega_1, \ldots, \Omega_n$ be finite sets, and let Ω denote their product $\prod \Omega_k$. Let $\mathbf{p} = (p_\omega : \omega \in \Omega)$ and $\mathbf{q} = (q_\omega : \omega \in \Omega)$ specify probability distributions on Ω. Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a family of random variables, with X_k taking values in Ω_k; and let $\mathbf{Y} = (Y_1, \ldots, Y_n)$ be another such family. We shall be interested in joint distributions for $\mathbf{X}$ and $\mathbf{Y}$ which have marginals $\mathbf{p}$ and $\mathbf{q}$; that is, such that

$$\mathbf{P}(\mathbf{X} = \omega) = \sum_{\omega' \in \Omega} \mathbf{P}((\mathbf{X}, \mathbf{Y}) = (\omega, \omega')) = p_\omega$$

for each $\omega \in \Omega$, and similarly for $\mathbf{Y}$ and $\mathbf{q}$. We shall define a notion of distance between the distributions $\mathbf{p}$ and $\mathbf{q}$ based on the expected Hamming distance between random points $\mathbf{X}$ and $\mathbf{Y}$. Observe that the expected Hamming distance between $\mathbf{X}$ and $\mathbf{Y}$ is given by

$$\mathbf{E}(d_H(\mathbf{X}, \mathbf{Y})) = \sum_k \mathbf{P}(X_k \neq Y_k).$$

We define $d_H(\mathbf{p}, \mathbf{q})$ to be the minimum value of $\mathbf{E}(d_H(\mathbf{X}, \mathbf{Y}))$, over all choices of joint distribution for $\mathbf{X}$ and $\mathbf{Y}$ with marginals $\mathbf{p}$ and $\mathbf{q}$. It turns out that we may obtain concentration results by giving an upper bound on $d_H(\mathbf{p}, \mathbf{q})$ when the distribution $\mathbf{q}$ is a product distribution (that is, corresponds to independent random variables).

For the key lemma, we need one last definition. The *informational divergence* of $\mathbf{p}$ with respect to $\mathbf{q}$ is

$$D(\mathbf{p}||\mathbf{q}) = \sum_{\omega \in \Omega} p_\omega \ln(p_\omega / q_\omega).$$

Lemma 4.13 *If $\mathbf{q}$ is a product distribution, then*

$$d_H(\mathbf{p}, \mathbf{q})^2 \leq (n/2) D(\mathbf{p}||\mathbf{q}).$$

Using this information-theoretic lemma we shall prove the following elegant symmetrical inequality, closely related to Theorem 3.5. Recall that the Hamming distance $d_H(A, B)$ between two subsets A and B of Ω is the minimum value of $d_H(x, y)$ over all choices of $x \in A$ and $y \in B$.

Theorem 4.14 *Let* $\mathbf{q}$ *be a product distribution. Then*

$$d_H(A,B) \le \left(\frac{n}{2}\ln\frac{1}{q(A)}\right)^{\frac{1}{2}} + \left(\frac{n}{2}\ln\frac{1}{q(B)}\right)^{\frac{1}{2}}.$$

Proof Let $\mathbf{p}$ denote the distribution with $p_\omega = q_\omega/q(A)$ for $\omega \in A$ and $p_\omega = 0$ otherwise; and define the distribution $\mathbf{r}$ similarly corresponding to B. Then

$$\begin{aligned} D(\mathbf{p}||\mathbf{q}) &= \sum_{\omega\in\Omega} p_\omega \ln(p_\omega/q_\omega) \\ &= \sum_{\omega\in\Omega} p_\omega \ln(1/q(A)) \\ &\le \ln(1/q(A)). \end{aligned}$$

