Concentration of Measure for the Analysis of Randomised Algorithms

Devdatt P. Dubhashi   Alessandro Panconesi

October 20, 1998

Contents

Chapter 1

Chernoff–Hoeffding Bounds

1.1 What is "Concentration of Measure"?

The basic idea of concentration of measure is well illustrated by the simplest of random experiments, and one lying at the fountain–head of probability theory: coin tossing. If we toss a fair coin once, the result is completely unpredictable – it can be "heads" or "tails" with equal probability. Now suppose we toss the same coin a large number of times, say, a thousand times. The outcome is now *sharply predictable*! Namely, the number of heads will be "very likely to be around 500". This apparent paradox, which is nevertheless familiar to everybody, is an instance of the phenomenon of the concentration of measure – although there are potentially a large number of possibilities, the ones that are likely to be observed are concentrated in a very narrow range, hence sharply predictable.

In more sophisticated forms, the phenomenon of the concentration of measure underlies much of our pysical world. As we know now, the world is made up of microscopic particles that are governed by probabilistic laws – those of quantum and statistical physics. The reason that the macroscopic properties determined by these large ensembles of particles nevertheless appear determinstic when viewed on our larger scales is precisely the concentration of measure: the observed possibilities are concentrated into a very narrow range.

Given the obvious importance of the phenomenon, it is no surprise that large parts of treatises on probability theory are devoted to its study. The various "Laws of Large Numbers" and the "Central Limit Theorem" are some of the most central results of modern probability theory.

We would like to use the phenomenon of concentration of measure in the analysis of probabilistic algorithms. In analogy with the physical situation described above, we would like to use it to argue that the observable behaviour of randomised algorithms is "almost deterministic". In this way, we can obtain the satisfaction of deterministic results, while at the same time retaining the benefits of randomised algorithms, namely their simplicity and efficiency.

In slightly more technical terms, the basic problem we want to study in this monograph is this: given a random variable X with mean $\mathtt{E}[X]$, what is the probability that X deviates far from its expectation? Furthermore, we would like to understand under what conditions the random variable X stays almost constant or, put in a different way, large deviation from the the mean are highly unlikely. This is the case for the familiar example of repeated coin tosses, but, as we shall see, it is a more general phenomenon.

There are several reasons that the results from probability theory are somewhat inadequate or inappropriate for studying these questions.

- First and foremost, the results in probability theory are *asymptotic limit laws* applying in the infinite limit. We are interested in laws that apply in finitary cases.
- The probability theory results are often *qualitative*: they ensure convergence in the limit, but do not consider the *rate* of convergence. We are interested in *quantitative* laws that determine the rate of convergence, or at least good bounds on it.
- The laws of probability theory are classically stated under the assumption of *independence.* This is a very natural and reasonable assumption in probability theory, and it greatly simplifies the statement and proofs of the results. However, in the analysis of randomised algorithms, whose outcome is the result of a complicated interaction of various processes, independence is the exception rather than the rule. Hence, we are interested in laws that are valid even without independence, or when certain known types of dependence obtain.

We shall now embark on a development of various tools and techniques that meet these criteria.

1.2 The Binomial Distribution

Let us start with an analysis of the simple motivating example of coin tossing. The number of "heads" or successes in repeated tosses of a fair coin is a very important distribution because it models a very basic paradigm of the probabilistic method, namely to repeat experiments to boost the confidence.

Let us analyse the slightly more general case of the number of "heads" in n trials with a coin of bias p, with $0 \leq p \leq 1$ i.e. $\texttt{Pr}[\texttt{Heads}] = p$ and $\texttt{Pr}[\texttt{Tails}] = 1-p$. This is a random variable $B(n,p)$ whose distribution is called the *Binomial distribution* with parameters n and p:

$$\texttt{Pr}[B(n,p) = i] = \binom{n}{i} p^i q^{n-i}, \quad 0 \leq i \leq n. \tag{1.1}$$

The general problem defined in the previous section here becomes the following: In the binomial case the expectation is $\texttt{E}[B(n,p)] = np$, we would like to get a bound on the probability that the variable does not deviate too far from this expected value. Are such large deviations unlikely for $B(n,p)$? A direct computation of the probabilites $\texttt{Pr}[B(n,p) \geq k] = \sum_{i \geq k} \binom{n}{i} p^i q^{n-i}$ is far too unwieldy. However, see Problem 1.2 for a neat trick that yields a good bound. We shall now introduce a general method that successfully solves our problem and is versatile enough to apply to many other problems that we shall encounter.

1.3 The Chernoff Bound

The random variable $B(n,p)$ can be written as a sum $X := \sum_{i \in [n]} X_i$, by introducing the indicator random variables $X_i, i \in [n]$ define by

$$X_i := \begin{cases} 1; & \text{if the } i\text{th trial is a success,} \\ 0; & \text{otherwise.} \end{cases}.$$

The basic Chernoff technique we are going to develop now applies in many situations where such a decomposition as a sum is possible.

The trick is to consider the so–called *moment–generating function* of X, defined as $\texttt{E}[e^{\lambda X}]$ where $\lambda > 0$ is a parameter. By formal expansion of the Taylor series, we see that

$$\begin{aligned} \texttt{E}[e^{\lambda X}] &= \texttt{E}[\sum_{i \geq 0} \frac{\lambda^i}{i!} X^i] \\ &= \sum_{i \geq 0} \frac{\lambda^i}{i!} \texttt{E}[X^i]. \end{aligned}$$

This explains the name as the function $\texttt{E}[e^{\lambda X}]$ is the exponential generating function of all the moments of X – it "packs" all the information about the moments of X into one function.

Now, for any $\lambda > 0$, we have

$$\begin{aligned}
\mathtt{Pr}[X > m] &= \mathtt{Pr}[e^{\lambda X} > e^{\lambda m}] \\
&\leq \frac{\mathtt{E}[e^{\lambda X}]}{e^{\lambda m}}.
\end{aligned} \tag{1.2}$$

The last step follows by *Markov's inequality*: for any non–negative random variable X, $\mathtt{Pr}[X > a] \leq \mathtt{E}[X]/a$.

Let us compute the moment generating function for our example:

$$\begin{aligned}
\mathtt{E}[e^{\lambda X}] &= \mathtt{E}[e^{\lambda \sum_i X_i}] \\
&= \mathtt{E}[\prod_i e^{\lambda X_i}] \\
&= \prod_i \mathtt{E}[e^{\lambda X_i}], \quad \text{by independence} \\
&= (pe^{\lambda} + q)^n
\end{aligned} \tag{1.3}$$

Substituting this back into (1.2), and using the parametrisation $m := (p+t)n$ which will lead to a convenient statement of the bound, we get:

$$\mathtt{Pr}[X > m] \leq \left(\frac{pe^{\lambda} + q}{e^{\lambda(p+t)}}\right)^n .$$

We can now pick $\lambda > 0$ to minimise the value between the paranthesis and by a simple application of Calculus, we arrive at the basic Chernoff bound:

$$\begin{aligned}
\mathtt{Pr}[X > (p+t)n] &\leq \left(\left(\frac{p}{p+t}\right)^{p+t} \left(\frac{q}{q-t}\right)^{q-t}\right)^n \\
&= \exp\left(-(p+t)\ln\frac{p+t}{p} - (q-t)\ln\frac{q-t}{q}\right)^n .
\end{aligned} \tag{1.4}$$

What shall we make of this mess? Certainly, this is not the most convenient form of the bound for use in applications! In § 1.6 we shall derive much simpler and more intellegible formulae that can be used in applications. For now we shall pause a while and take a short detour to make some remarks on (1.4). This is for several reasons: First, it is the strongest form of the bound. Second, and more importantly, this same bound appears in many other situations. This is no accident for it is a very natural and insightful bound – when properly viewed! For this, we need a certain concept from Information Theory.

Given two (discrete) probability distributions $p_1, \ldots, p_n$ and $q_1, \ldots, q_n$ on a space of cardinality n, the *relative entropy distance* between them, $H(\mathbf{p}, \mathbf{q})$ is defined by [1]:

$$H(\mathbf{p}, \mathbf{q}) := \sum_i -p_i \log \frac{p_i}{q_i}.$$

The expression multiplying n in the exponent in (1.4) is exactly the relative entropy distance of the distribution $p+t, q-t$ from the distribution p, q on the two point space $\{1, 0\}$. So (1.4) seen from the statistician's eye says: the probability of getting the "observed" distribution $\{p+t, q-t\}$ when the *a priori* or *hypothesis* distribution is $\{p, q\}$ falls exponentially in n times the relative entropy distance between the two distributions.

By considering $-X$, we get the same bound symmetrically for $\texttt{Pr}[X < (p-t)n]$.

1.4 Heterogeneous Variables

As a first example of the versatility of the Chernoff technique, let us consider the situation where the trials are heterogeneous: probabilities of success at different trials need not be the same. In this case, Chvatal's proof in Problem 1.2 is inapplicable, but the Chernoff method works with a simple modification. Let p_i be the probability of success at the ith trial. Then we can repeat the calculation of the moment–generating function $\texttt{E}[e^{\lambda X}]$ exactly as in (1.3) except for the last line to get:

$$\texttt{E}[e^{\lambda X}] = \prod_i (p_i e^\lambda + q_i). \tag{1.5}$$

Recall that the arithmetic–geometric mean inequality states that

$$\frac{1}{n}\sum_{i=1}^{n} a_i \le \left(\prod_{i=1}^{n} a_i\right)^{1/n}$$

for all $a_i \le 0$. Now employing the arithmetic–geometric mean inequality, we get:

$$\begin{aligned}\texttt{E}[e^{\lambda X}] &= \prod_i (p_i e^\lambda + q_i)\\ &\le \left(\frac{\sum_i (p_i e^\lambda + q_i)}{n}\right)^n\\ &= (pe^\lambda + q)^n,\end{aligned}$$

where $p := \frac{\sum_i p_i}{n}$, and $q := 1-p$. This is the same as (1.3) with p taken as the arithmetic mean of the p_is. The rest of the proof is as before and we conclude that the basic Chernoff bound (1.4) holds.

[1] Note that when $\mathbf{q}$ is the uniform distribution, this is just the usual *entropy* of the distribution $\mathbf{p}$ up to an additive term of $\log n$.

1.5 The Hoeffding Extension

A further extension by the same basic technique is possible to heterogeneous variables that need not even be discrete. Let $X := \sum_i X_i$ where each $X_i, i \in [n]$ takes values in $[0,1]$ and has mean p_i. To calculate the moment generating function $e^{\lambda X}$, we need, as before, to compute each individual $e^{\lambda X_i}$. This is no longer as simple as it was with the case where X_i took only two values.

However, the following convexity argument gives a simple upper bound. The graph of the function $e^{\lambda x}$ is convex and hence, in the interval $[0,1]$, lies always below the straight line joining the endpoints $(0,1)$ and $(1, e^{\lambda})$. This line has the equation $y = \alpha x + \beta$ where $\beta = 1$ and $\alpha = e^{\lambda} - 1$. Thus

$$\begin{aligned} \mathtt{E}[e^{\lambda X_i}] &\leq \mathtt{E}[\alpha X_i + \beta] \\ &= p_i e^{\lambda} + q_i. \end{aligned}$$

Thus we have

$$\mathtt{E}[e^{\lambda X}] \leq \prod_i \mathtt{E}[e^{\lambda X_i}] =\leq \prod_i (p_i e^{\lambda} + q_i),$$

which is the same bound as in (1.5) and the rest of the proof is concluded as before.

1.6 Useful Forms of the Bound

The following forms of the Chernoff–Hoeffding bound are most useful in applications (see also Problem 1.6).

Theorem 1.1 *Let $X := \sum_{i \in [n]} X_i$ where $X_i, i \in [n]$ are independently distributed in $[0,1]$. Then*

- *For all $t > 0$,*

$$\mathtt{Pr}[X > \mathtt{E}[X] + t], \mathtt{Pr}[X < \mathtt{E}[X] - t] \leq e^{-2t^2/n}. \tag{1.6}$$

- *For $0 < \epsilon < 1$,*

$$\mathtt{Pr}[X > (1+\epsilon)\mathtt{E}[X]] \leq \exp\left(-\frac{\epsilon^2}{3}\mathtt{E}[X]\right), \quad \mathtt{Pr}[X < (1-\epsilon)\mathtt{E}[X]] \leq \exp\left(-\frac{\epsilon^2}{2}\mathtt{E}[X]\right). \tag{1.7}$$

- *If* $t > 2e\mathtt{E}[X]$, *then*

$$\mathtt{Pr}[X > t] \leq 2^{-t}. \tag{1.8}$$

Proof. We shall manipulate the bound in (1.4). Set

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

We successively compute

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

and

$$f''(t) = \frac{1}{(p+t)(q-t)}.$$

Now, $f(0) = 0 = f'(0)$ and furthermore $f''(t) \geq 4$ for all $0 \leq t \leq q$ because $xy \leq \frac{1}{4}$ for any two non–negative reals summing to 1. Hence by Taylor's Theorem with remainder,

$$\begin{aligned} f(t) &= f(0) + f'(0)t + f''(\xi)\frac{t^2}{2!}, \quad 0 < \xi < t \\ &\geq 2t^2. \end{aligned}$$

This gives, after simple manipulations, the bound (1.6).

Now consider $g(x) := f(px)$. Then $g'(x) = pf'(px)$ and $g''(x) = p^2 f''(px)$. Thus, $g(0) = 0 = g'(0)$ and $g''(x) = \frac{p^2}{(p+px)(q-px)} \geq \frac{p}{1+x} \geq \frac{2p}{3x}$. Now by Taylor's theorem, $g(x) \geq px^2/3$. This gives the upper tail in (1.7).

For the lower tail in (1.7), set $h(x) := g(-x)$. Then $h'(x) = -g'(-x)$ and $h''(x) = g''(-x)$. Thus $h(0) = 0 = h'(0)$ and $h''(x) = \frac{p^2}{(p-px)(q+px)} \geq p$. Thus by Taylor's theorem, $h(x) \geq px^2/2$ and this gives the result.

For the (1.8), see Problem 1.6 ∎

1.7 A Variance Bound

Finally, we shall give an application of the basic Chernoff technique to develop a form of the bound in terms of the varainces of the individual summands, a form that can be considerably sharper than those derived above, and one which will be especially useful for applications we will encounter in later chapters.

Let us return to the basic Chernoff technique with $X := X_1 + \cdots + X_n$ and $X_i \in [0,1]$ for each $i \in [n]$. Set $\mu_i := \mathtt{E}[X_i]$ and $\mu := \mathtt{E}[X] = \sum_i \mu_i$. Then

$$\begin{aligned}
\mathtt{Pr}[X > \mu + t] &= \mathtt{Pr}[\sum_i (X_i - \mu_i) > t] \\
&= \mathtt{Pr}[e^{\lambda \sum_i (X_i - \mu_i)} > e^{\lambda t}] \\
&\leq \mathtt{E}[e^{\lambda \sum_i (X_i - \mu_i)}]/e^{\lambda t},
\end{aligned}$$

for each $\lambda > 0$. The last line follows again from Markov's inequality.

We shall now use the simple inequalities that $e^x \leq 1 + x + x^2$ for $0 < |x| < 1$, and $e^x \geq 1 + x$. Now, if $\lambda \max(\mu_i, 1 - \mu_i) < 1$ for each $i \in [n]$, we have,

$$\begin{aligned}
\mathtt{E}[e^{\lambda \sum_i (X_i - \mu_i)}] &= \prod_i \mathtt{E}[e^{\lambda (X_i - \mu_i)}] \\
&\leq \prod_i \mathtt{E}[1 + \lambda(X_i - \mu_i) + \lambda^2 (X_i - \mu_i)^2] \\
&= \prod_i (1 + \lambda^2 \sigma_i^2) \\
&\leq \prod_i e^{\lambda^2 \sigma_i^2} \\
&= e^{\lambda^2 \sigma^2},
\end{aligned}$$

where σ_i^2 is the variance of X_i for each $i \in [n]$ and σ^2 is the variance of X. Thus,

$$\mathtt{Pr}[X > \mu + t] \leq e^{\lambda^2 \sigma^2}/e^{\lambda t},$$

for λ satisfying $\lambda \max(\mu_i, 1 - \mu_i) < 1$ for each $i \in [n]$. By calculus, take $\lambda := \frac{t}{2\sigma^2}$ and we get the bound:

$$\mathtt{Pr}[X > \mu + t] \leq \exp\left(\frac{-t^2}{4\sigma^2}\right),$$

Shall we add a comment saying in what cases this is a good bound?

for $t < 2\sigma^2 / \max_i \max(\mu_i, 1 - \mu_i)$. This is a bound that is valid for relatively small deviations around the mean.

1.8 The Normal Approximation

Work in progress...

1.9 CH Bounds with Limited Independence

Work in progress...

1.10 Bibliographic Notes

The original technique is from Chernoff [4] although the idea of using the moment–generating function to derive tail bounds is attributed to S.N. Bernstein. The extension to continuous variables is due to W. Hoeffding [8]. Our presentation was much influenced by [10]. The quick derivation in Problems 1.2 and 5.7 are due to V. Chvátal [16].

1.11 Problems

Problem 1.1 *A set of n balls is drawn by sampling with replacement from an urn containing N balls, M of which are red. Give a sharp concentration result for the number of red balls in the sample drawn.*

Problem 1.2 *In this problem, we outline a simple proof of the Chernoff bound due to V. Chvátal.*
(a) Argue that for all $x \geq 1$, we have

$$\mathtt{Pr}[B(n,p) \geq k] \leq \sum_{i \geq 0} \binom{n}{i} p^i q^{n-i} x^{i-k}.$$

(b) Now use the Binomial Theorem and thereafter Calculus to optimise the value of x.

Problem 1.3 (Hypergeometric Distribution) *A set of n balls is drawn by sampling* **without** *replacement from an urn containing N balls, M of which are red. The random variable $H(N,M,n)$ of the number of red balls drawn is said to have the* **hypergeometric distribution**.
(a) What is $\mathtt{E}[H(N,M,n)]$?
(b) Can you apply CH bounds to give a sharp concentration result for $H(N,M,n)$? Now we outline a direct proof due to V. Chvátal for the tail of the hypergeometric distribution along the lines of the previous problem.
(c) Show that

$$\mathtt{Pr}[H(N,M,n) = k] = \binom{M}{k} \binom{N-M}{n-k} \binom{N}{n}^{-1}.$$

(d) Show that

$$\binom{N}{n}^{-1} \sum_{i \geq j} \binom{M}{i} \binom{N-M}{n-i} \binom{i}{j} \leq \binom{n}{j} \left(\frac{M}{N}\right)^j.$$

(e) Use the previous part to show that for every $x \geq 1$,

$$\sum_{i \geq 0} \binom{M}{i} \binom{N-M}{n-i} \binom{N}{n}^{-1} x^i \leq (1 + (x-1)M/N)^n.$$

(f) Combine parts (c) through (e) and optimise the value of x to derive the same relative entropy bound (1.4):

$$\mathtt{Pr}[H(N,M,n) \geq (p+t)n] \leq \left(\left(\frac{p}{p+t}\right)^{p+t} \left(\frac{q}{q-t}\right)^{q-t} \right)^n,$$

where $p := M/N$ and $q := 1-p$.