Similarly, $D(\mathbf{r}||\mathbf{q}) \le \ln(1/q(B))$. Next we use the observation that, since $d_H(\mathbf{p},\mathbf{r})$ is the expected Hamming distance between certain random points in A and in B, it must be at least the minimum value $d_H(A,B)$. Hence, by a triangle inequality and the above lemma,

$$\begin{aligned} d_H(A,B) &\le d_H(\mathbf{p},\mathbf{r}) \\ &\le d_H(\mathbf{p},\mathbf{q}) + d_H(\mathbf{r},\mathbf{q}) \\ &\le \left(\frac{n}{2}\ln\frac{1}{q(A)}\right)^{\frac{1}{2}} + \left(\frac{n}{2}\ln\frac{1}{q(B)}\right)^{\frac{1}{2}}, \end{aligned}$$

as required. □

Finally let us see that Theorem 3.5 follows directly from the last result. Let $t > 0$ and let $B = \Omega \setminus A_t$, the complement of the t-fattening of A – see the comments immediately after Theorem 3.5. We shall take $q(A)$ to be $\mathbf{P}(\mathbf{X} \in A)$, in the notation there. Since $d_H(A,B) \ge t$, by Theorem 4.14 above we have

$$\left(\frac{n}{2}\ln\frac{1}{q(B)}\right)^{\frac{1}{2}} \ge t - t_0,$$

where

$$t_0 = \left(\frac{n}{2}\ln\frac{1}{q(A)}\right)^{\frac{1}{2}}.$$

and so

$$\mathbf{P}(d_H(\mathbf{X},A) \ge t) \;=\; q(B) \;\ge\; 1 - e^{-2(t-t_0)^2/n}.$$

But this is exactly the inequality (22) in the proof of Theorem 3.5, and so the theorem follows.

Acknowledgements I am pleased to acknowledge very helpful comments from the referees.

References

[1] Ahlswede R., Gács P. and Körner J. (1976): Bounds on conditional probabilities with applications in multi-user communication, Z. Wahrscheinlichkeitstheorie verw. Geb. **34**, 157 – 177. (Erratum (1977) **39**, 353 – 354.)

[2] Alon N., Kim J.H. and Spencer J. (1997): Nearly perfect matchings in regular simple hypergraphs, Israel J. Math. **100**, 171 – 188.

[3] Alon N. and Spencer N. (1992): The Probabilistic Method, John Wiley & Sons.

[4] Angluin D. and Valiant L. (1979): Fast probabilistic algorithms for Hamiltonian circuits and matchings, J. Computer and System Sciences **18**, 155 – 193.

[5] Avram F. and Bertsimas D. (1992): The minimum spanning tree constant in geometrical probability and under the independent model: a unified approach, Annals of Applied Probability **2**, 113 - 130.

[6] Azuma K. (1967): Weighted sums of certain dependent random variables, Tôkuku Math. J. **19**, 357 – 367.

[7] Bennett G. (1962): Probability inequalities for the sum of independent random variables, J. Amer. Statist. Assoc. **57**, 33 – 45.

[8] Beveridge A., Frieze A. and McDiarmid C. (1998): Random minimum length spanning trees in regular graphs, Combinatorica, to appear.

[9] Bollobás B. (1985): Random Graphs, Academic Press.

[10] Bollobás B. (1997): Martingales, isoperimetric inequalities and random graphs, Colloq. Math. Soc. János Bolyai **52**, 113-139.

[11] Bollobás B. (1988): The chromatic number of random graphs, Combinatorica **8**, 49 – 55.

[12] Bollobás B. (1990): Sharp concentration of measure phenomena in the theory of random graphs, in Random Graphs '87, (M. Karoński, J. Jaworski and A. Ruciński, editors), John Wiley and Sons, 1 – 15.

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

[14] Chernoff H. (1952): A measure of the asymptotic efficiency for tests of a hypothesis based on the sum of observables, Ann. Math. Statist. **23**, 493 – 509.

[15] Chvátal V. (1979): The tail of the hypergeometric distribution, Discrete Mathematics **25**, 285-287.