Problem 1.4 *Show that for* $0 < \alpha \le 1/2$,

$$\sum_{0 \le k \le \alpha n} \binom{n}{k} \le 2^{H(\alpha)n},$$

where $H(\alpha) := -\alpha \log \alpha - (1-\alpha)\log(1-\alpha)$ *is the binary entropy function.*

Problem 1.5 (Weierstrass Approximation Theorem) *Prove:* For every continuous function $f : [0,1] \to R$ and every $\epsilon > 0$, there is a polynomial p such that $|f(x) - p(x)| < \epsilon$ for every $x \in [0,1]$. *(*HINT: *Consider* $p_n(x) := \sum_{0 \le i \le n} \binom{n}{i} x^i (1-x)^{n-i} f(i/n)$.*)*

Problem 1.6 *Repeat the basic proof structure of the CH bounds to derive the following bound: if* $X_1, \ldots, X_n$ *are independent* $0/1$ *variables (not necessarily identical), and* $X := \sum_i X_i$, *then for any* $\epsilon > 0$,

$$\mathtt{Pr}\left[X \ge (1+\epsilon)\mathtt{E}[X]\right] \le \left(\frac{e^{\epsilon}}{(1+\epsilon)^{(1+\epsilon)}}\right)^{\mathtt{E}[X]}.$$

(a) Compare this bound to the one obtained by setting $t := \epsilon\mathtt{E}[X]/n$ *in the relative entropy bound derived in class.*
(b) Argue further that the right side is bounded by $(\frac{e}{1+\epsilon})^{(1+\epsilon)\mathtt{E}[X]}$ *and hence infer that if* $\epsilon > 2e - 1$, *then*

$$\mathtt{Pr}\left[X \ge (1+\epsilon)\mathtt{E}[X]\right] \le 2^{-(1+\epsilon)\mathtt{E}[X]}.$$

Problem 1.7 *In this problem, we explore a further extension of the CH bounds, namely to variables that are bounded in some arbitrary intervals, not necessarily* $[0,1]$. *Let* $X_1, \ldots, X_n$ *be independent variables such that for each* $i \in [n]$, $X_i \in [a_i, b_i]$ *for some reals* a_i, b_i.
(a) Suppose $a_i = a$ *and* $b_i = b$ *for each* $i \in [n]$. *Derive a bound by rescaling the Hoeffding bound for* $[0,1]$.
(b) Does the rescaling work for non–identical intervals?
(c) Derive the following general bound for non–identical intervals by repeating the basic proof technique:

$$\mathtt{Pr}[|X - \mathtt{E}[X]| \ge t] \le 2\exp\left(\frac{-2t^2}{\sum_i (b_i - a_i)^2}\right).$$

Problem 1.8 (Sums of Exponential Variables) *Let* $Z := Z_1 + \cdots + Z_n$ *where* $Z_i, i \in [n]$ *are independent and identically distributed with the* **exponential distribution** *with parameter* $\alpha \in (0,1)$. *The probability density function for this distribution is*

$$f(x) = \alpha e^{-\alpha x},$$

and the corresponding cumulative distribution function is

$$F(x) = \int_0^x f(t)dt = 1 - e^{-\alpha x}.$$

Give a sharp concentration result for the upper tail of Z.

Solution. Note that for each $i \in [n]$,

$$\mathtt{E}[Z_i] = \int_0^\infty x f(x)dx = \alpha \int_0^\infty x e^{-\alpha x} dx = \frac{1}{\alpha}.$$

Hence

$$\mathtt{E}[Z] = \sum_i \mathtt{E}[Z_i] = \frac{n}{\alpha}.$$

We cannot apply the Chernoff–Hoeffding bounds directly to Z since the summands are not bounded. One solution is to use the method of *truncation*. Let $Z_i', i \in [n]$ be defined by

$$Z_i' := \min(Z_i, n^\beta), \quad i \in [n],$$

for some $0 < \beta < 1$ to be chosen later. Let $Z' := \sum_i Z_i'$. Observe first that $\mathtt{E}[Z'] \leq \mathtt{E}[Z]$. Second, that since for each $i \in [n]$, $\mathtt{Pr}[Z_i > n^\beta] \leq 1 - F(n^\beta) = e^{-n^\beta}$,

$$\mathtt{Pr}[\bigwedge_i Z_i = Z_i'] \geq 1 - n e^{-n^\beta}.$$

Finally, since the summands Z_i' are bounded ($0 \leq Z_i' \leq n^\beta$, for each $i \in [n]$), one can apply the Chernoff–Hoeffding bounds from Problem 1.7. Hence,

$$\begin{aligned}
\mathtt{Pr}[Z > \mathtt{E}[Z] + t] &\leq \mathtt{Pr}[Z' > \mathtt{E}[Z] + t] + n e^{-n^\beta} \\
&\leq \mathtt{Pr}[Z' > \mathtt{E}[Z'] + t] + n e^{-n^\beta} \\
&\leq \exp\left(\frac{-t^2}{n^{1+2\beta}}\right) + n e^{-n^\beta}.
\end{aligned}$$

For $t := \epsilon \mathtt{E}[Z] = \epsilon \frac{n}{\alpha}$, this gives

$$\mathtt{Pr}[Z > (1+\epsilon)\mathtt{E}[Z]] \leq \exp\left(\frac{-\epsilon^2 n^{1-2\beta}}{\alpha^2}\right) + n e^{-n^\beta}.$$

To (approximately) optimise this, choose $\beta = \frac{1}{3}$. Then,

$$\mathtt{Pr}[Z > (1+\epsilon)\mathtt{E}[Z]] \leq \exp\left(\frac{-\epsilon^2 n^{1/3}}{\alpha^2}\right) + n e^{-n^{1/3}}.$$

Another approach is to apply the Chernoff technique directly. Compute the moment generating function

$$\mathtt{E}[e^{\lambda Z_i}] = \alpha \int_0^\infty e^{\lambda x} e^{-\alpha x} dx = \frac{\alpha}{\alpha - \lambda},$$

for $) < \lambda < \alpha$. Thus

$$\begin{aligned} \mathtt{E}[e^{\lambda Z}] &= \left(\frac{\alpha}{\alpha - \lambda}\right)^n \\ &= \left(\frac{1}{1 - \frac{\lambda}{\alpha}}\right)^n \end{aligned}$$

Hence,

$$\begin{aligned} \mathtt{Pr}[Z > t] &\leq \frac{\mathtt{E}[e^{\lambda Z}]}{e^{\lambda t}} \\ &= \frac{1}{e^{\lambda t}\left(1 - \frac{\lambda}{\alpha}\right)^n}. \end{aligned}$$

Using Calculus,we find the optimal value of λ to be $\lambda^* := \alpha - \frac{n}{t}$ and substituting this gives the bound:

$$\mathtt{Pr}[Z > t] \leq \left(\frac{\alpha t}{n}\right)^n e^{-\alpha t + n}.$$

With $t := (1 + \epsilon)\mathtt{E}[Z]$, this gives

$$\mathtt{Pr}[Z > (1 + \epsilon)\mathtt{E}[Z]] \leq \left(\frac{e^\epsilon}{1 + \epsilon}\right)^{-n}.$$

This is a much better bound than that achieved by truncation. $\triangle$

Problem 1.9 *Give a sharp concentration bound on the upper tail of* $Z_1^2 + \ldots + Z_n^2$ *where* $Z_i, i \in [n]$ *are i.i.d. variables with the exponential distribution as in the previous problem.*

Solution. For each $i \in [n]$, note that

$$\mathtt{E}[Z_i^2] = \alpha \int_0^\infty x^2 e^{-\alpha x} dx = \frac{2}{\alpha^2},$$

and hence

$$\mathtt{E}[Z_1^2 + \cdots + Z_n^2] = \frac{2n}{\alpha^2}.$$

Denote this by μ.

Apply the method of truncation as in the previous problem. With the same notation as in the previous problem,

$$\mathtt{Pr}[Z_1^2 + \cdots + Z_n^2 > \mu + t] \leq \mathtt{Pr}[Z_1^2 + \cdots + Z_n^2 > \mu + t] \leq \exp\left(\frac{-t^2}{n^{1+2\beta}}\right) + ne^{-n^\beta}.$$

With $t := \epsilon\mu$, this gives,

$$\mathtt{Pr}[Z_1^2 + \cdots + Z_n^2 > (1+\epsilon)\mu] \leq \exp\left(-4(\frac{\epsilon}{\alpha})^2 n^{1-2\beta}\right) + ne^{-n^\beta}.$$

Pick $\beta := \frac{1}{3}$ to approximately optimise this. △

Problem 1.10 *Suppose a fair die is tossed n times and let X be the total sum of all the throws.*
(a) Compute $\mathtt{E}[X]$.
(b) Give a sharp concentration estimate on X by applying the result of the previous problem.
(c) Can you improve this by deriving the bound from scratch using the basic technique?

Problem 1.11 *In this problem, we shall explore the following question: How does the concentration bound on non–identically distributed variables depend on the individual probabilities $p_1, \ldots, p_n$? Abbreviate $(p_1, \ldots, p_n)$ by $\mathbf{p}$. Let $B(n, \mathbf{p})$ denote the sumber of successes in n independent trials where the probability of success at the ith trial is p_i. Let*

$$L(c, \mathbf{p}) := \mathtt{Pr}[B(n, \mathbf{p}) \leq c], \quad U(c, \mathbf{p}) := \mathtt{Pr}[B(n, \mathbf{p}) \geq c].$$

Fix some $\lambda > 0$. We shall explore how L and U are related for different $\mathbf{p}$ in the region

$$D(\lambda) := \{\mathbf{p} \mid \mathbf{0} \leq \mathbf{p} \leq \mathbf{1}, \sum_i p_i = \lambda\}.$$

Let

$$\mathbf{p}^*(\lambda) := n^{-1}(\lambda, \ldots, \lambda), \quad \hat{\mathbf{p}}(\lambda) := (1, \ldots, 1, \lambda - [\lambda], 0, \ldots, 0).$$

The first corresponds to the identical uniform case and the second (with $[\lambda]$ ones and $n - [\lambda] - 1$ zeroes) to the other extreme. Note that both $\mathbf{p}^*(\lambda), \hat{\mathbf{p}}(\lambda) \in D(\lambda)$.
(a) Show that for any $\mathbf{p} \in D(\lambda)$,

$$L(c, \hat{\mathbf{p}}) \leq L(c, \mathbf{p}) \leq L(c, \mathbf{p}^*) \quad \text{if } 0 \leq c \leq \lfloor \lambda - 2 \rfloor,$$

and

$$U(c, \mathbf{p}) \leq U(c, \mathbf{p}^*) \leq U(c, \hat{\mathbf{p}}) \quad \text{if } \lfloor \lambda + 2 \rfloor \leq c \leq n.$$

(b) More generally, let $\mathbf{p}, \mathbf{p}' \in D_\lambda$ *be such that there is a doubly stochastic matrix* Π *with* $\mathbf{p}' = \mathbf{p}\Pi$. *Equivalently, if*

$$p_{\sigma(1)} \geq p_{\sigma(2)} \geq \ldots \geq p_{\sigma(n)}, \quad p'_{\sigma'(1)} \geq p'_{\sigma'(2)} \geq \ldots \geq p'_{\sigma'(n)},$$

then for each $1 \leq k \leq n$,

$$\sum_{i \leq k} p_{\sigma(i)} \geq \sum_{i \leq k} p'_{\sigma'(i)}.$$

The vector $\mathbf{p}$ *is said to* **majorise** *the vector* $\mathbf{p}'$. *Show that*

$$L(c, \mathbf{p}) \leq L(c, \mathbf{p}') \quad \text{if } 0 \leq c \leq \lfloor \lambda - 2 \rfloor,$$

and

$$U(c, \mathbf{p}) \leq U(c, \mathbf{p}') \quad \text{if } \lfloor \lambda + 2 \rfloor \leq c \leq n.$$

Verify that this generalises part (a).

Chapter 2

Applications of the Chernoff-Hoeffding Bounds

In this chapter we present some non-trivial applications of the Chernoff-Hoeffding bounds arising in the design and analysis of randomised algorithms. The examples are quite different, a fact that illustrates the usefullness of these bounds, and were chosen because of their importance and elegance.

2.1 Data Structures: Skip Lists and Treaps

The first set of examples concerns the design and analysis of data structures. We shall discuss two data structures known as Skip Lists and Treaps.

2.1.1 Skip Lists: The Data Structure

We want to devise a data structure that efficiently supports the operations of inserting, deleting and searching for an element. Elements are drawn from a totally ordered universe X of size n, which can be assumed to be a finite set of natural numbers. The basic idea is as follows. Order the elements and arrange them in a linked list. We call this the 0-th level and denote it by L_0. It is convenient to assume that the list starts with the element $-\infty$. With this convention

$$L_0 = -\infty \to x_1 \to x_2 \to \ldots \to x_n.$$

We now form a new linked list L_1 by selecting every second element from L_0 and putting $-\infty$ in front.

$$L_1 = -\infty \to x_2 \to x_4 \to \ldots \to x_m.$$

Identical elements in the two lists are joined by double pointers, including $-\infty$'s. Continuing in this fashion we obtain a structure with $O(\log n)$ levels like the one in Figure 0.1. This structure resembles a binary tree and likewise allows for

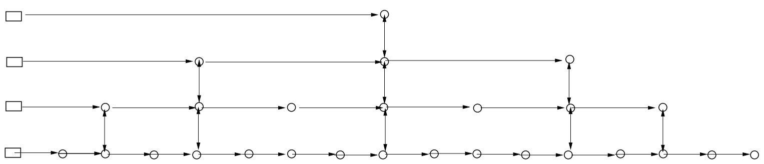

Figure 0.1: A skip list for 16 elements. Boxes store $-\infty$'s, circles store the data.

efficient searches. To search for an element y we start from the top list L_t and determine the largest element of L_t which is smaller than or equal to y. Denote such element by e_t. Then we go down one level, position ourselves on the copy of e_t and look for the largest element of L_{t-1} smaller than or equal to y. To do so we only need to scan L_{t-1} to the right of e_t. Continuing in this fashion we generate a sequence $e_t, e_{t-1}, \ldots, e_0$ where e_0 is the largest element in X smaller than or equal to y. Clearly, y is present in the data structure if and only if $e_0 = y$. Although an element could be encountered before reaching L_0, we assume that the search continues all the way down. This makes sense in applications for which the elements are keys pointing to records. In such cases one might not want to copy a whole record at higher levels. This convention also simplifies the probabilistic analysis to follow.

When performing a search we traverse the data structure in a zig-zag fashion, making only downturns and left turns (see Figure 0.2). The cost of the traversal is proportional to the sum of the height and the width of this path, both of which are $O(\log n)$. The width is $O(\log n)$ because each time we go down one level we roughly halve the search space. Searches are inexpensive as long as the data structure stays balanced. The problem is that insertions and removals can destroy the symmetry, making maintenance both cumbersome and expensive. By using randomization we can retain the advantages of the data structure while keeping the cost of reorganizations low.

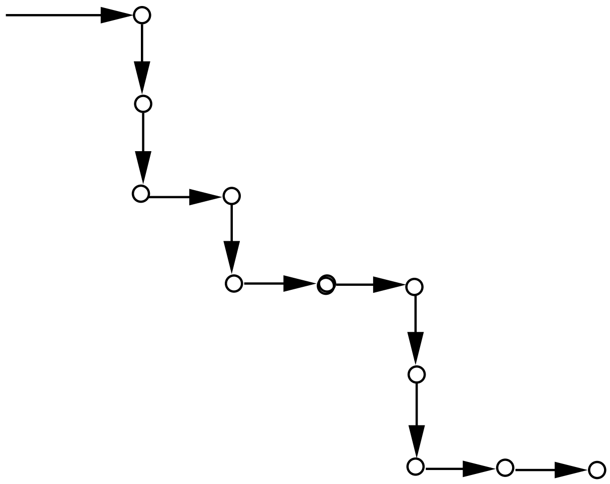

Figure 0.2: A zig-zag path through the data structure generated by a search.

2.1.2 Skip Lists: Randomization makes it easy

As before, L_0 is an ordered list of all the elements. Subsequent levels are built according to the following probabilistic rule: Given that an element x appears in level i, it is chosen to appear in level $i+1$ with probability p, independently of the other elements. Thus, the highest level that an element appears in obeys a geometric distribution with parameter p. If we denote by H_i the highest level to which x_i belongs, then

$$\Pr[H_i = k] = p^k(1-p). \tag{2.1}$$

H_i is called the *height* of x_i. The data structure is organized as before, with each level being an ordered list of elements starting with $-\infty$, and with copies of the same element at different levels arranged in a doubly linked list. Such a data structure is called a *skip list*.

A search is implemented as before. To insert an element x we do as follows. First, we search for x. This generates the sequence $e_t, e_{t-1}, \ldots, e_0$. Second, we insert x in L_0 between e_0 and its successor. Then we flip a coin; if the outcome is TAIL we stop, otherwise we insert x in L_1 between e_{t-1} and its successor. And so on, until the first TAIL occurs. Although we could stop when the last level is reached, we do not do so because this would slightly complicated the probabilistic analysis.

To remove an element x, we first locate it by means of a search and then remove all occurrences of x from all levels, modifying the pointers of the various lists in the obvious way.

How expensive are these operations? When inserting or deleting an element x_i

the cost is proportional to that of a search for x_i plus H_i. As we shall see, the cost of each search is upperbounded by the *height* of the data structure, defined as

$$H := \max_i H_i. \tag{2.2}$$

The cost of a search for x_i is proportional to the length of a zig-zag path of height $H_i \leq H$ and width W_i. We will prove that with high probability the orders of magnitude of W_i and H are the same. Intuitively, this is because the data structure stays roughly balanced. For we expect one in every $1/p$ elements of L_k to belong to L_{k+1} When L_k is sizeable large deviations are unlikely.

We now study the random variable H. First, we prove that $H = O(\log n)$ with high probability.

Proposition 2.1 $\Pr[H > a \log n] \leq n^{-a+1}$, *for any* $a > 0$.

Proof. The height H_i of any element $i \in [n]$ in the list is a geometrically distributed random variable with the parameter p:

$$\mathtt{Pr}[H_i = k] = p^k q, \quad k \geq 0. \tag{2.3}$$

Hence for $\ell \geq 1$,

$$\begin{aligned} \mathtt{Pr}[H_i > \ell] &= \sum_{k>\ell} \mathtt{Pr}[H_i = k] \\ &= \sum_{k>\ell} p^k q \\ &= p^{\ell+1}. \end{aligned} \tag{2.4}$$

The height of the skip list, H is given by

$$H = \max_i H_i. \tag{2.5}$$

Hence,

$$\begin{aligned} \mathtt{Pr}[H > \ell] &= \mathtt{Pr}[\bigvee_i H_i > \ell] \\ &\leq \sum_i \mathtt{Pr}[H_i > \ell] \\ &= np^{\ell+1}. \end{aligned} \tag{2.6}$$

In particular, for $\ell := a \log_p n - 1, (a > 0)$,

$$\mathtt{Pr}[H > a \log_p n] \leq n^{-a+1}. \tag{2.7}$$

■

Refer now to Figure 0.2. The cost of a traversal is equal to the number of $\downarrow$'s plus the number of $\rightarrow$'s. If an $\downarrow$-edge is traversed then the element x must be stored in the two consecutive levels L_k and L_{k+1} of the data structure. This means that when x flipped its coin to determine whether to percolate up from L_k to L_{k+1} the outcome was HEAD. Similarly, if an element x is entered from the left with a $\rightarrow$ it means that when x flipped its coin at level L_k the outcome was TAIL. We label each $\downarrow$ with p– denoting *success*– and each $\rightarrow$ with q— denoting *failure*. *Then, the number of arrows in the path is equal to the number of times needed to toss a coin with bias p in order to obtain H successes.* The distribution of the random variable defined as the number of tosses of a coin with bias p needed to obtain k successes, is called *negative binomial* and the random variable is denoted by $W(k,p)$. $W(k,p)$ is closely related to the binomial distribution $B(n,p)$.

In order to show that $W(k,p) = O(\log n)$ with high probability we can proceed in several different ways, some of which are is explored in the problem section. Perhaps the simplest approach is to start by establishing a connection with the binomial distribution.

Proposition 2.2 $\Pr(W(k,p) \leq m) = \Pr(B(m,p) \geq k)$.

Proof. See Exercise 2.1. ■

Let a and b be two parameters to be fixed later. Define

$$\begin{aligned} k &:= a \log n \\ m &:= b \log n. \end{aligned}$$

The first of these two values will upperbound H while the second will upperbound the time of a search, i.e. the total number of $\downarrow$'s and $\rightarrow$'s of a traversal. By Proposition 2.2

$$\Pr(W(k,p) > m) = \Pr(B(m,p) < k)$$

which translates the problem into that of estimating deviations of the binomial distribution below the mean. Recalling Theorem 1.1, i.e. the CH-bounds in usable forms,

$$\Pr(B(m,p) < E[B(m,p)] - t) = \Pr(B(m,p) < pm - t) \leq e^{-2t^2/m}.$$

By setting

$$k = pm - t$$

and solving for t, we get

$$t = (pb - a)\log n,$$

which gives

$$\Pr(B(m,p) < k) \leq \frac{1}{n^{(pb-a)^2/b}}.$$

Recalling Proposition 2.1, and setting $a = 2$, $b = 8$, and $p = 1/2$

$$\begin{aligned}
&\Pr(\text{cost of search} > m) = \\
&\qquad \Pr(\textit{cost of search} > m \mid H \leq k)\Pr(H \leq k) + \\
&\qquad \Pr(\textit{cost of search} > m \mid H > k)\Pr(H > k) \\
&\quad \leq \Pr(\textit{cost of search} > m \mid H \leq k) + \Pr(H > k) \\
&\quad \leq \Pr(W(k,p) > m) + \Pr(H > k) \\
&\quad \leq \frac{1}{n^{(pb-a)^2/b}} + \frac{1}{n^{a-1}} \\
&\quad = \frac{2}{n}.
\end{aligned}$$

Therefore with probability at least $\left(1 - \frac{2}{n}\right)$ no search ever takes more than $8\log n$ steps. Furthermore, with at least the same probability, no insert or delete ever takes more than $W(H,p) + H \leq (a+b)\log n = 10\log n$ steps.

2.1.3 Treaps

Work in progress...

2.2 Randomized rounding

Work in progress...

2.3 Bibliographic Notes

Skiplists are an invention of W. Pugh [12]

2.4 Problems

In the next three problems, we shall derive bounds on the sums of independent *geometrically distributed* variables. Let $W(1,p)$ denote the number of tosses required to obtain a "heads" with a coin of bias p (i.e. $\mathtt{Pr}(\text{heads}) = p, \mathtt{Pr}(\text{tails}) = 1 - p =: q$). Note that $\mathtt{Pr}[W(1,p) = \ell] = q^{\ell-1}p$, for $\ell \geq 1$. Let $W(n,p)$ denote the number of tosses needed to get n heads. Note that $W(n,p) = \sum_{i\in[n]} W_i(1,p)$, where the $W_i, i \in [n]$ are independent geometrically distributed variables with parameter p. The variable $W(n,p)$ is said to have a *negative binomial distribution.*

I'd like to insert a two lines comment here

Problem 2.1 *Prove that* $\Pr(W(k,p) \leq m) = \Pr(B(m,p) \geq k)$.

This is actually an exercise...

Problem 2.2 *A second approach to derive concentration results on $W(n,p)$ is to apply the basic Chernoff technique. Consider for simplicity the case $p = \frac{1}{2} = q$.*
(a) Show that for any integer $r \geq 1$, and for any $0 < \lambda < \ln 2$,

Does this work if we replace 2 with $1/p$?

$$\mathtt{Pr}[W(n,p) \geq (2+r)n] \leq \left(\frac{e^{-\lambda(r+1)}}{2-e^{\lambda}}\right)^n.$$

(b) Use Calculus to find the optimal λ and simplify to derive the bound that for $r \geq 3$,

$$\mathtt{Pr}[W(n,p) \geq (2+r)n] \leq e^{-rn/4}.$$

You may find it useful to note that $1 - x \leq e^{-x}$ and that $1 + r/2 \leq e^{r/4}$ for $r \geq 3$. Compare this bound with the one from the previous problem.

Solution. Work in progress... △

Problem 2.3 *Here is a third approach to the negative binomial distribution.*
(a) By explicit computation, show that

$$\mathtt{Pr}[W(n,p) \geq \ell] = \left(\frac{p}{q}\right)^n \sum_{t\geq\ell} q^t \binom{t-1}{n}.$$

(b) Let $S_n := \sum_{t\geq\ell} q^t \binom{t-1}{n}$. Show that

$$S_n = \frac{q}{p}\left(q^{\ell-1}\binom{\ell-1}{n} + S_{n-1}\right).$$

Hence deduce that

$$\mathtt{Pr}[W(n,p) \geq \ell] = q^{\ell-1} \sum_{0\leq i\leq n} \left(\frac{q}{p}\right)^{n+1-i} \binom{\ell-1}{i}.$$

(c) Consider the case $p = 1/2 = q$ and find a bound for $\mathtt{Pr}[W(n,p) \geq (2+r)n]$ *and compare with the previous problem.*

Solution. Work in progress... △

Problem 2.4 *Prove a sharp concentration result for the space used by a skip list.*

Solution. The space used by the data structure is

$$S_n = \sum_{i \in [n]} H_i \tag{2.8}$$

In view of (2.1), this is therefore a sum of geometric random variables i.e. $S_n = W(n,p)$. Thus $\mathtt{E}[S_n] = n/p$. Recalling Proposition 2.2 (or, equivalently, Problem 2.1), for $r > 0$,

$$\begin{aligned}
\mathtt{Pr}[S_n \geq (r+1/p)n] &= \mathtt{Pr}[B((r+1/p)n,p) \leq n] \\
&= \mathtt{Pr}[B((r+1/p)n,p) \leq (rpn+n) - rpn] \\
&= \mathtt{Pr}[B((r+1/p)n,p) \leq \mathtt{E}[B((r+1/p)n,p)] - rpn] \\
&\leq \exp\left(-2r^2p^2n\right)
\end{aligned}$$

where the last line follows by the Chernoff bound on the binomial distribution. △

Problem 2.5 *What is the best value of p in order to minimize the expected time of a search operation?*

Chapter 3

Generalization of the Chernoff-Hoeffding bounds

This chapter, which is still **work in progress**, discuses two topics:

3.1 Negative Dependence:

A brief treatment of an incipient theory (partly developed by one of the authors) to systematically deal with a kind of dependence that occurs frequently in resource–sharing applications. The emphasis here is on how the Chernoff–Hoeffding bounds can be salvaged in this context.

3.2 Limited Independence:

A brief treatment of Chernoff–Hoeffding bounds in situations where the underlying variables are only partially independent. This involves an insightful new approach to the Chernoff bound due to J. Schmidt, A. Siegel and A. Srinivasan. This and the previous topic are a genuine Computer Science contribution to the theory of concentration of measure.

Chapter 4

Martingales and the Method of Bounded Differences

The Chernoff-Hoeffding bounds provide very sharp concentration estimates when the random variable X under consideration can be expressed as the sum $X = X_1 + \ldots + X_n$ of independent (and bounded) random variables. However in many applications, to do this might be very difficult or impossible. It would therefore be useful to obtain sharp concentration results for the case when X is some complicated function of not necessarily independent variables. Such a generalization would be useful in many diverse contexts but especially in the analysis of randomized algorithms where the parameters that characterize the behaviour of the algorithm are the result of a complicated interaction among a base set of (often non–independent) random variables. Our goal then is to study the case when

$$X := f(X_1, X_2, \ldots, X_n),$$

where f is a function that may not even be explicitly specified in a "closed form". We seek a set of conditions on f so that one can assert that the probability of a large deviation of f from its expected value is exceedingly small– ideally, exponentially small in the amount of the deviation. In general, we would like to be able to do this even without assuming that the X_i's are independent.

We will present a number of such inequalities, all of which rest upon a well-studied concept of Probability Theory known as *martingales.* We shall see that once the appropriate concept to replace independence is properly formulated, the proofs of these inequalities are quite similar to the basic structure of the proofs of the Chernoff–Hoeffding bounds we have already seen.

4.1 Review of Conditional Probabilities and Expectations

The concept of a martingale requires a good understanding of the notions of conditional probability and expectation, so we first provide a quick review from an elementary standpoint.

Given two events $\mathcal{E}$ and $\mathcal{F}$ in a probability space with measure $\mathtt{Pr}$, the *conditional probability* of $\mathcal{E}$ with respect to $\mathcal{F}$ is defined by

$$\mathtt{Pr}[\mathcal{E} \mid \mathcal{F}] := \frac{\mathtt{Pr}[\mathcal{E} \wedge \mathcal{F}]}{\mathtt{Pr}[\mathcal{F}]},$$

provided that $\mathtt{Pr}[\mathcal{F}] \neq 0$. If $\mathtt{Pr}[\mathcal{F}] = 0$, then, by convention we shall set $\mathtt{Pr}[\mathcal{E} \mid \mathcal{F}] = 0$.

Often we will be interested in events of the form $X = a$, that a random variable X takes the value a, or that a sequence $X_1, \ldots, X_n$ takes the values $a_1, \ldots a_n$ respectively. For economy of notation, we shall use the vector boldface notation to stand for a finite or infinite sequence of the appropriate type. Thus a sequence of variables $X_1, X_2, \ldots$ will be denoted by $\mathbf{X}$ and a sequence of real values $a_1, a_2, \ldots$ by $\mathbf{a}$. When given such a sequence, we shall use the subscript n to denote the prefix of length n; thus $\mathbf{X}_n$ will denote $X_1, \ldots, X_n$ and $\mathbf{a}_n$ will denote $a_1, \ldots, a_n$. If n is less than the starting index of the sequence under consideration, the prefix sequence is empty. With these conventions the event $X_1 = a_1, \ldots, X_n = a_n$ can be abbreviated by $\mathbf{X}_n = \mathbf{a}_n$. We can always assume that such an event occurs with non–zero probability by discarding from the domain, the values for which it is zero.

The *conditional expectation* of a random variable Y with respect to an event $\mathcal{E}$ is defined by

$$\mathtt{E}[Y \mid \mathcal{E}] := \sum_b b \cdot \mathtt{Pr}[Y = b \mid \mathcal{E}]. \tag{4.1}$$

In particular, if the event $\mathcal{E}$ is $X = a$, this equation defines a function f, namely

$$f(a) := \mathtt{E}[Y \mid X = a].$$

Thus $\mathtt{E}[Y \mid X]$ is a random variable, namely the variable $f(X)$. In the same way, if the event $\mathcal{E}$ in (4.1) is $\mathbf{X} = \mathbf{a}$, we have a multivariate function

$$f(\mathbf{a}) := \mathtt{E}[Y \mid \mathbf{X} = \mathbf{a}],$$

and $\mathtt{E}[Y \mid \mathbf{X}]$ can be regarded as the random variable $f(\mathbf{X})$.

Regarding $\mathtt{E}[Y \mid X]$ as a random variable, we can ask what is its expectation? The answer involves some fundamental properties of conditional expectation that are listed in the next proposition and whose verification we leave as an exercise.

Proposition 4.1 *Let X, Y and Z be random variables defined on a probability space. Then, for arbitrary functions f and g,*

$$\mathtt{E}[\mathtt{E}[f(X)g(X,Y) \mid X]] = \mathtt{E}[f(X)\mathtt{E}[g(X,Y)] \mid X].$$

Also,

$$\mathtt{E}[X] = \mathtt{E}[\mathtt{E}[X|Y]],$$

and,

$$\mathtt{E}[X \mid Z] = \mathtt{E}[\mathtt{E}[X \mid Y, Z] \mid Z].$$

The formal verification of these is left as an exercise to the reader. Nevertheless it is perhaps appropriate to give an intuitive justification of these formulae which at first might appear somewhat obscure. The first equality is based on the simple fact that

$$\mathtt{E}[f(X)g(X,Y) \mid X = a] = f(a)\mathtt{E}[g(X,Y) \mid X = a]$$

which simply says that once the value of X is given $f(X)$ becomes a constant and can be taken out of the expectation. The second equality can be intuitively explained as follows. Suppose that X is a random variable representing, say, the height of individuals of a given population, and that Y is the age of an individual. In order to compute $\mathtt{E}[X]$– the average height– we can either do it directly or proceed as follows. Partition the population according to age, recording for each age group the fraction of the total population. To make things concrete, the 15 year olds could be 7% of the total population, the 32 year olds 11%, etc. Then, compute the average height in each age group– the average height of 15 year olds, of 32 year old, and so on. Finally, compute the weighted average of these averages by weighing each age group according to its share of the total population. This will give the average height of the whole population. The third equality is the same as the second one, except that we focus on a particular subset of the whole population. For instance Z could represent the sex of an individual. Sticking to our example, the formula asserts that in order to compute the average height of, say, the male population we can proceed as just described.

Proposition 4.1 generalises smoothly to the multivariate case. Once again we leave the verification as an exercise.

Proposition 4.2 (Fundamental Facts about Conditional Expectation) *Let* $\mathbf{X}$, $\mathbf{Y}$ *and Z be random variables defined on a probability space. For arbitrary* *functions f and g,*

$$\mathtt{E}[\mathtt{E}[f(\mathbf{X})g(\mathbf{X},\mathbf{Y}) \mid \mathbf{X}]] = \mathtt{E}[f(\mathbf{X})\mathtt{E}[g(\mathbf{X},\mathbf{Y}) \mid \mathbf{X}]]. \tag{4.2}$$

Also,

$$\mathtt{E}[X] = \mathtt{E}[\mathtt{E}[X \mid \mathbf{Y}]], \tag{4.3}$$

and

$$\mathtt{E}[X \mid \mathbf{Z}] = \mathtt{E}[\mathtt{E}[X \mid \mathbf{Y},\mathbf{Z}] \mid \mathbf{Z}]. \tag{4.4}$$

These facts will be heavily used in this chapter.

4.2 Martingales and Azuma's Inequality.

Martingales are a well-studied concept in classical probability. Here we will develop them in a discrete setting in the simplest form, which is sufficient for our purposes.

Definition 4.3 *A* martingale *is a sequence of random variables $X_0, X_1, \ldots$ such that*

$$\mathtt{E}[X_i | X_0, X_1, \ldots, X_{i-1}] = X_{i-1} \quad i \geq 1.$$

With the vector notation, the martingale condition is succintly expressed as

$$\mathtt{E}[X_i | \mathbf{X}_{i-1}] = X_{i-1}, \quad i \geq 1$$

The next examples and exercises should help clarify the definition.

Example 4.4 A fair coin is flipped n times. Let $X_i \in \{-1, 1\}$ denote the outcome of the i-th trial (with -1 standing for "tails" and $+1$ for "heads"). Let $S_0 :=$ 0 and $S_n := \sum_{i \leq n} X_i$. The variables $S_i, i \geq 0$ define a martingale. First, observe that they satisfy the so-called Markov property, $\mathtt{E}[S_n | S_0, \ldots, S_{n-1}] = \mathtt{E}[S_n | S_{n-1}]$, which intuitively says that the future outcome depends only on the current state. Hence,

$$\begin{aligned}
\mathtt{E}[S_n | S_0, \ldots, S_{n-1}] &= \mathtt{E}[S_n | S_{n-1}] \\
&= \mathtt{E}[S_{n-1} + X_n | S_{n-1}] \\
&= S_{n-1} + \mathtt{E}[X_n | S_{n-1}] \\
&= S_{n-1} + \mathtt{E}[X_n], \quad \text{by independence of the coin tosses} \\
&= S_{n-1}, \quad \text{since } \mathtt{E}[X_n] = 0.
\end{aligned}$$

Think of a gambler who starts with an initial fortune of $S_0 := 0$ and repeatedly bets an amount of 1 unit on a coin toss. Thus his fortune can go up or down by one unit equiprobably on each toss. His fortune after n tosses is S_n. Think of the sequence $S_0, S_1, \ldots$ as a sequence of dependent random variables. Before his nth wager, the gambler knows only the numerical values of $S_0, \ldots, S_{n-1}$ but can only guess at the future $S_n, S_{n+1}, \ldots$. If the game is fair, then conditional on the past information, he will expect no change in his present capital on average. This is exactly the martingale condition. $\bigtriangledown$

Example 4.5 Suppose now that $X_i \in \{0, 1\}$ equiprobably for each $i \in [n]$. Now it is no longer true that $\mathtt{E}[X_i] = 0$. Nevertheless, a martingale can be defined by letting $S_k := \sum_{i \leq k} X_i - \frac{k}{2}$ with $S_0 := 0$. The straightforward verification is left as an exercise. $\bigtriangledown$

Exercise 4.6 *Let $X_i \in \{0, 1\}$ $(1 \leq i \leq n)$ be a set of n variables such that $\Pr[X_i = 1] = p_i$. Can you generalize example 4.4?*

The following definition is central.