[16] Coffman E.G. and Lueker G.S. (1991): Probabilistic Analysis of Packing and Partitioning Algorithms, Wiley, New York.

[17] Csiszár I. and Körner J. (1981): Information Theory: Coding Theorems for Discrete Memoryless Systems, Academic Press, New York.

[18] Dembo A. (1997): Information inequalities and concentration of measure, Ann. Probab. **25**, 927 – 939.

[19] Dembo A. and Zeitouni O. (1993): Large Deviation Techniques, Jones and Bartlett.

[20] Durrett R. (1996): Probability: Theory and Examples, Second edition, Duxbury Press.

[21] Freedman D.A. (1975): On tail probabilities for martingales, Ann. Probab. **3**, 100 – 118.

[22] Feller W.J. (1968): An Introduction to Probability Theory and its Applications, Volume 1, Third Edition, John Wiley & Sons, New York.

[23] Frieze A.M. (1985): On the value of a random minimum spanning tree problem, Discrete Applied Mathematics **10**, 47 - 56.

[24] Frieze A.M. and McDiarmid C.J.H. (1989): On random minimum length spanning trees, Combinatorica **9**, 363 - 374.

[25] Frieze A.M. and McDiarmid C.J.H. (1997): Algorithmic theory of random graphs, Random Structures and Algorithms **10**, 5 – 42.

[26] Grable D.A. (1998): A large deviation inequality for functions of independent multi-way choices, Combinatorics, Probability and Computing **7**, 57 – 63.

[27] Grable D.A. and Panconesi A. (1997): Nearly optimal distributed edge colouring in $O(\log \log n)$ rounds, Random Structures and Algorithms **10**, 385 – 405.

[28] Grimmett G.R. and Stirzaker D.R. (1992): Probability and Random Processes, Second edition, Oxford University Press.

[29] Hoeffding W.J. (1963): Probability inequalities for sums of bounded random variables, J. Amer. Statist. Assoc. **58**, 713-721.

[30] Janson S. (1995): The minimal spanning tree in a complete graph and a functional limit theorem for trees in a random graph, Random Structures and Algorithms **7**, 337 - 355.

[31] Johnson W. and Schechtman G. (1991): Remarks on Talagrand's deviation inequality for Rademacher's functions, Lecture Notes in Mathematics **1470**, Springer-Verlag, 72 – 77.

[32] Kahn J. (1996): Asymptotically good list colorings, J. Combinatorial Theory A **73**, 1 –59.

[33] Kamath A., Motwani R., Palem K. and Spirakis P. (1995): Tail bounds for occupancy and the satisfiability threshold conjecture, Random Structures and Algorithms **7**, 59 – 80.

[34] Kim J.H. (1995): On Brooks' theorem for sparse graphs, Combinatorics, Probability and Computing **4**, 97 – 132.

[35] Kim J.H. (1995): The Ramsey number $R(3,t)$ has order of magnitude $t^2/\log t$, Random Structures and Algorithms **7**, 173 – 207.

[36] Knuth D.E. (1973): The Art of Computer Programming Volume 3: Sorting and Searching, Addison-Wesley.

[37] Leader I. (1991): Discrete isoperimetric inequalities, Proc. Sympos. Appl. Math. **44**, 57 – 80.

[38] Ledoux M. and Talagrand M. (1991): Probability in Banach Spaces, Springer-Verlag.

[39] Lipster R. Sh. and Shiryayev A.N. (1989): Theory of Martingales Kluwer, Dordrecht.

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

[41] Marton K. (1996): Bounding $\bar{d}$-distance by informational divergence: a method to prove measure concentration, Ann. Probab. **24**, 857 – 866.

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

[43] Marton K. and Shields P.C. (1994): The positive divergence and blowing-up properties, Israel J. Math. **86**, 331 – 348.

[44] Maurey B. (1979): Construction de suites symétriques, Compt. Rend. Acad. Sci. Paris **288**, 679 – 681.