Definition 4.7 (Bounded Differences). *Let $X_0, X_1, \ldots$ be a martingale. The X_i's satisfy the* Bounded Difference Condition *(BDC) with parameters c_i if*

$$|X_i - X_{i-1}| \leq c_i$$

for some non-negative constants $c_i, i \geq 1$.

The following concentration result for martingales is known as Azuma's Inequality although it appears also in an earlier paper by Hoeffding. It will provide us with a basic tool for our generalization.

Theorem 4.8 (Azuma's Inequality). *Let $X_0, X_1, \ldots$ be a martingale satisfying the Bounded Difference Condition with parameters $c_i, i \geq 1$. Then,*

$$\Pr(X_n > X_0 + t) \leq \exp\left(-\frac{t^2}{2c}\right)$$

and

$$\Pr(X_n < X_0 - t) \leq \exp\left(-\frac{t^2}{2c}\right)$$

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

Before proving the theorem some comments are in order. First, notice that there is no assumption of independence. Second, if we think of a martingale sequence as keeping track of a process evolving through time– where X_i is the measurement at time i– the Bounded Difference Condition roughly states that the process does not makes big jumps. Azuma's inequality roughly says that if this is so, then it is unlikely that the process wanders very far from its starting point. Clearly this crucially depends on the martingale property. Notice also that c appears in the denominator, which means that the smaller the c_i's the sharper the concentration.

In the proof of Azuma's inequality we shall use several ideas already encountered in the derivation of various forms of the CH-bounds. The assumption of independence will be replaced by the martingale property, while the assumption that the summands are bounded is replaced by the Bounded Difference Condition (BDC).

Now to the proof. We shall prove the statement for the upper tail, the proof for the lower tail is symmetrical with the martingale $\mathbf{X}$ replaced by $-\mathbf{X}$. To start with, we can assume without loss of generality that $X_0 := 0$. Otherwise we can define the translated sequence $X_i' := X_i - X_0$ which satisfies the conditions equally well. We then apply the Chernoff Method starting with Markov's Inequality:

$$\begin{aligned} \Pr(X_n > t) &= \Pr(e^{\lambda X_n} > e^{\lambda t}) \\ &\le \frac{\mathtt{E}[e^{\lambda X_n}]}{e^{\lambda t}} \end{aligned} \tag{4.5}$$

for all $\lambda > 0$. Now we focus on the numerator $\mathtt{E}[e^{\lambda X_n}]$: we want to find a good upper bound in terms of λ and then find the value of λ that minimizes the ratio $\mathtt{E}[e^{\lambda X_n}]/e^{\lambda t}$.

We define the *martingale difference sequence*:

$$Y_i := X_i - X_{i-1}, \quad i \ge 1$$

which allows us to express the martingale as the sum of increments:

$$X_k = \sum_{i \le k} Y_i.$$

Note that the martingale condition can be rephrased as follows:

$$\mathtt{E}[Y_i \mid \mathbf{X}_{i-1}] = 0.$$

Applying the basic equality (4.3), we get

$$\mathtt{E}[e^{\lambda X_n}] = \mathtt{E}[\exp(\lambda \sum_{i \le n} Y_i)]$$

$$
\begin{aligned}
&= \mathtt{E}[\mathtt{E}[\exp(\lambda \sum_{i \le n} Y_i) | \mathbf{X}_{n-1}]] \\
&= \mathtt{E}[\mathtt{E}[\exp(\lambda X_{n-1}) e^{\lambda Y_n} | \mathbf{X}_{n-1}]] \\
&= \mathtt{E}[\exp(\lambda X_{n-1}) \mathtt{E}[e^{\lambda Y_n} | \mathbf{X}_{n-1}]].
\end{aligned}
$$

The last equality follows from (4.2).

Now the proof continues by looking for a good upper bound for $\mathtt{E}[e^{\lambda Y_n} | \mathbf{X}_{n-1}]$. Denoting such a good upperbound by $U_n(\lambda)$ we obtain by induction,

$$
\begin{aligned}
\mathtt{E}[e^{\lambda X_n}] &= \mathtt{E}[\exp(\lambda X_{n-1}) \mathtt{E}[e^{\lambda Y_n} | \mathbf{X}_{n-1}]] \\
&\le \mathtt{E}[\exp(\lambda X_{n-1})] \cdot U_n(\lambda) \qquad (4.6) \\
&\le \prod_{i=1}^{n} U_i(\lambda) =: U(\lambda).
\end{aligned}
$$

As it happens there are two different ways to find good upperbounds $U_i(\lambda)$. The first, to be used in the next lemma, is based on the convexity of the e^x function– a fact already exploited to derive the Hoeffding generalization of the Chernoff bounds. The second uses a different idea, which we used in § 1.7. This second approach leads to another useful generalization of the CH-bounds which we shall call the *the Method of Bounded Variances*, and to which we return in a later chapter.

Lemma 4.9 *Let Z be a random variable such that $\mathtt{E}[Z] = 0$ and $|Z| \le c$. Then, $\mathtt{E}[e^{\lambda Z}] \le e^{\lambda^2 c^2/2}$.*

Even if this is repeated some redundancy which makes chapters self-contained might be good?

Proof. Let $f(x) := e^{\lambda x}$, $P_- := (-c, f(-c))$, $P_+ := (c, f(c))$, and let $y := mx + q$ be the straight line going through P_- and P_+. Since f is convex we have that

$$f(x) \le mx + q$$

for all x in the interval $(-c, c)$. By setting $x := Z$, we have that

$$e^{\lambda Z} \le mZ + q \qquad (4.7)$$

where

$$q = \frac{e^{\lambda c} + e^{-\lambda c}}{2}.$$

Taking expectations on both sides of Equation (4.7) we have

$$\mathtt{E}[e^{\lambda Z}] \le m\,\mathtt{E}[Z] + \mathtt{E}[q] = \frac{e^{\lambda c} + e^{-\lambda c}}{2} \le e^{(\lambda c)^2/2}$$

The last inequality follows from the fact that, for all x,

$$\frac{e^x + e^{-x}}{2} \leq e^{x^2/2}$$

which can be easily verified by taking the Taylor series expansion of both sides. ∎

We apply the lemma to the random variable

$$Z := (Y_n | \mathbf{X}_{n-1}).$$

Z satisfies the hypotheses of the lemma since

$$\mathtt{E}[Z] = \mathtt{E}[Y_n|\mathbf{X}_{n-1}] = \mathtt{E}[X_n - X_{n-1}|\mathbf{X}_{n-1}] = 0$$

by the martingale property, and

$$|Z| = |(Y_n|\mathbf{X}_{n-1})| \leq |X_n - X_{n-1}| \leq c_n$$

by the Bounded Difference Condition. Therefore

$$\mathtt{E}[e^{\lambda Y_n}|\mathbf{X}_{n-1}] \leq e^{\lambda^2 c_n^2/2}$$

which, after substituting into Equation (4.6), yields by induction

$$\begin{aligned} \mathtt{E}[e^{\lambda X_n}] &\leq \mathtt{E}[\exp(\lambda X_{n-1})] \cdot e^{\lambda^2 c_n^2/2} \\ &\leq \prod_{i=1}^{n} e^{\lambda^2 c_i^2/2} =: e^{\lambda^2 c/2} \end{aligned}$$

where

$$c := \sum_{i=1}^{n} c_i^2.$$

An elementary application of Calculus shows that the ratio $e^{\lambda^2 c/2}/e^{\lambda t}$ attains the minumum when $\lambda = t/c$. Therefore, substituting into Equation (4.5),

$$\begin{aligned} \Pr(X_n > t) &\leq \min_{\lambda > 0} \frac{\mathtt{E}[e^{\lambda X_n}]}{e^{\lambda t}} \\ &= \exp\left(-\frac{t^2}{2c}\right) \end{aligned}$$

which ends the proof of Theorem 4.8.

Exercise 4.10 *Derive a similar inequality for the case when the martingale condition is replaced by the following:*

$$|\mathtt{E}[Y_i \mid \mathbf{X}_{i-1}]| \leq m,$$

for some non–negative real number m.

Exercise 4.11 *Suppose the bounded differences condition is satisfied with probability at least* $1-\epsilon$ *for some* $\epsilon > 0$ *i.e.*

$$\mathtt{Pr}[\bigvee_{i\in[n]} |X_i - X_{i-1}| > c_i] \leq \epsilon.$$

Show that

$$\Pr(X_n > X_0 + t) \leq \exp\left(-\frac{t^2}{2c}\right) + \epsilon$$

and

$$\Pr(X_n < X_0 - t) \leq \exp\left(-\frac{t^2}{2c}\right) + \epsilon$$

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

4.3 Generalizing Martingales and Azuma's Inequality.

It is useful to generalize the definition of martingale to the case when the random variables under study might depend on another set of random variables.

Definition 4.12 *A sequence of random variables* $\mathbf{Y} := Y_0, Y_1, \ldots$ *is a* martingale with respect to another sequence $\mathbf{X} := X_0, X_1, \ldots$ *if for each* $i \geq 0$,

$$Y_i = g_i(\mathbf{X}_i),$$

for some function g_i *and*

$$\mathtt{E}[Y_i|\mathbf{X}_{i-1}] = Y_{i-1} \quad i \geq 1.$$

Example 4.13 Let us consider again example 4.4, where a gambler starts with an initial fortune of 0 and wagers a unit amount at repeated throws of a fair die. In the notation of that example, the sequence $S_0, S_1, \ldots$ is a martingale with respect to the sequence $0 =: X_0, X_1, X_2, \ldots$, where $S_n = \sum_{i\leq n} X_i$ for each $n \geq 0$. $\triangledown$

A very important example of this general definition of a martingale is provided by the following definition and lemma.

Definition 4.14 *The* **Doob sequence** *of a function f with respect to a sequence of random variables $X_1, \ldots, X_n$ is defined by*

$$Y_i := \mathtt{E}[f|\mathbf{X}_i], \quad 0 \le i \le n.$$

In particular, $Y_0 := \mathtt{E}[f]$ and $Y_n = f(X_1, \ldots, X_n)$.

Proposition 4.15 *The Doob sequence of a function defines a martingale. That is,*

$$\mathtt{E}[Y_i|\mathbf{X}_{i-1}] = Y_{i-1} \quad 0 \le i \le n.$$

The proof is an immediate consequence of (4.4).

Example 4.16 [Edge exposure martingale] An important special case of definition 4.14 occurs in the context of the *random graph* $G_{n,p}$. This is the graph with vertex set $[n]$ and each edge $\{i,j\}, i \ne j$ present with probability p independently of all other edges. Let $f : \binom{[n]}{2} \to \mathbb{R}$ be a function on the edge set of the complete graph K_n. For instance f could be the chromatic number or the size of the largest clique. Number the edges from 1 to $\binom{n}{2}$ in some arbitrary order and let $X_j := 1$ if the jth edge is present and 0 otherwise. The Doob sequence of f with respect to the variables $X_j, j \in [\binom{n}{2}]$ is called the **Edge exposure martingale**. Intuitively, we are exposing the edges one by one and observing the average value of f under this partial information. $\triangledown$

Azuma's inequality can be generalized to a sequence $\mathbf{Y}$ which is a martingale w.r.t. another sequence $\mathbf{X}$ of r.v.'s.

Theorem 4.17 (Azuma's Inequality– general version) *Let $Y_0, Y_1, \ldots$ be a martingale w.r.t. the sequence $X_0, X_1, \ldots$. Suppose also that the $\mathbf{Y}$ satisfies the Bounded Difference Condition with parameters $c_i, i \ge 1$. Then,*

$$\Pr(Y_n > Y_0 + t) \le \exp\left(-\frac{t^2}{2c}\right)$$

and

$$\Pr(Y_n < Y_0 - t) \le \exp\left(-\frac{t^2}{2c}\right)$$

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

Proof. The proof is almost identical to that of Theorem 4.8. Assume without loss of generality that $Y_0 := 0$ and define the martingale difference sequence $D_i := Y_i - Y_{i-1}, i \geq 1$. Then $Y_n = Y_{n-1} + D_n$. As before,

$$\Pr(Y_n > t) \leq \min_{\lambda > 0} \frac{\mathtt{E}[e^{\lambda Y_n}]}{e^{\lambda t}}.$$

Focus on the numerator $\mathtt{E}[e^{\lambda Y_n}]$.

$$\begin{aligned} \mathtt{E}[e^{\lambda Y_n}] &= \mathtt{E}[e^{\lambda(Y_{n-1}+D_n)}] \\ &= \mathtt{E}[\mathtt{E}[e^{\lambda(Y_{n-1}+D_n)} \mid \mathbf{X}_{n-1}]] \\ &= \mathtt{E}[e^{\lambda Y_{n-1}} \mathtt{E}[e^{\lambda D_n} | \mathbf{X}_{n-1}]]. \end{aligned}$$

The last line, the only place where the proof differs from that of Theorem 4.8, follows form (4.2) because $Y_{n-1} = g_{n-1}(\mathbf{X}_{n-1})$. The proof now proceeds identical to that of Theorem 4.8 provided Lemma 4.9 is invoked for the variables $Z := (D_n | \mathbf{X}_{n-1})$. The verification that Z satisfies the hypotheses of Lemma 4.9 is straightforward and is left as an exercise. ∎

4.4 The Method of Bounded Differences

We shall now see how to apply Azuma's Inequality to obtain a very powerful and useful generalization of the CH-bounds. The link is provided by the Doob martingale from which the following theorem emerges naturally.

Theorem 4.18 *[The Method of Bounded Martingale Differences]*
Let $X_1, \ldots, X_n$ be an arbitrary set of random variables and let f be a function satisfying the property that for each $i \in [n]$, there is a non–negative c_i such that

$$|\mathtt{E}[f | \mathbf{X}_i] - \mathtt{E}[f | \mathbf{X}_{i-1}]| \leq c_i. \tag{4.8}$$

Then,

$$\Pr[f > \mathtt{E}f + t] \leq \exp\left(-\frac{t^2}{2c}\right)$$

and

$$\Pr[f < \mathtt{E}f - t] \leq \exp\left(-\frac{t^2}{2c}\right)$$

where $c := \sum_{i \leq n} c_i^2$.

This theorem is just a restatement of Theorem 4.17 for the special case of the Doob sequence $Y_i := \mathtt{E}[f|\mathbf{X}_i], 0 \leq i \leq n$. *Notice that the X_i's are not assumed to be independent.*

Some weaker but often more convenient versions of this bound will now be deduced.

Definition 4.19 (Average Bounded Differences) *A function f satisfies the* Average Bounded Difference Condition *(henceforth* ABDC*) with parameters $c_i, i \in [n]$ with respect to the random variables $X_1, \ldots, X_n$ if for any a_i, a_i',*

$$|\mathtt{E}[f|\mathbf{X}_{i-1}, X_i = a_i] - \mathtt{E}[f|\mathbf{X}_{i-1}, X_i = a_i']| \leq c_i \tag{4.9}$$

for $1 \leq i \leq n$.

In words, the condition ABDC in (4.9) says: fix the first $i-1$ variables to some values, let the ith variable take two different values and set the remaining variables at random (according to the given distribution conditioned on the previous settings). Then the difference between the two corresponding partial averages of f must be bounded uniformly by c_i.

Corollary 4.20 (The Method of Bounded Average Differences) *Let f satisfy the* ABDC *condition with respect to the variables $X_1, \ldots, X_n$ with parameters $c_i, i \in [n]$. Then*

$$\Pr[f > \mathtt{E}f + t] \leq \exp\left(-\frac{t^2}{2c}\right)$$

and

$$\Pr[f < \mathtt{E}f - t] \leq \exp\left(-\frac{t^2}{2c}\right)$$

where $c := \sum_{i \leq n} c_i^2$.

Proof. We shall show that if (4.9) holds then so does (4.8). To see this, expand using total conditional probability:

$$\mathtt{E}[f \mid \mathbf{X}_{i-1}] = \sum_a \mathtt{E}[f \mid \mathbf{X}_{i-1}, X_i = a]\mathtt{Pr}[X_i = a \mid \mathbf{X}_{i-1}],$$

and write

$$\mathtt{E}[f \mid \mathbf{X}_i] = \sum_a \mathtt{E}[f \mid \mathbf{X}_i]\mathtt{Pr}[X_i = a \mid \mathbf{X}_{i-1}].$$

Hence,

$$\begin{aligned}
&|\mathtt{E}[f \mid \mathbf{X}_{i-1}] - \mathtt{E}[f \mid \mathbf{X}_i]| = \\
&\qquad |\sum_a (\mathtt{E}[f \mid \mathbf{X}_{i-1}, X_i = a] - \mathtt{E}[f \mid \mathbf{X}_i])\mathtt{Pr}[X_i = a \mid \mathbf{X}_{i-1}]| \\
&\leq \sum_a |\mathtt{E}[f \mid \mathbf{X}_{i-1}, X_i = a] - \mathtt{E}[f \mid \mathbf{X}_i]|\mathtt{Pr}[X_i = a \mid \mathbf{X}_{i-1}] \\
&\leq \sum_a c_i \cdot \mathtt{Pr}[X_i = a \mid \mathbf{X}_{i-1}] \\
&= c_i.
\end{aligned}$$

∎

A further significant simplification obtains from the following definition.

Definition 4.21 *A function $f(x_1, \ldots, x_n)$ satisfies the* **Lipshitz property** *or the* **Bounded Differences Condition** (BDC) *with constants $d_i, i \in [n]$ if*

$$|f(\boldsymbol{a}) - f(\boldsymbol{a}^0)| \leq d_i, \tag{4.10}$$

whenever $\boldsymbol{a}$ and $\boldsymbol{a}^0$ differ in just the i-th coordinate, $i \in [n]$.

In words, the condition BDC says: the difference between the values of f on two inputs that differ in only the ith co–ordinate is bounded uniformly by d_i. This is exactly like the usual Lipschitz condition in the setting where the underlying metric is the Hamming distance.

Corollary 4.22 (Method of Bounded Differences) *If f satisfies the Lipshitz property with constants $d_i, i \in [n]$ and $X_1, \ldots, X_n$ are independent random variables, then,*

$$\Pr[f > \mathtt{E}f + t] \leq \exp\left(-\frac{t^2}{2d}\right)$$

and

$$\Pr[f < \mathtt{E}f - t] \leq \exp\left(-\frac{t^2}{2d}\right)$$

where $d := \sum_{i \leq n} d_i^2$.

Proof. For typographical convenience let $\mathbf{X}^{i+1}$ be shorthand notation for the sequence $X_{i+1}, \ldots, X_n$. And let $\mathbf{X}^{i+1} = \mathbf{a}^{i+1}$ denote the componentwise equality for the two sequences. We show that if f satisfies the Lipschitz condition with parameters $c_i, i \in [n]$, then (4.9) holds. To see this, expand using total conditional probability to get

$$\begin{aligned}
&\mathtt{E}[f \mid \mathbf{X}_{i-1}, X_i = a] = \\
&\qquad \sum_{a_{i+1},\ldots,a_n} \mathtt{E}[f \mid \mathbf{X}_{i-1}, X_i = a, \mathbf{X}^{i+1} = \mathbf{a}^{i+1}]\mathtt{Pr}[\mathbf{X}^{i+1} = \mathbf{a}^{i+1} \mid \mathbf{X}_{i-1}, X_i = a] \\
&= \sum_{a_{i+1},\ldots,a_n} \mathtt{E}[f \mid \mathbf{X}_{i-1}, X_i = a, \mathbf{X}^{i+1} = \mathbf{a}^{i+1}]\mathtt{Pr}[\mathbf{X}^{i+1} = \mathbf{a}^{i+1}], \text{by independence,} \\
&= \sum_{a_{i+1},\ldots,a_n} f(\mathbf{X}_{i-1}, a, a_{i+1}, \ldots, a_n)\mathtt{Pr}[\mathbf{X}^{i+1} = \mathbf{a}^{i+1}].
\end{aligned}$$

Put $a := a_i, a'_i$ successively and take the difference. Then,

$$\begin{aligned}
&|\mathtt{E}[f \mid \mathbf{X}_{i-1}, X_i = a_i] - \mathtt{E}[f \mid \mathbf{X}_{i-1}, X_i = a'_i]| = \\
&\qquad |\sum_{a_{i+1},\ldots,a_n} f(\mathbf{X}_{i-1}, a_i, \mathbf{a}^{i+1}) - f(\mathbf{X}_{i-1}, a'_i, \mathbf{a}^{i+1})\mathtt{Pr}[\mathbf{X}^{i+1} = \mathbf{a}^{i+1}]| \\
&\leq \sum_{a_{i+1},\ldots,a_n} |f(\mathbf{X}_{i-1}, a_i, a_{i+1}, \ldots, a_n) - f(\mathbf{X}_{i-1}, a'_i, a_{i+1}, \ldots, a_n)|\mathtt{Pr}[\mathbf{X}^{i+1} = \mathbf{a}^{i+1}] \\
&\leq \sum_{a_{i+1},\ldots,a_n} c_i \cdot \mathtt{Pr}[\mathbf{X}^{i+1} = \mathbf{a}^{i+1}], \text{by the Lipschitz property,} \\
&= c_i.
\end{aligned}$$

■

Some comments are in order about the three different versions of the "Method of Bounded Differences".

Corollary 4.22 is usually referred to in the literature as *the* Method of Bounded Differences. This is because it is the most convenient one to apply. The BDC condition is very attractive and easy to check. It also makes the result intuitive: if f does not depend on any one argument, then it is not likely to be far from its expectation when the inputs are set at random. However, there are two drawbacks: first the variables $X_1, \ldots, X_n$ must be independent. Second, the parameters d_i in the BDC condition might be too large and consequently the bound might turn out too weak to be useful.

It might be the case that the BDC condition holds for f with small parameters d_i except for a small set of exceptional instances. In that case, it is unfair to "penalise" f with the "worst–case" larger parameters strictly demanded by the BDC condition. Rather, one should take an average, and this is the purpose of the ABDC condition. The parameters c_i required for the BDC condition are always bounded by the parameters d_i required for the BDC condition, and often $c_i \ll d_i$. In the latter case, the bound obtained from Corollary 4.20, The Method Of Average Bounded Differences will be significantly better than that from Corollary 4.22, the Method of Bounded Differences.

Theorem 4.18 is the most powerful version of the method: the parameters required for the martingale differences condition are always bounded by the parameters required by the ABDC condition, and hence the probability bound is always stronger.

The price to be paid however, is that both the martingale differences condition and the ABDC condition can be quite difficult to check for an arbitrary f compared with the simple BDC condition. If f can be decomposed as a sum, linearity of expectation can be used to simplify the computation as we shall demonstrate on some examples in the next chapter.

Note crucially, that in both Theorem 4.18 and in Corollay 4.20, *the variables are not required to be independent.* This greatly increases the scope of their applicability as we shall demonstrate in several examples in the next chapter.

We now develop familiarity with these tools by applying them to several different situations in the next chapter.

4.5 Bibliographic Notes

Martingales ae a classic subject treated in many standard texts on Probability such as Grimmett and Stirzaker [7][Ch.7,12]. The Method of Bounded Differences and its applications to problems of combinatorics and discrete mathematics is covered in a well–known survey of the same name by C. McDiarmid [10]. Both these are couched in measure–theoretic terminology. A more elementary account can be found in Alon and Spencer [1].

4.6 Problems

Problem 4.1 *Let $X_0, X_1, \ldots$ be random variables such that the partial sums $S_n := X_1 + \ldots + X_n$ determine a martingale with respect to $\mathbf{X}$. Show that $\mathtt{E}[X_i X_j] = 0$ if $i \neq j$.*

Problem 4.2 (Sampling without replacement). *Consider an urn containing N balls out of which M are red. Balls are drawn without replacement. Show that the sequence of random variables denoting the fraction of red balls remaining in the urn defines a martingale. Derive a concentration result.*

Problem 4.3 *Let $X_0, X_1, \dots$ be a sequence of random variables with finite means satisfying*

$$\mathtt{E}[X_{n+1} \mid X_0, \dots, X_n] = aX_n + bX_{n-1}, \quad n \geq 1$$

where $0 < a, b < 1$ and $a + b = 1$. Find a value of α for which $S_n := \alpha X_n + X_{n-1}$ determines a martingale with respect to $\mathbf{X}$.

We shall generalise the definition of a martingale even further to be able to define the so–called vertex exposure martingale in a random graph.

Definition 4.23 *A sequence $\mathbf{Y} := Y_0, Y_1, \dots$ is a martingale with respect to a sequence $\mathbf{X} := X_0, X_1, \dots$ if there is an increasing sequence $0 \leq k_0 \leq k_1 \leq \dots$ such that $Y_i = g_i(\mathbf{X}_{k_i}), i \geq 0$ for some function g_i and $\mathtt{E}[Y_i \mid \mathbf{X}_{k_{i-1}}] = Y_{i-1}$.*

Problem 4.4 (Vertex Exposure Martingale) *Use Definition 4.23 to define a martingale in the random graph $G_{n,p}$ corresponding to revealing the edges in n stages where at the ith stage we reveal all edges incident on the first i vertices.*

Problem 4.5 (Azuma generalised further, I) *Show that Azuma's inequality can be generalised to apply to the Definition 4.23 of a martingale.*

Problem 4.6 (Onward from Azuma) *Prove the following generalisation of Azuma's inequality. Let $X_0, X_1, \dots, X_n, \dots$ be a martingale with*

$$|X_i - X_{i-1}| \leq W_i, \quad , i \geq 1,$$

where $W_1, \dots, W_n$ are random variables such that for some constant $c > 0$, and some $\delta = \delta(n)$,

$$\mathtt{Pr}[W_1^2 + \cdots + W_n^2 > c] \leq \delta.$$

Then for any $t > 0$,

$$\mathtt{Pr}[|X_n - X_0| > t] \leq 2 \exp\left(\frac{-t^2}{2c}\right) + \delta.$$

Chapter 5

Martingales and the Method of Bounded Differences: Applications

In this chapter we shall see the "Method of Bounded Differences" in action by applying it to various examples. We shall see that in some cases, it suffices to apply the method in the simplest form whereas in others, the more powerful version is required to get meaningful bounds.

5.1 Chernoff–Hoeffding Revisited

Let $X_1, \ldots, X_n$ be independent variables taking values in $[0, 1]$, and consider $f(x_1, \ldots, x_n) := \sum_i x_i$. Then of course f has the Lipshitz property with each $d_i = 1$ in (4.10) and we get for $X := X_1 + \cdots + X_n$, the bound:

$$\mathtt{Pr}[|X - \mathtt{E}[X]| > t] \leq 2e^{\frac{-t^2}{2n}},$$

which is only off by a factor of 4 in the exponent from the Chernoff–Hoeffding bound.

Exercise 5.1 *Remove the factor of* 4 *by applying the method of bounded martingale differences.*

5.2 Stochastic Optimization: Bin Packing

The bin packing problem is a well–studied combinatorial optimization problem: we are given n items of sizes in the interval $[0, 1]$ and are required to pack them into the fewest number of unit–capacity bins as possible. In the stochastic version, the item sizes are independent random variables in the interval $[0, 1]$. Let $B_n = B_n(x_1, \ldots, x_n)$ denote the optimum value, namely the minimum number of bins that suffice. Then clearly the Lipshitz condition holds with constant 1 and we get the concentration result:

$$\mathtt{Pr}[|B_n - \mathtt{E}[B_n]| > t] \leq 2e^{\frac{-t^2}{2n}}.$$

Is this hard to show? Citation needed

It can be shown that $\mathtt{E}[B_n] = \beta n$ for some constant $\beta > 0$, hence we deduce that $\mathtt{Pr}[|B_n - \mathtt{E}[B_n]| > \epsilon \mathtt{E}[B_n]]$ decreases exponentially in n. This straighforward application of the martingale technique vastly improved previous results on this problem.

Exercise 5.2 *Let B_n^{FF} denote the number of bins that would be needed if one applied the* first–fit *heuristic. Give a sharp concentration result on B_n^{FF}. (The first-fit heuristic places the items one by one, with the current item being placed in the first available bin.)*

5.3 Balls and Bins

In the classical balls and bins experiment, m balls are thrown independently at random into n bins (usually $m \geq n$) and we are interested in various statistics of the experiment, for instance, the number of empty bins. Let $Z_i, i \in [n]$ denote the indicator variables

$$Z_i := \begin{cases} 1, & \text{if bin } i \text{ is empty,} \\ 0, & \text{otherwise.} \end{cases}$$

Then, the variable we are interested in is the sum $Z := \sum_i Z_i$.

Exercise 5.3 *Show that $\mu := \mathtt{E}[Z] = n\left(1 - \frac{1}{n}\right)^m \approx ne^{-m/n}$.*

Exercise 5.4 *Show that the Z_i's are not independent.*

In view of Exercise 5.4 we cannot apply the Chernoff bounds. In order to get a sharp concentration result, we can use the method of bounded differences in simple form. Consider Z as a function $Z(X_1, \ldots, X_m)$ where, for $k \in [m]$, X_k is a random variable taking values in the set $[n]$ and indicating the bin that ball k lands in.

Let's check that the function Z satisfies the Lipschitz condition with constant 1. Denoting by b_k the bin in which the k-th balls falls into, the condition

$$|Z(b_1, \ldots, b_{i-1}, b_i, b_{i+1}, \ldots, b_m) - Z(b_1, \ldots, b_{i-1}, \hat{b}_i, b_{i+1}, \ldots, b_m)| \leq 1$$

simply says that if the i-th ball is moved from one bin to another, keeping all other balls where they are, the number of empty bins can at most either go up by one or down by one. Hence, we have the bound:

$$\texttt{Pr}[|Z - \texttt{E}[Z]| > t] \leq 2 \exp\left(\frac{-t^2}{2m}\right). \tag{5.1}$$

A better bound can be obtained by applying the method of bounded average differences. Now we need to compute

$$c_k := |\texttt{E}[Z \mid \mathbf{X}_{k-1}, X_k = b_k] - \texttt{E}[Z \mid \mathbf{X}_{k-1}, X_k = b'_k]|$$

for $k \in [m]$. By linearity of expectation, this reduces to computing for each $i \in [n]$, $c_{i,k} := |\texttt{E}[Z_i \mid \mathbf{X}_{k-1}, X_k = b_k] - \texttt{E}[Z_i \mid \mathbf{X}_{k-1}, X_k = b'_k]|$.

Let us therefore consider for each $i \in [n]$,and for some fixed set of bins $b_1, \ldots, b_k, b'_k$ $(b_k \neq b'_k)$,

$$c_{i,k} = |\texttt{E}[Z_i \mid \mathbf{X}_{k-1} = \mathbf{b}_{k-1}, X_k = b_k] - \texttt{E}[Z_i \mid \mathbf{X}_{k-1} = \mathbf{b}_{k-1}, X_k = b'_k]|.$$

Let $S := \{b_1, \ldots, b_{k-1}\}$.

- If $i \in S$, then of course,

  $$\texttt{E}[Z_i \mid \mathbf{X}_{k-1} = \mathbf{b}_{k-1}, X_k = b_k] = 0 = \texttt{E}[Z_i \mid \mathbf{X}_{k-1} = \mathbf{b}_{k-1}, X_k = b'_k].$$

  and so $c_{i,k} = 0$.

- If $i \notin S$ and $i \neq b$, then

  $$\texttt{E}[Z_i \mid \mathbf{X}_{k-1} = \mathbf{b}_{k-1}, X_k = b] = (1 - 1/n)^{m-k},$$

  Hence, for $i \notin S \cup \{b_k, b'_k\}$, we have $c_{i,k} = 0$.

- Finally, if $i = b_k \notin S$, , then of course $\texttt{E}[Z_i \mid \mathbf{X}_{k-1} = \mathbf{b}_{k-1}, X_k = b_k] = 0$ but if $b'_k \notin S$

  $$\texttt{E}[Z_i \mid \mathbf{X}_{k-1} = \mathbf{b}_{k-1}, X_k = b'_k] = (1 - 1/n)^{m-k}.$$

  Hence $c_{i,k} = (1 - 1/n)^{m-k}$ in this case.

Overall, we see that $c_k = \sum_i c_{i,k} \leq (1 - 1/n)^{m-k}$ and

$$\sum_k c_k^2 \leq \frac{1 - (1 - 1/n)^{2m}}{1 - (1 - 1/n)^2} = \frac{n^2 - \mu^2}{2n - 1}.$$

This gives the bound:

$$\texttt{Pr}[|Z - \texttt{E}[Z]| > t] \leq 2 \exp\left(-\frac{t^2(n - 1/2)}{n^2 - \mu^2}\right).$$

How does it compare to the previous bound?

Asymptotically in terms of $r := m/n$, this is

$$2 \exp\left(-\frac{t^2}{n(1 - e^{-2r})}\right).$$

5.4 Stochastic Optimization: TSP

A travelling salesman is required to visit n towns and must choose the shortest route to do so. This is a notoriously difficult combinatorial optomization problem. A stochastic version in two dimensions asks for the shortest route when the points $P_i := (X_i, Y_i), i \in [n]$ are chosen uniformly and independently in the unit square, $[0,1]^2$ (i.e. each X_i and Y_i is distributed uniformly and independently in $[0,1]$).

By whom is the result? And, who is celebrating?

Let $T_n = T_n(P_i, i \in [n])$ denote the length of the optimal tour. A celebrated result shows that $\texttt{E}[T_n] = \beta\sqrt{n}$ for some $\beta > 0$. What about a sharp concentration result? A straightforward approach is to observe that T_n has the Lipschitz property with constant at most $2\sqrt{2}$ (imagine that all except one point are in one corner and the last is in the opposite corner). Hence, we have the bound

Is the computation correct? Shouldn't be the denominator be 8 ?

$$\texttt{Pr}[|T_n - \texttt{E}[T_n]| > t] \leq \exp\left(\frac{-t^2}{2\sqrt{2}n}\right). \tag{5.2}$$

Note that since $\texttt{E}[T_n] = \beta\sqrt{n}$, this bound is no good for small deviations around the mean i.e. for $t = \epsilon\texttt{E}[T_n]$.

For a better bound, we shall turn to the method of bounded martingale differences. Let $T_n(i)$ denote the length of the shortest tour through all points except the ith for $i \in [n]$.

Now we observe the crucial inequality that

$$T_n(i) \leq T_n \leq T_n(i) + 2Z_i, \quad i < n, \tag{5.3}$$

where Z_i is the shortest distance from point P_i to one of the points P_{i+1} through P_n. The first inequality follows because, denoting the neighbours of P_i in T_n by

P and Q, the tour obtained by joining P and Q directly excludes P_i and, by the triangle inequality, has length less than T_n. For the second inequality, suppose $P_j, j > i$ is the closest point to P_i. Now take an optimal tour of all points except P_i and convert it into a tour including P_i by visiting P_i after reaching P_j and returning to P_j. This is not a tour but can be converted into one by taking a short–cut to the next point after P_j. The length of the resulting tour is no more than $T_n(i) + 2Z_i$ by the triangle inequality.

Taking conditional expectations in (5.3), we get:

$$\mathtt{E}[T_n(i) \mid \mathbf{P}_{i-1}] \leq \mathtt{E}[T_n \mid \mathbf{P}_{i-1}] \leq \mathtt{E}[T_n(i) \mid \mathbf{P}_{i-1}] + 2\mathtt{E}[Z_i \mid \mathbf{P}_{i-1}],$$

$$\mathtt{E}[T_n(i) \mid \mathbf{P}_i] \leq \mathtt{E}[T_n \mid \mathbf{P}_i] \leq \mathtt{E}[T_n(i) \mid \mathbf{P}_i] + 2\mathtt{E}[Z_i \mid \mathbf{P}_i].$$

Note that $\mathtt{E}[T_n(i) \mid \mathbf{P}_i] = \mathtt{E}[T_n(i) \mid \mathbf{P}_{i-1}]$. Hence, we conclude,

$$|\mathtt{E}[T_n \mid \mathbf{P}_i] - \mathtt{E}[T_n \mid \mathbf{P}_{i-1}]| \leq 2\max(\mathtt{E}[Z_i \mid \mathbf{P}_{i-1}], \mathtt{E}[Z_i \mid \mathbf{P}_i]), \quad i \leq n.$$

Computing $\mathtt{E}[Z_i \mid \mathbf{P}_i]$ is the following question: given a point Q in $[0,1]$, what is its shortest distance to one of a randomly chosen set of $n-i$ points? Computing $\mathtt{E}[Z_i \mid \mathbf{P}_{i-1}]$ is the same, except the point Q is also picked at random. This exercise is relegated to Problem 5.4. The answer is that $\mathtt{E}[Z_i \mid \mathbf{P}_i], \mathtt{E}[Z_i \mid \mathbf{P}_{i-1}] \leq c/\sqrt{n-i}$ for some constant $c > 0$. Finally, taking the trivial bound $|\mathtt{E}[T_n \mid \mathbf{P}_n] - \mathtt{E}[T_n \mid \mathbf{P}_{n-1}]| \leq 2\sqrt{2}$ we get

$$\begin{aligned} \mathtt{Pr}[|T_n - \mathtt{E}[T_n]| > t) &\leq 2\exp\left(\frac{-t^2}{2(8 + \sum_{i<n} 4c^2/(n-i))}\right) \\ &\leq 2\exp\left(\frac{-at^2}{\log n}\right), \end{aligned} \tag{5.4}$$

for some $a > 0$. Compare (5.4) to (5.2); in particular, note that the former together with $\mathtt{E}[T_n] = \beta\sqrt{n}$ yields

$$\mathtt{Pr}[|T_n - \beta\sqrt{n}| > \epsilon\sqrt{n}] \leq 2\exp\left(\frac{-b\epsilon^2 n}{\log n}\right),$$

for some $b > 0$ and all $\epsilon > 0$.