[45] McDiarmid C. (1989): On the method of bounded differences, in Surveys in Combinatorics, ed J. Siemons, London Mathematical Society Lecture Note Series 141, Cambridge University Press.

[46] McDiarmid C. (1990): On the chromatic number of random graphs, Random Structures and Algorithms **1**, 435 – 442.

[47] McDiarmid C. (1997): Centering sequences with bounded differences, Combinatorics, Probability and Computing **6**, 79 – 86.

[48] McDiarmid C. (1998): Concentration for minimum spanning tree lengths, manuscript.

[49] McDiarmid C. and Hayward R. (1996): Large deviations for quicksort, J. Algorithms **21**, 476 – 507.

[50] Milman V. and Schechtman G. (1986): Asymptotic theory of finite dimensional normed spaces, Lecture Notes in Math. 1200, Springer-Verlag.

[51] Motwani R. and Raghavan P. (1995): Randomized Algorithms, Cambridge University Press.

[52] Penrose M. (1998): Random minimum spanning tree and percolation on the n-cube, Random Structures and Algorithms **12**, 369 – 382.

[53] Platzman L.K. and Bartholdi J.J. (1989): Spacefilling curves and the planar traveling salesman problem, J. Assoc. Comput. Mach. **36**, 719 – 737.

[54] Rhee W.T. and Talagrand M. (1987): Martingale inequalities and NP-complete problems, Math. Oper. Res. **12**, 177 – 181.

[55] Rhee W.T. and Talagrand M. (1989): Martingale inequalities, interpolation and NP-complete problems, Math. Oper. Res. **14**, 189 – 202.

[56] Rhee W.T. and Talagrand M. (1989): A sharp deviation for the stochastic traveling salesman problem, Ann. Probab. **17**, 1 – 8.

[57] Ross S.M. (1996): Stochastic Processes, Second edition, Wiley.

[58] Schmidt J., Siegel A. and Srinivasan A. (1995): Chernoff-Hoeffding bounds for applications with limited independence, SIAM J. Discrete Math. **8**, 223 – 250.

[59] Sedgewick R. and Flajolet P. (1996): Analysis of Algorithms, Addison-Wesley.

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

[61] Shiryaev A.N. (1996): Probability, Second edition, Graduate Texts in Mathematics 95, Springer.

[62] Steele J.M. (1995): Variations on the long increasing subsequence theme of Erdős and Szekeres, in Discrete Probability and Algorithms, D. Aldous, P. Diaconis and J.M. Steele, eds., Volumes in Mathematics and its Applications **72**, Springer-Verlag, New York, 111 – 131.

[63] Steele J.M. (1997): Probability Theory and Combinatorial Optimization, SIAM CBMS 69.

[64] Steiger W.L. (1967): Some Kolmogoroff-type inequalities for bounded random variables, Biometrika **54**, 641 – 647.

[65] Steiger W.L. (1969): A best possible Kolmogoroff-type inequality for martingales and a characteristic property, Ann. Math. Statist. **40**, 764 – 769.

[66] Steiger W.L. (1970): Bernstein's inequality for martingales, Z. Wahrscheinlichkeitstheorie verw. Geb. **16**, 104 – 106.

[67] Talagrand M. (1991): A new isoperimetric inequality for product measure and the tails of sums of independent random variables, Geometric and Functional Analysis **1**, 211 – 223.

[68] Talagrand M. (1995): Concentration of measure and isoperimetric inequalities in product spaces, Publ. Math. Institut des Hautes Études Scientifiques **81**, 73 – 205.

[69] Talagrand M. (1996): A new look at independence, Annals of Probability **24**, 1 – 31.

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

[71] Talagrand M. (1996): Transportation cost for Gaussian and other product measures, Geometric and Functional Analysis **6**, 587 – 600.

[72] Williams D. (1991): Probability with Martingales, Cambridge University Press.