We shall see later that this bound can be further improved by removing the $\log n$ factor. But that will need a new method!

5.5 Isoperimetric Inequalities

The new method has to do with a certain discrete isoperimetric inequality that we will discuss later. The application of martingales here will prepare you for the "things to come".

The setting for a discrete isoperimetric inequality is a space Ω equipped with both a measure μ and a metric d. In many natural spaces of this kind, concentration of measure obtains in the following form: for any large set (in the sense of μ), almost all the measure is concentrated in a small neighbourhood (in the sense of d) around it. More precisely, for $t > 0$, define the t–neighbourhood A_t of a subset $A \subseteq \Omega$ by

$$A_t := \{x \mid d(x, A) \leq t\},$$

where $d(x, A) := \inf_{y \in A} d(x, y)$. Define the function $\alpha(t)$ so that whenever $\mu(A) \geq 1/2$, then $\mu(A_t) \geq 1 - \alpha(t)$. Then if α decreases sufficiently rapidly (exponentially, for instance), we will get a sharp concentration of measure around A.

Such an inequality has immediate applications to show that a real–valued function f on Ω that is well behaved will be "almost constant". Precisely, let f be Lipshitz, i.e.

$$|f(x) - f(y)| \leq d(x, y), \quad x, y \in \Omega.$$

Let $\mathtt{M}[f]$ denote a *median* of f i.e.

$$\mathtt{Pr}[f(x) \leq \mathtt{M}[f]] \geq 1/2 \quad \text{and} \quad \mathtt{Pr}[f(x) \geq \mathtt{M}[f]] \geq 1/2.$$

Set $A := \{x \mid f(x) \leq \mathtt{M}[f]\}$ and $B := \{x \mid f(x) \geq \mathtt{M}[f]\}$. Note that by the Lipschitz property of f, $f(x) \leq \mathtt{M}[f] + t$ for $x \in A_t$ and $f(x) \geq \mathtt{M}[f] - t$ for $x \in B_t$. Hence $|f(x) - \mathtt{M}[f]| \leq t$ for $x \in A_t \cap B_t$ and we conclude that

$$\mathtt{Pr}[|f(x) - \mathtt{M}[f]| \leq t] \geq 1 - 2\alpha(t).$$

Thus we get a sharp concentration result for f around its median value. In Problem 5.3, you are asked to show that if α decreases exponentially, this also implies a sharp concentration around the mean.

As a concrete illustration of this abstract framework, consider the space $\{0, 1\}^n$ with the uniform measure (which is also the product measure with $p = 1/2$ in each co–ordinate) and the Hamming metric, d_H. Let A be a subset of size at least 2^{n-1} so that $\mu(A) \geq 1/2$. Consider the function $f(x) := d_H(x, A)$, the Hamming distance of x to A. Surely f is Lipshitz. Let $X_1, \ldots, X_n$ be independent and uniformly distributed in $\{0, 1\}$. Then, by applying the method of bounded differences,

$$\mu[f > \mathtt{E}[f] + t], \mu[f < \mathtt{E}[f] - t] \leq e^{\frac{-t^2}{2n}}.$$

In particular,

$$\begin{aligned} 1/2 &\leq \mu(A) \\ &= \mu(f = 0) \\ &\leq \mu(f < \mathtt{E}[f] - \mathtt{E}[f]) \\ &\leq e^{\frac{-\mathtt{E}[f]^2}{2n}}. \end{aligned}$$

Thus $\mathtt{E}[f] \leq t_0 := \sqrt{2 \ln 2n}$. Finally then,

$$\mu(A_t) \geq 1 - \exp\left(\frac{-(t - t_0)^2}{2n}\right).$$

In Problem 5.5, you are asked to deduce a cleaner form by an application of the Chernoff bound.

Exercise 5.5 *Give a similar result for a general subset A with $\mu(A) = \alpha < 1$.*

5.6 Distributed Edge Colouring

In this section, we consider algorithms for the problem of edge–colouring a graph. Apart from its intrinsic interest as a classical combinatorial problem, edge colouring is often useful because of its connection to scheduling. Here we will discuss a distributed edge colouring algorithm that allows a distributed network to compute an edge colouring of its own (unknown) topology. In distributed networks or architectures this might be useful, for a matching often corresponds to a set of data transfers that can be executed simultaneously. So, a partition of the edges into a small number of matchings– i.e. a "good" edge colouring– gives an efficient schedule to perform data transfers (for more details, see [11, 6]).

Vizing's Theorem shows that every graph G can be edge coloured in polynomial time with Δ or $\Delta + 1$ colours, where Δ is the maximum degree of the input graph (see, for instance, [2]). It is a challenging open problem whether colourings as good as these can be computed fast in parallel. In absence of such a result one might aim at the more modest goal of computing reasonably good colourings, instead of optimal ones. By a trivial modification of a well-known *vertex* colouring algorithm of Luby it is possible to edge colour a graph using $2\Delta - 2$ colours [9]. In this section we shall present and analize a simple algorithm that computes an edge colouring of the input graph using at most $1.6\Delta + 32 \log^{1+\epsilon} n$ colours, where $\epsilon > 0$ is arbitrary, but fixed. We shall keep the discussion informal except for the point which is relevant to us in this context. Namely, the high probability analyis of a subroutine to edge colour bipartite graphs. The details can be found in [11].

The algorithm is based on the following idea.

- Each vertex, in parallel, flips fair a coin and enters a set T, if the outcome is heads, or B, if the outcome is tails (T and B stand for top and bottom respectively).
- The bipartite graph $G[T, B]$ induced by the edges connecting top and bottom vertices is coloured by invoking the procedure BIPARTITE-COLORING described below.
- The algorithm is invoked recursively in parallel on $G[T]$ and $G[B]$, the graphs respectively induced by the top and bottom vertices. Both graphs are coloured useing the same set of colours.

The correctness of the last step follows by observing that $G[T]$ and $G[B]$ are vertex disjoint.

By using the techniques developed in Chapter 1 it is easy to see that, with high probability, the maximum degrees of $G[T, B]$, $G[T]$ and $G[b]$ are $(1 + o(1))\Delta/2$, provided that Δ is high enough. Here "high enough" means that it should be $\Delta \geq \log^{1+c} n$, for an arbitrary, but fixed, $c > 0$. When this condition is not satisfied any longer the algorithm switches to a mopping up phase by invoking the algorithm of Luby mentioned earlier. If initially $\Delta \gg \log n$, using twice as many colours as the max degree at the end of the recursion will not affect the quality of the solution too much.

The procedure BIPARTITE-COLORING exploits the fact that each vertex knows whether it is top or bottom. The procedure executes repeatedly the following round as long as the degree of the graph is high enough. (As before, "high enough" means $\Delta \gg \log n$.) At the beginning of the round each bottom vertex is given a set of Δ fresh new colours (i.e. colours not used in previous rounds). At the end of the round some of the edges will have their final color and will be removed from the bipartite graph. The leftover graph is passed on as the input to the next round.

ONE ROUND OF BIPARTITE-COLORING

- Each bottom vertex, in parallel, makes a *proposal* independently of other bottom vertices by assigning a random permutation of the colours to their incident edges.

- Each top vertex, in parallel, then picks a *winner* out of every set of incident edges that have the same colour. Tentative colours of winner edges become final.
- The *losers*– edges who are not winners– are decoloured and passed to the next round.

Similar to the outer recursive scheme, BIPARTITE-COLORING switches to Luby's algorithm when the condition $\Delta \gg \log n$ is no longer satisfied.

It is apparent that the overall algorithm is *distributed*. That is to say, each vertex need only exchange information with the neighbours to execute the algorithm. This and its simplicity make the algorithm amenable for implementations in a distributed environment.

Our aim now is to analize one round of BIPARTITE-COLORING. Let Δ denote the maximum degree of the graph at the beginning of the round and Δ' denote the maximum degree of the leftover graph. First, one can show the following.

Exercise 5.6 *Show that, for both top and bottom vertices, $\mathtt{E}[\Delta'] \approx \Delta/e$, where e is the Euler number.*

We want to show that, if $\Delta \gg \log n$, then the probability that $\Delta' > (1+\epsilon)\mathtt{E}[\Delta']$ is inverse polynomial, for any fixed $\epsilon > 0$.

Given this, the total number of colours used by BIPARTITE-COLORING is, with high probability,

$$\begin{aligned} \text{BC}(\Delta) &= \Delta + \frac{(1+\epsilon)\Delta}{e} + \frac{(1+\epsilon)^2\Delta^2}{e} + \dots \\ &\le \frac{1}{1-(1+\epsilon)e} \approx 1.59, \text{ for small enough } \epsilon. \end{aligned}$$

To this, one should add $O(\log n)$ colours at the end of the recursion. As it can be seen by analizing the simple recursion describing the number of colours used by the outer level of the recursion, the overall numbers of colours is the same $1.59\Delta + O(\log n)$, [11].

We now switch to the high probability analysis. The analysis which is published in the literature is extremely complicated and uses a certain *ad hoc* extension of the Chernoff–Hoeffding bounds [11]. The ease with which the algorithm can be analyses with the method of bounded average differences, as shown below, testifies to its power.

The analysis is particularly easy when v is a top vertex. For, in this case, the incident edges all receive colours independently of each other. This is exactly the situation of the balls and bins experiment from § 5.3: the incident edges are the "balls" that are falling at random independently into the colours that represent the "bins". One can apply the method of bounded differences in the simplest form. Let $T_e, e \in E$, be the random variables taking values in $[\Delta]$ that represent the tentative colours of the edges. Then the number of edges successfully coloured around v is a function $f(T_e, e \in N^1(v))$, where $N^1(v)$ denotes the set of edges incident on v.

Exercise 5.7 *Show that this function has the Lipschitz property with constant* 1.

Moreover, the variables $T_e, e \in N^1(v)$ are independent when v is a "top" vertex. Hence, by the method of bounded differences,

$$\mathtt{Pr}[|f - \mathtt{E}[f]| > t] \leq \exp\left(\frac{-t^2}{2\Delta}\right),$$

which for $t := \epsilon\Delta$ $(0 < \epsilon < 1)$ gives an exponentially decreasing probability for deviations around the mean. If $\Delta \gg \log n$ then the probability that the new degree of any vertex deviates far from its expected value is inverse polynomial, i.e. the new max degree is sharply concentrated around its mean.

The analysis for the "bottom" vertices is more complicated in several respects. It is useful to see why so that one can appreciate the need for using a more sophisticated tool such as the MOBD in average form. To start with, one could introduce an indicator random variable X_e for each edge e incident upon a bottom vertex v. These random variable are not independent however. Consider a four cycle with vertices v, a, w, b, where v and w are bottom vertices and a and b are top vertices. Let's refer to th eprocess of selecting the winner (step 2 of the bipartite colouring algorithm) as "the lottery". Suppose that we are given the information that edge va got tentative colour red and lost the lottery— i.e. $X_{va} = 0$— and that edge vb got tentative colour green. We'll argue intuitively that given this, it is more likely that $X_{vb} = 0$. Since edge va lost the lottery, the probability that edge wa gets tentative colour red increases. In turn, this increases the probability that edge wb gets tentative colour green, which implies that edge vb is more likely to lose the lottery. So, not only are the X_e's not independent, but the dependency among them is nefarious.

One could hope to bound this effect by using the MOBD in it simplest form. This is also ruled out however, for two reasons. The first is that the tentative colour choices of the edges around a vertex are not independent. This follows from the fact that edges are assigned a permutation of the colours. Or, put in another way, each edge is given a colour at random, but colours are drawn without replacement. The second reason is that the new degree of a bottom vertex v is a function f which might depend on as many as $\Delta(\Delta - 1) = \Theta(\Delta^2)$ edges. Even if f is Lipshitz with constants $d_i = 1$, this is not enough to get a strong enough bound because $d = \sum_i d_i^2 = \Theta(\Delta^2)$. Applying the method of bounded difference in simple form (Corollary 4.22) would give the bound

$$\mathtt{Pr}[|f - \mathtt{E}[f]| > t] \leq 2\exp\left(-\frac{t^2}{\Theta(\Delta^2)}\right).$$

This bound however is useless for $t = \epsilon\mathtt{E}[f]$ since $\mathtt{E}[f] \approx \Delta/e$.

We will use the method of bounded average differences to get a much better bound. We shall invoke the two crucial features of this more general method. Namely that it does not presume that the underlying variables are independent, and that, as we shall see, it allows us to bound the effect of individual random choices with constants much smaller than those given by the MOBD in simple form.

Let's now move on to the analysis. In what follows, we shall focus on a generic bottom vertex v. Let $N^1(v)$ denote the set of "direct" edges– i.e. the edges incident on v– and let $N^2(v)$ denote the set of "indirect edges" that is, the edges incident on a neighbour of v. Let $N(v) := N^1(v) \bigcup N^2(v)$. The number of edges successfully coloured at vertex v is a function $f(T_e, e \in N(v))$. Let us number the variables so that the direct edges are numbered *after* the indirect edges (this will be important for the calculations to follow). We need to compute

$$a_k := |\mathtt{E}[f \mid \mathbf{T}_{k-1}, T_k = c_k] - \mathtt{E}[f \mid \mathbf{T}_{k-1}, T_k = c'_k]|.$$

We decompose f as a sum to ease the computations later. Introduce the indicator variables $X_e, e \in E$:

$$X_e := \begin{cases} 1; & \text{if edge } e \text{ is successfully coloured,} \\ 0; & \text{otherwise.} \end{cases}$$

Then $f = \sum_{v \in e} X_e$.

Hence we are reduced, by linearity of expectation, to computing for each $e \in N^1(v)$,

$$|\mathtt{Pr}[X_e = 1 \mid \mathbf{T}_{k-1}, T_k = c_k] - \mathtt{Pr}[X_e = 1 \mid \mathbf{T}_{k-1}, T_k = c'_k]|.$$

For the computations that follows we should keep in mind that bottom vertices assign colours independently of each other. This implies that the colour choices of the edges incident upon a neighbour of v are independent of each other.

Let $e = vw$ be an edge incident on the bottom vertex v.

- First, let us consider the case when $e_1, \ldots, e_k \in N^2(v)$, i.e. only the choices of indirect edges are exposed. Let $e_k = (w, x)$. Notice that e and e_k are incident on the same top vertex w. Then it is clear that for our direct edge $e = vw$, $\texttt{Pr}[X_e = 1 \mid \mathbf{T}_{k-1}, T_k = c_k] = \texttt{Pr}[X_e = 1 \mid \mathbf{T}_{k-1}, T_k = c'_k]$. So we only need to compute

$$|\texttt{Pr}[X_{v,w} = 1 \mid \mathbf{T}_{k-1}, T_k = c_k] - \texttt{Pr}[X_{v,w} = 1 \mid \mathbf{T}_{k-1}, T_k = c'_k]|.$$

Now,

$$\begin{aligned}
&\texttt{Pr}[X_{v,w} = 1 \mid \mathbf{T}_{k-1}, T_k = c^*] = \\
&\qquad \sum_c \texttt{Pr}[X_{v,w} = 1 \mid \mathbf{T}_{k-1}, T_k = c^*, T_{v,w} = c]\texttt{Pr}[T_{v,w} = c \mid \mathbf{T}_{k-1}, T_k = c^*]. \\
&= \sum_c \texttt{Pr}[X_{v,w} = 1 \mid \mathbf{T}_{k-1}, T_k = c^*, T_{v,w} = c]\texttt{Pr}[T_{v,w} = c], \\
&:= \sum_c p(c, c^*)\texttt{Pr}[T_{v,w} = c]
\end{aligned}$$

where $p(c, c^*) := \texttt{Pr}[X_{v,w} = 1 \mid \mathbf{T}_{k-1}, T_k = c^*, T_{v,w} = c]$ and the penultimate line follows since $T_{v,w}$ is independent of $\mathbf{T}_k$. Now, if $c \neq c_k, c'_k$ then, recalling that colour choices of edges incident on top vertices are independent, $p(c, c_k) = p(c, c'_k)$. Hence,

$$\begin{aligned}
&|\texttt{Pr}[X_{v,w} = 1 \mid \mathbf{T}_{k-1}, T_k = c_k] - \texttt{Pr}[X_{v,w} = 1 \mid \mathbf{T}_{k-1}, T_k = c'_k]| = \\
&\qquad |(p(c_k, c_k) - p(c_k, c'_k))\texttt{Pr}[T_{v,w} = c_k] + (p(c'_k, c_k) - p(c'_k, c'_k))\texttt{Pr}[T_{v,w} = c'_k]| \\
&\leq \texttt{Pr}[T_{v,w} = c_k] + \texttt{Pr}[T_{v,w} = c'_k] \\
&= \frac{2}{\Delta}.
\end{aligned}$$

In the last line, we used the fact that the probability of an edge receiving a given colour is $\frac{1}{\Delta}$.

- Now let us consider the case when $e_k \in N^1(v)$, i.e. choices of all indirect edges and of some direct edges have been exposed. We shall now decompose f in a different way: introduce the indicators $Y_c, c \in \Delta$,

$$Y_c := \begin{cases} 1; & \text{if an edge in } N^1(v) \text{ is coloured } c, \\ 0; & \text{otherwise.} \end{cases}$$

and note that $f = \sum_c Y_c$. Further note that, if $c \neq c_k, c'_k$,

$$\texttt{Pr}[Y_c = 1 \mid \mathbf{T}_{k-1}, T_k = c_k] = \texttt{Pr}[Y_c = 1 \mid \mathbf{T}_{k-1}, T_k = c'_k].$$

Thus, we only need to consider the terms corresponding to the colours c_k, c'_k. However, since for any c,

$$\texttt{Pr}[Y_c = 1 \mid \mathbf{T}_{k-1}, T_k = c_k], \texttt{Pr}[Y_c = 1 \mid \mathbf{T}_{k-1}, T_k = c'_k] \leq 1,$$

the absolute value of their difference is at most 2. We conclude that in the case $e_k \in N^1(v)$,

$$|\mathtt{E}[f \mid \mathbf{T}_{k-1}, T_k = c_k] - \mathtt{E}[f \mid \mathbf{T}_{k-1}, T_k = c'_k]|, \leq 2.$$

Overall, denoting by a_k the coefficients of the ABDC, we get:

$$\sum_k a_k^2 = \sum_{e \in N^2(v)} \frac{4}{\Delta^2} + \sum_{e \in N^1(v)} 4 \leq 4(\Delta + 1).$$

Now the concentration bound on f is

$$\mathtt{Pr}[|f - \mathtt{E}[f]| > t] \leq 2 \exp\left(-\frac{t^2}{8(\Delta+1)}\right),$$

which for $t = \epsilon\Delta$ gives a probability that decreases exponentially in Δ. As remarked, if $\Delta \gg \log n$, this implies that the new max degree is sharply concentrated around the mean.

5.7 Hypergeometric Distribution

The hypergeometric distribution describes the number of red balls drawn in an experiment where n balls are sampled without replacement from a bin containing N balls, M of which are red. This can be regarded as a function $f(X_1, \ldots, X_n)$ where $X_1, \ldots, X_n$ are independent random variables, the variable X_i taking values in the set $[N - i + 1]$ for $i \in [n]$ giving the number of the ball drawn on the ith trial. To estimate $|\mathtt{E}[f \mid \mathbf{X}_i] - \mathtt{E}[f \mid \mathbf{X}_{i-1}]|$, let N_{i-1} be the total number of balls and M_{i-1} the number red balls at the stage when the ith ball is drawn, for $i \in [n]$. Thus $N_0 = N$, $M_0 = M$ and $N_i = N - i$. Observe that

$$\mathtt{E}[f \mid \mathbf{X}_i] = (M - M_i) + \frac{M_i}{N_i}(n - i),$$

and furthermore that $M_{i-1} - M_i \leq 1$. From this, we conclude that

$$\begin{aligned} |\mathtt{E}[f \mid \mathbf{X}_i] - \mathtt{E}[f \mid \mathbf{X}_{i-1}]| &\leq \max\left(\frac{M_{i-1}}{N_{i-1}}, 1 - \frac{M_{i-1}}{N_{i-1}}\right)\frac{N-n}{N-i} \\ &\leq \frac{N-n}{N-i}. \end{aligned}$$

Furthermore

$$
\begin{aligned}
\sum_i \left(\frac{N-n}{N-i}\right)^2 &= (N-n)^2 \sum_i \frac{1}{(N-i)^2} \\
&= (N-n)^2 \sum_{N-n \le j \le N-1} \frac{1}{j^2} \\
&\approx (N-n)^2 \int_{N-n}^{N-1} \frac{1}{x^2} dx \\
&= (N-n)\frac{n-1}{N-1}.
\end{aligned}
$$

Thus we get the bound:

$$
\mathtt{Pr}[|f - \mathtt{E}[f]| > t] \le \exp\left(\frac{-(N-1)t^2}{2(N-n)(n-1)}\right).
$$

Commentary? Thus with $t := \epsilon \mathtt{E}[f]$ and $\mathtt{E}[f] = \frac{M}{N}n$, we get

$$
\mathtt{Pr}[|f - \frac{M}{N}n| > \epsilon \frac{M}{N}n] \le \exp\left(-\epsilon^2 \frac{M}{N}\frac{M}{N-n}n\right).
$$

5.8 Consensus

Work in progress...

5.9 Problems

Work in progress...

Problem 5.1 *(From [1], p.92) Let ρ be the hamming metric on $H := \{0,1\}^n$. For $A \subseteq H$, let $B(A,s)$ denote the set of $y \in H$ so that $\rho(x,y) \leq s$ for some $x \in A$. ($A \subseteq B(A,s)$ as we may take $x = y$.) Show that if $\epsilon, \lambda > 0$ satisfy $e^{-\lambda^2/2}$ then,*

$$|A| \geq \epsilon 2^n \;\Rightarrow\; |B(A, 2\lambda\sqrt{n})| \geq (1-\epsilon)2^n.$$

Problem 5.2 *(From [1], p.91) Let B be a normed space and let $v_1, \ldots, v_n \in B$ with $|v_i| \leq 1$ for each $i \in [n]$. Let $\epsilon_1, \ldots, \epsilon_n$ be independent and uniform in $\{-1,+1\}$. Set $f := |\sum_i \epsilon_i v_i|$. Show that f is Lipschitz and deduce a sharp concentration result. Can you improve this by using the method of bounded martingale differences?*

Problem 5.3 *[Concentration around the Mean and the Median]*
Show that the following forms of the concentration of measure phenomenon for a function f defined on a space are all equivalent:

- *There exists a a such that for all $t > 0$,*
$$\mathtt{Pr}[|f - a| > t] \leq k_1 e^{-\delta_1 t^2}.$$
- *For all $t > 0$,*
$$\mathtt{Pr}[|f - \mathtt{E}[f]| > t] \leq k_2 e^{-\delta_2 t^2}.$$
- *For all $t > 0$*
$$\mathtt{Pr}[|f - \mathtt{M}[f]| > t] \leq k_3 e^{-\delta_3 t^2}.$$
(Here $\mathtt{M}[f]$ is a median of f.)

Moreover, show that all k_is are linearly related to each other and so are the δs.

Problem 5.4 (Geometric Probability) *Let Q be a given point in the unit square $[0,1]^2$ and let $P_1, \ldots, P_l$ be l points chosen uniformly and independently at random in the unit square. Let Z denote the shortest distance from Q to one of the points $P_1, \ldots, P_l$.*
(a) Observe that if $Z > x$, then no P_i lies within the circle $C(Q,x)$ centered at Q with radius x. Note that $x \leq \sqrt{2}$.

(b) Argue that there is a constant c such that for all $x \in (0, \sqrt{2}]$, the intersection of $C(Q, x)$ with the unit square has area at least cx^2. Hence deduce that

$$\mathtt{Pr}[Z > x] \leq (1 - cx^2)^l, \quad x \in (0, \sqrt{2}].$$

(c) Integrate to deduce that $\mathtt{E}[Z] \leq d/\sqrt{l}$ for some constant $d > 0$.

Problem 5.5 (Isoperimetry in the Cube) *A Hamming ball of radius r centered at a point c in the cube $\{0,1\}^n$ is the set of all points at distance at most $r-1$ and some points at distance r from c. A beautiful result of Harper states that for any two subsets X and Y in the cube, one can find Hamming balls B_0 centered at 0 and B_1 centered at 1 such that $|B_0| = |X|$, $|B_1| = |Y|$, and $d_H(B_0, B_1) \geq d_H(X, Y)$. Use this result and the Chernoff bound to show that if A is a subset in the cube of size at least 2^{n-1}, then $|A_t| \geq (1 - e^{-t^2/2n})2^n$.*

Problem 5.6 (Sampling without replacement) *Show that the sequence $\frac{M_i}{N_i}, i \geq 0$ is a martingale. Apply Azuma's inequality to deduce a sharp concentration result on the number of red balls drawn in a sample of size n.*

Problem 5.7 (Sampling without replacement) *Apply the result of Problem 4.6 to strengthen the bound for sampling without replacement.*

Chapter 6

Talagrand's Isoperimetric Inequality

6.1 The inequality

Recall that the setting for an isoperimetric inequality is a space Ω equipped with a probabilty measure P and a metric d. An isoperimetric inequality in this scenario states that if $A \subseteq \Omega$ is such that $P(A) \geq 1/2$ then $P(A_t^d) \geq 1 - \alpha(t)$ for some rapidly decreasing function α. (Recall that the neighbourhood set $A_t^d := \{x \in \Omega \mid d(x, A) \leq t\}$.

Talagrand's inequality applies in the setting where $\Omega = \prod_{i \in I} \Omega_i$ is a product space indexed by some finite index set I with the product measure $\prod_{i \in I} P_i$ where P_i is an arbitrary measure on the Ω_i, for $i \in I$. Below we will always assume this setting.

The crucial matter however is the distance. Talagrand's *convex distance* is defined by

$$d^T(x, A) := \sup_{0 \neq \alpha \geq 0} \min_{y \in A} \frac{\sum_{x_i \neq y_i} \alpha_i}{\left(\sum_i \alpha_i^2\right)^{1/2}}. \tag{6.1}$$

The sup is taken over all non–negative reals α_i which are not all zero. By normalising the α_is we can write this also as:

$$d^T(x, A) := \sup_{\sum_i \alpha_i^2 = 1} \min_{y \in A} \sum_{x_i \neq y_i} \alpha_i.$$

A caveat: we can define a distance $\hat{d}^T(x, y)$ between points by setting $\hat{d}^T(x, y) := d^T(x, \{y\})$. However, if as usual we then define $\hat{d}^T(x, A) := \min_{y \in A} \hat{d}^T(x, y)$, we *do not recover* Talagrand's distance $d^T(x, A)$. In fact

$$d^T(x, A) := \sup_{\sum_i \alpha_i^2 = 1} \min_{y \in A} \sum_{x_i \neq y_i} \alpha_i \leq \min_{y \in A} \sup_{\sum_i \alpha_i^2 = 1} \sum_{x_i \neq y_i} \alpha_i = \hat{d}^T(x, A).$$

Theorem 6.1 (Talagrand's Inequality) *Let A be a subset ina product space. Then for any $t > 0$,*

$$\mathtt{Pr}[A]\mathtt{Pr}[\overline{A_t^T}] \leq e^{-t^2/4}. \tag{6.2}$$

6.2 *Deja vu*

To gain some intuition about the Talagrand distance, let us set $\alpha_i := 1$ for each i; then we get

$$\begin{aligned} d^T(x, A) &\geq \min_{y \in A} \frac{\sum_{x_i \neq y_i} 1}{\sqrt{n}} \\ &= d^H(x, A)/\sqrt{n}, \end{aligned} \tag{6.3}$$

where d^H is the familiar Hamming distance. This implies that for any $t > 0$,

$$A_{t/\sqrt{n}}^T \subseteq A_t^H. \tag{6.4}$$

These two simple observations give us some notable consequences.

Consider the simplest product space, $\{0, 1\}^n$ equipped with the product measure where $\mathtt{Pr}[0] = 1/2 = \mathtt{Pr}[1]$ in each co–ordinate (this is the same as the uniform measure on the whole space). Take

$$A := \{x \in \{0, 1\}^n \mid \sum_i x_i \geq n/2\}.$$

Note that

$$A_t^H = \{x \in \{0, 1\}^n \mid \sum_i x_i \geq n/2 - t\},$$

and by (6.4) and Talagrand's inequality (6.2), we get

$$\begin{aligned} \mathtt{Pr}[\overline{A_t^H}] &\leq \mathtt{Pr}[\overline{A_{t/\sqrt{n}}^T}] \\ &\leq \frac{1}{\mathtt{Pr}[A]} e^{-t^2/4n} \\ &\leq 2e^{-t^2/4n} \quad \text{since } \mathtt{Pr}[A] \geq 1/2. \end{aligned}$$

This is a disguised form of the Chernoff bound (except for small constant factors) for deviations below the mean! By considering $A' := \{x \in \{0, 1\}^n \mid \sum_i x_i \leq n/2\}$, we can similarly get the Chernoff bound for deviations above the mean.

Exercise 6.2 *Show that one can extend this to the heterogeneous case as well (once again upto constant factors).*

Now let $A \subseteq \{0,1\}^n$ be an arbitrary set with $\mathtt{Pr}[A] \geq 1/2$. By the same reasoning as above, we get:

$$\mathtt{Pr}[\overline{A_t^H}] \leq 2e^{-t^2/4n},$$

a cleaner form of the isoperimetric inequality we derived using martingales and the method of bounded differences.

6.3 Hereditary Properties

We will develop a general framework to analyse a certain class of functions on product spaces which are defined by hereditary (i.e. monotone) properties of index sets. This framework generalises slightly the results implicit in Talagrand [15] and explicit in in Steele [14] and Spencer [13]. We then illustrate the versatality of this framework by several examples.

6.3.1 A General Framework

For $x, y \in \Omega$ and $J \subseteq I$, we use the notation $x_J = y_J$ to mean $x_j = y_j, j \in J$, and for a family J of subsets of I, by $x_{\mathrm{J}} = y_{\mathrm{J}}$ we mean $x_J = y_J, J \in \mathrm{J}$. Let $\mathrm{J}_{x=y} := \{J \in \mathrm{J} \mid x_J = y_J\}$ and note that $x_{\mathrm{J}_{x=y}} = y_{\mathrm{J}_{x=y}}$.

Let $\phi(x, \mathrm{J})$ be a boolean property such that it is

- a *property of index sets* i.e. if $x_{\mathrm{J}} = y_{\mathrm{J}}$, then $\phi(x, \mathrm{J}) = \phi(y, \mathrm{J})$, and
- *non–increasing on the index sets*, i.e. if $\mathrm{J} \subseteq \mathrm{J}'$ then $\phi(x, \mathrm{J}) \leq \phi(x, \mathrm{J}')$.

. We shall say that ϕ is a *hereditary property of index sets.*

Let $w(i, J), i \in I, J \subseteq I$ be a matrix of non–negative weights. Set $w(J) := \sum_i w(i, J)$. The function f_ϕ defined by a hereditary property ϕ of index sets is given by:

$$f_\phi(x) := \max_{\mathrm{J}:\phi(x,\mathrm{J})} \sum_{J \in \mathrm{J}} w(J). \tag{6.5}$$

By $\mathrm{J}(x)$, we shall denote a family achieving the maximum in (6.5). Thus,

$$f_\phi(x) := \sum_{J \in \mathrm{J}(x)} w(J). \tag{6.6}$$

A function f such that $f = f_\phi$ for some hereditary property ϕ of index sets will be called a *hereditary function of index sets*.

Theorem 6.3 *Let f be a hereditary function defined by a hereditary property ϕ and weights $w(i, J), i \in I, J \subseteq I$. Then for all $t > 0$,*

$$\mathtt{Pr}[f > \mathtt{M}[f] + t] \leq 2 \exp\left(\frac{-t^2}{4wk(M[f]+t)}\right),$$

and

$$\mathtt{Pr}[f < \mathtt{M}[f] - t] \leq 2 \exp\left(\frac{-t^2}{4wkM[f]}\right),$$

where $\mathtt{M}[f]$ is a median of f and

$$k := \max_{x\in\Omega} \max_{J \in \mathrm{J}(x)} |J|,$$

and

$$w := \max_{x\in\Omega} \max_{i\in I} \sum_{\substack{J\in \mathrm{J}(x) \\ i \in J}} \sum_J w(J)$$

.

Proof. First note that for any $x, y \in \Omega$,

$$\begin{aligned} 1 &= \phi(x, \mathrm{J}(x)) \\ &\leq \phi(x, \mathrm{J}(x)_{x=y}), \quad \text{by monotonicity} \\ &= \phi(y, \mathrm{J}(x)_{x=y}), \quad \text{index set property of } \phi. \end{aligned}$$

Hence,

$$\begin{aligned} f(y) &= \max_{\substack{\mathrm{J} \\ \phi(y,\mathrm{J})}} \sum_{J\in\mathrm{J}} w(J) \\ &\geq \sum_{J\in \mathrm{J}(x)_{x=y}} w(J) \\ &= \sum_{J\in\mathrm{J}(x)} w(J) - \sum_{J\in\mathrm{J}(x)_{x\neq y}} w(J) \\ &= f(x) - \sum_{\substack{J\in\mathrm{J}(x) \\ x_J \neq y_J}} w(J) \end{aligned} \tag{6.7}$$

Define $\alpha_i = \alpha(x)_i := \sum_{J \in \mathrm{J}(x)} 1[i \in J] w(J)$ for $i \in I$. Then,

$$\begin{aligned}
\sum_{x_i \neq y_i} \alpha_i &= \sum_{x_i \neq y_i} \sum_{J \in \mathrm{J}(x)} 1[i \in J] w(J) \\
&= \sum_{J \in \mathrm{J}(x)} w(J) \sum_{i \in J} 1[x_i \neq y_i] y_i] \\
&\geq \sum_{J \in \mathrm{J}(x)} w(J) 1[x_J \neq y_J].
\end{aligned}$$

Hence, by (6.7), we have,

$$f(y) \geq f(x) - \sum_{x_i \neq y_i} \alpha_i. \tag{6.8}$$

Now set $A = A(a) := \{y \in \Omega \mid f(y) \leq a\}$. Then we have by the definition of Talagrand's distance,

$$\begin{aligned}
d_T(x, A) &\geq \frac{\min y \in A \sum_{x_i \neq y_i}}{(\sum_i \alpha_i^2)} \alpha(x)_i \\
&= \frac{\sum_{x_i \neq y_i^*} \alpha_i}{(\sum_i \alpha_i)},
\end{aligned}$$

for some $y^* \in A$. Further note that

$$\begin{aligned}
\sum_i \alpha_i^2 &= \sum_i \left(\sum_{J \in \mathrm{J}(x)} 1[i \in J] w(J) \right)^2 \\
&\leq \sum_i w \sum_{J \in \mathrm{J}(x)} 1[i \in J] w(J) \\
&= w \sum_{J \in \mathrm{J}(x)} w(J) \sum_{i \in J} 1 \\
&\leq wk \sum_{J \in \mathrm{J}(x)} w(J) \\
&= wkf(x).
\end{aligned}$$

Thus,

$$\sum_{x_i \neq y_i} \alpha(x)_i \leq \sqrt{wkf(x)} d_T(x, A),$$

and plugging into (6.8), we get:

$$d_T(x, A) \geq \frac{f(x) - a}{\sqrt{wkf(x)}}.$$

Now applying Talagrand's inequality, we have,

$$\begin{aligned}
\mathtt{Pr}[f(X) > a+t] &= \mathtt{Pr}\left[\frac{f(X)-a}{\sqrt{wk(a+t)}} > \frac{t}{\sqrt{wk(a+t)}}\right] \\
&\leq \mathtt{Pr}\left[\frac{f(X)-a}{\sqrt{wkf(x)}} > \frac{t}{\sqrt{wk(a+t)}}\right] \\
&\leq \mathtt{Pr}\left[d_T(x,A) > \frac{t}{\sqrt{k(a+t)}}\right] \\
&\leq \frac{1}{\mathtt{Pr}[A]}\exp\left(\frac{-t^2}{4wk(a+t)}\right).
\end{aligned}$$

Let us rewite this as

$$\mathtt{Pr}[f(X) \leq a]\mathtt{Pr}[f(X) > a+t] \leq \exp\left(\frac{-t^2}{4wk(a+t)}\right).$$

Setting $a := \mathtt{M}[f]$ and $a := \mathtt{M}[f] - t$ successively,we get the result. ■

Setting $t := \epsilon\mathtt{M}[f]$ in Proposition 6.3, we get

$$\mathtt{Pr}[f > (1+\epsilon)\mathtt{M}[f]] \leq 2\exp\left(\frac{-\epsilon^2}{4k(1+\epsilon)}\mathtt{M}[f]\right), \quad \mathtt{Pr}[f < (1-\epsilon)\mathtt{M}[f]] \leq 2\exp\left(\frac{-\epsilon^2}{4k}\mathtt{M}[f]\right).$$

One can obtain a concentration of measure inequality around the mean rather than the median via the following genral result:

Proposition 6.4 *The following are equivalent for an arbitrary function f and random variables $X_1, \ldots, X_n$:*

- *For all $t > 0$, there are positive constants c and α such that*

$$\mathtt{Pr}[|f - \mathtt{M}[f]| > t] \leq ce^{-\alpha t}.$$

-

- *For all $t > 0$, there are positive constants c' and α' such that*

$$\mathtt{Pr}[|f - \mathtt{E}[f]| > t] \leq c'e^{-\alpha' t}.$$

6.3.2 Examples

We illustrate the general framework with several examples. The example of increasing subsequences in § 6.3.2 was one of the motivating examples for Talagrand's inequality [15], see also [14]. The examples involving counting extensions in § 6.3.2 and counting representations in § 6.3.2 were pointed out by Spencer [13].

In the examples below, we will always identify a singleton set with the element itself. WE will usually give the weighted version of the statement from which the unweighted version follows easily by setting all weights to unity.

Increasing subsequences

Let $w_1, \ldots, w_n$ be given non–negative weights and let $I(x_1, \ldots, x_n)$ denote the weight of the heaviset increasing subsequence from $x_1, \ldots, x_n$. Let $X_1, \ldots, X_n$ be chosen independently at random from $[0, 1]$. Proposition 6.3 can be applied immediately to give a sharp concentration result on $I(X_1, \ldots, X_n)$.

Set $w(i, J) := [i \in J]w_i$. Take the hereditary property $\phi(x, \mathrm{J})$ to be:

- $|J| = 1$ for all $J \in \mathrm{J}$.
- For $j, j' \in \mathrm{J}$, if $j < j'$ then $x(j) \leq x(j')$.

Here $k = 1$ and $w := \max_i w_i$. Taking $w_i = 1$ for all $i \in [n]$ gives the original result.

Balls and Bins

Consider the probabilistic experiment where m balls are thrown independently at random into n bins and we are interested in a sharp concentration result on the number of empty bins. Equivalently, we can give a sharp concentration result on the number of non–empty bins.

To cast this in the framework of configuration functions, consider the product space $[n]^m$ with the product measure where $\mathtt{Pr}[X_k = i]$ is the probability that ball k is thown into bin i. We shall give a sharp concentration result on the number of non–empty bins. Take $w(i, J) = 1[i \in J]$ and the hereditary property $\phi(x, \mathrm{J})$ to be:

- $|J| = 1$ for all $J \in \mathrm{J}$.
- For distinct $j, j' \in \mathrm{J}$, $x(j) \neq x(j')$.

Discrete Isoperimetric Inequalities

Let A be a downward closed subset of the cube $\{0,1\}^n$ equipped with the product measure, and let us consider the Hamming distance $d_H(x, A)$ from a point x to the set A. This is in fact a function of hereditary index sets. Take the weight system $w(i, J) := 1[i \in J]$ and the hereditary property $\phi(x, \mathrm{J})$ to be:

- $|J| = 1$ for all $J \in \mathrm{J}$.
- For all $j \in \mathrm{J}$, $x(j) = 1$ and $y(j) = 0$ for all $y \in A$.

We have $k = 1 = w$ and the bound obtained is comparable with the bounds obtained directly by isoperimetric inequalities in the theory of hereditary sets [3] (see also [14, p. 132].

Counting Extensions

Consider the random graph $G(n, p)$ with non–negative weights $w_e, e \in E$ on the edges. A subset E' of the edges has weight $w(E') := \sum_{e \in E'} w_e$. For a fixed vertex u, let let $T(u)$ denote the weight of the triangles containing a given vertex u. One obtains a sharp concentration result as follows. Consider the product space $\{0,1\}^E$, take $w(e, J) := 1[e \in J]w_e$ and take the hereditary property $\phi(x, \mathrm{J})$ to be:

- $|J| = 3$ for all $J \in \mathrm{J}$.
- The edges in $\{e_1, e_2, e_3\}$ in J form a triangle including the vertex u.

With all weights $w_e = 1$, it is easy to see that the number of triangles containing u is $\mathtt{E}[N(u)] = \binom{n-1}{2}p^3$ and we get sharp concentration around this value. Similarly one can prove generalisations due to Spencer for the number of extensions $N(x_1, \ldots, x_r)$ of a given set of vertices to a copy of a given graph H.

Counting Representations

For a given set S of natural numbers, let $f_S(n)$ denote the number of representations $n = x + y$ for distinct $x, y \in S$. An old result of Erdös shows that there is a set S for which $f_S(n) = \theta(\ln n)$. This is obtained by a probabilistic construction: define S randomly setting $\mathtt{Pr}[x \in S] := \min\left(1, 10\sqrt{\frac{\ln x}{x}}\right)$. One can show that $\mathtt{E}[f_S(n)] \sim 100\pi \ln n$. A sharp concentration result about this value is obtained by taking unit weights $w_i = 1, i \geq 1$ and the hereditary property $\phi(x, \mathrm{J})$ to be:

- $|J| = 2$ for all $J \in \mathrm{J}$.
- $x + y = n$ for all $\{x, y\} \in \mathrm{J}$.

Note that in this example, $k = 2$ and also $w = 2$ since an element x can be in at most one subset in J.

Extending this let $g_S(n)$ denote the number of representations $n = x + y + z$ for distinct $x, y, z \in S$. Erdös and Tetali choose S randomly by letting $\mathtt{Pr}[i \in S] := \min(10 \left(\frac{\ln i}{i^2}\right)^{1/3} 1/2)$ and show that $\mathtt{E}[g_S(n)] = K \ln n$ for some constant $K > 0$. In fact one obtains a sharp concentration result around this value by bootstrapping the previous result. Set all weights to be unity as before and take the hereditary property $\phi(x, \mathrm{J})$ to be:

- $|J| = 3$ for all $J \in \mathrm{J}$.
- $x + y + z = n$ for all $\{x, y, z\} \in \mathrm{J}$.

This time $k = 3$ and w is bounded by $f_S(n)$ before.

Similarly, one can bootstrap upwards to k–ary representations for $k > 3$.

Edge Colouring

In this example, we shall considre some simple randomised algorithms for edge colouring a graph and illustrate the flexibility of our framework for giving sharp concentration results.

Given a graph G and a palette Δ of colours, we would like to assign colours to the edges in such a way that no two edges incident on a vertex have the same colour. We would also like the algorithm to be truly distributed, so we would like it to have a local character. This leads naturally to randomised algorithms of the type considered below. These algorithms run in stages. At each stage, some edges are successfully coloured. The others pass on to the next stage. Typically one analyses the algorithm stage by stage; in each stage, we would like to show that a significant number of edges are successfully coloured, so that the graph passed to the next stage is significantly smaller.

For simplicity, we assume that the graph G is bipartite with bipartition U, V (note that even colouring bipartite graphs in a distributed fashion is non–trivial).

Algorithm 1: each edge picks a colour independently from the common palette $[\Delta]$. Conflicts are resolved in a two steps:

- First the V vertices resolve conflicts: if there are two edges (u_i, v) and (u_j, v) with the same colour with $i < j$, then (u_j, v) "loses" and is decoloured.
- Next the U vertices resolve any remaining conflicts by choosing one "winner" out of the remaining conflicting edges for each colour.

We are interested in a sharp concentration result on the number of edges around a fixed vertex $u \in U$ that are successfuly coloured (A similar analysis works for a vertex in V). Alternatively, we can give a sharp concentration result on the number of edges around u that are *not* successfully coloured.

The underlying product space is $[\Delta]^E(u)$ where $E(u)$ is the set of edges that are incident to u or to a neighbour of u. Let $w_(e, J) := [e \in J][u \in e]$. Take the hereditary property $\phi(x, \mathrm{J})$ to be:

- The sets in J are all disjoint and $|J| = 2$ for all $J \in \mathrm{J}$.
- For each $\{e, e'\} \in \mathrm{J}$, at least one of e, e' is incident on u and e, e' are incident on a common vertex.
- For each $\{e, e'\} \in \mathrm{J}$, $x(e) = x(e')$.

Some comments are in order about how to find $\mathrm{J}(x)$ in this case.

- First, for each edge $e = (u, v)$ which is unsuccessful beacuse of an edge $e' = (u', v)$ which received the same colour, pick the set $\{e, e'\}$.
- This leaves, for each colour, a bunch of edges incident on u with the same colour. Group these in pairs except for thepick a conflicting edge e' (i.e. e and e' are adjacent and $x(e) = x(e')$). In the case that e and e' are both

incident on u and are the only edges in conflict with each other, pick the pair $\{e, e'\}$ with multiplicity 2 in J. One can verify that the above conditions are indeed satisfied.

6.4 Talagrand's Inequality with Negative Dependence

In this section we give an extension of Talagrand's inequality to setting where the the underlying measure in the product space is not the product measure but satisfies a negative dependence property.

6.4.1 A General Framework

A function $f(x_i, i \in i)$ is said to be non–decreasing (or non–increasing) if it is non–decreasing (non–increasing) in each co–ordinate. A subset A is non–decreasing (non–increasing) if its characteristic function $\chi(x_i, i \in I) := 1[(x_i, i \in I) \in A]$ is.

A set of variables $X_1, \ldots, X_n$ satisfies the *negative regression* condition $(-R)$ if for all disjoint index sets I and J of $[n]$ and all non–decreasing functions $f(x_i, i \in I)$,

$$E[f(X_i, i \in I) \mid X_j = t_j, j \in J],$$

is non–increasing in each $t_j, j \in J$. A measure in a product space is said to satisfy negative regression if the co–ordinate functions $X_k(\omega_i, i \in I) = \omega_k$ satisfy negative regression.

A set of variables $X_1, \ldots, X_n$ is said to be *exchangeable* or *symmetric* if for all permutations $\sigma : [n] \to [n]$ and all $a_1, \ldots, a_n$,

$$\mathtt{Pr}[X_i = a_i, i \in [n]] = \mathtt{Pr}[X_i = a_{\sigma(i)}, i \in [n]].$$

A product space is symmetric if the co–ordinate functions are symmetric.

Theorem 6.5 (Talagrand's Inequality with Negative Dependence) *Let $\Omega = \{0, 1\}^I$ be a symmetric product space satisfying $(-R)$. Then for any non–increasing subset $A \subseteq \Omega$,*

$$\mathtt{Pr}[A]\mathtt{Pr}[\overline{A_t}] \leq e^{-t^2/4}.$$

The proof follows very much along the lines of the usual proof of Talagrand's inequality. There are two main ingredients of the standard proof. The first is a general probabilistic fact, namely *Holder's inequality*: for any two random variables X and Y (whether independent or not), and any two reals p, q with $1/p + 1/q = 1$,

$$\mathtt{E}[|XY|] \le \mathtt{E}[X^p]^{1/p}\mathtt{E}[Y^q]^{1/q}. \tag{6.9}$$

The second is a key inequality having to do with the convexity of Talagrand's distance function. For a subset $A \subseteq \Omega = \prod_{i\in[n]} \Omega_i$, denote by

$$\pi(A) := \{x \mid (x, x_n) \in A\},$$

the *projection* of A on all coordinates except the last, and for $x_n \in \Omega_n$, let

$$A(x_n) := \{x \mid (x, x_n) \in A\},$$

denote the x_n–*section* of A. Then for $(x, x_n) \in \Omega$, and all $0 \le \lambda \le 1$,

$$d_T^2((x, x_n), A) \le (1-\lambda)d_T^2(x, \pi(A)) + \lambda d_T^2(x, A(x)) + (1-\lambda)^2. \tag{6.10}$$

We will also use crucially, two more simple observations. First are some general properties of the Talagrand distance valid in any product space of ordered spaces.

Proposition 6.6 *Let $A \subseteq \Omega$ be a non–increasing subset. Then*

- *Any projection $\pi(A)$ as well as any section $A(\omega)$ are also non–increasing.*
- *The Talagrand distance $d_T(x, A)$ is a non–decreasing function of x.*
- *$d_T(x_1, \ldots, x_{n-1}, x_n, A(x_n))$ is a non–decreasing function of x.*

The second is a property of the "influence" of Boolean variables from [5] tht is of independent interest and has other applications. Let $f(x_i, i \in [n]$ be a Boolean function. A variable $x_i, i \in [n]$ is said to have *positive influence* for f and a given diatribution if

$$\mathtt{E}[f(X_1, \ldots, X_n) \mid X_i = x],$$

is non–decreasing in x.

Proposition 6.7 *For any Boolean function $f(x_1, \ldots, x_n)$ and any distribution of its arguments, there is always a variable of positive influence i.e. an $i \in [n]$ such that*

$$\mathtt{E}[f(X_1, \ldots, X_n) \mid X_i = 0] \le \mathtt{E}[f(X_1, \ldots, X_n) \mid X_i = 1].$$

Proof. (Of Theorem 6.5): We have,

$$\begin{aligned}\texttt{Pr}[d_T(X,A) > t] &= \texttt{Pr}[e^{\frac{1}{4}d_T^2(X,A)} > e^{\frac{1}{4}t^2}] \\ &\le \texttt{E}[e^{\frac{1}{4}d_T^2(X,A)}]e^{-\frac{1}{4}t^2}.\end{aligned}$$

Thus to complete the proof, we only need to show that $\texttt{E}[e^{\frac{1}{4}d_T^2(X,A)}] \le \frac{1}{\texttt{Pr}[A]}$.

This will be proved by induction on the dimension n. For $n = 1$, we have $d_T(x,A) = 1[x \notin A]$ and hence

$$\begin{aligned}\texttt{E}[e^{\frac{1}{4}d_T^2(X,A)}] &= e^{1/4}(1 - \texttt{Pr}[A]) + \texttt{Pr}[A] \\ &\le 1/\texttt{Pr}[A],\end{aligned}$$

since $e^{1/4}(1-u) + u \le 1/u$ for $0 \le u \le 1$ by elementary calculus.

For the induction step, we shall use the key convexity inequality (6.10) to write

$$\begin{aligned}&\texttt{E}[e^{\frac{1}{4}d_T^2(X_1,\ldots,X_{n+1},A)}] \le \\ &\qquad \texttt{E}\left[e^{\frac{1}{4}(1-\lambda)^2}e^{(1-\lambda)\frac{1}{4}d_T^2(X_1,\ldots,X_n,\pi(A))}e^{\lambda\frac{1}{4}d_T^2(X_1,\ldots,X_n,A(X_{n+1}))}\right] \\ &= \texttt{E}\left[e^{\frac{1}{4}(1-\lambda)^2}\texttt{E}\left[e^{(1-\lambda)\frac{1}{4}d_T^2(X_1,\ldots,X_n,\pi(A))}e^{\lambda\frac{1}{4}d_T^2(X_1,\ldots,X_n,A(X_{n+1}))} \mid X_{n+1}\right]\right] \\ &\le \texttt{E}\left[e^{\frac{1}{4}(1-\lambda)^2}\left(\texttt{E}\left[e^{\frac{1}{4}d_T^2(X_1,\ldots,X_n,\pi(A))} \mid X_{n+1}\right]\right)^{1-\lambda}\left(\texttt{E}\left[e^{\frac{1}{4}d_T^2(X_1,\ldots,X_n,A(X_{N+1}))} \mid X_{n+1}\right]\right)^{\lambda}\right],\end{aligned}$$

using Holder's inequality in the last line.

Now, using Proposition 6.6 and the $(-R)$ property, we have that

$$\texttt{E}\left[e^{\frac{1}{4}d_T^2(X_1,\ldots,X_n,\pi(A))} \mid X_{n+1} = x_{n+1}\right],$$

is a non–increasing function of x_{n+1} whereas by Proposition 6.7,

$$\texttt{E}\left[e^{\frac{1}{4}d_T^2(X_1,\ldots,X_n,A(x_{n+1}))} \mid X_{n+1} = x_{n+1}\right]$$

is a non–decreasing function of x_{n+1}. Hence,

$$\begin{aligned}&\texttt{E}[e^{\frac{1}{4}d_T^2(X_1,\ldots,X_{n+1},A)}] \le \\ &\qquad \texttt{E}\left[e^{\frac{1}{4}(1-\lambda)^2}\texttt{E}\left[e^{\frac{1}{4}d_T^2(X_1,\ldots,X_n,\pi(A))} \mid X_{n+1}\right]\texttt{E}\left[e^{\frac{1}{4}d_T^2(X_1,\ldots,X_n,A(X_{n+1}))} \mid X_{n+1}\right]\right] \\ &= e^{\frac{1}{4}(1-\lambda)^2}\texttt{E}\left[\texttt{E}\left[e^{\frac{1}{4}d_T^2(X_1,\ldots,X_n,\pi(A))} \mid X_{n+1}\right]\right]\texttt{E}\left[\texttt{E}\left[e^{\frac{1}{4}d_T^2(X_1,\ldots,X_n,A(X_{n+1}))} \mid X_{n+1}\right]\right] \\ &= e^{\frac{1}{4}(1-\lambda)^2}\texttt{E}\left[e^{\frac{1}{4}d_T^2(X_1,\ldots,X_n,\pi(A))}\right]\texttt{E}\left[e^{\frac{1}{4}d_T^2(X_1,\ldots,X_n,A(X_{n+1}))}\right] \\ &\le \texttt{E}\left[e^{\frac{1}{4}(1-\lambda)^2}\left(\frac{1}{\texttt{Pr}[\pi(A)]}\right)^{1-\lambda}\left(\frac{1}{\texttt{Pr}[A(X_{n+1})]}\right)^{\lambda}\right],\end{aligned}$$

where the last line follows by induction. applied to the sets $\pi(A)$ and $A(X_n)$.

Now, we appeal to simple calculus once again to show that for $0 \leq r \leq 1$, $\inf_{0\leq\lambda\leq 1} r^{-\lambda}e^{\frac{1}{4}(1-\lambda)^2} \leq 2-r$. Hence, continuing the inequalities from above, we have

$$\begin{aligned}
\mathtt{E}[e^{\frac{1}{4}d_T^2(X_1,\ldots,X_{n+1},A)}] &\leq \mathtt{E}\left[e^{\frac{1}{4}(1-\lambda)^2}\left(\frac{1}{\mathtt{Pr}[\pi(A)]}\right)^{1-\lambda}\left(\frac{1}{\mathtt{Pr}[A(X_{n+1})]}\right)^{\lambda}\right] \\
&= \frac{1}{\mathtt{Pr}[\pi(A)]}\mathtt{E}\left[\left(\frac{\mathtt{Pr}[A(X_{n+1})]}{\mathtt{Pr}[\pi(A)]}\right)^{-\lambda} e^{\frac{1}{4}(1-\lambda)^2}\right] \\
&\leq \frac{1}{\mathtt{Pr}[\pi(A)]}\mathtt{E}\left[2-\frac{\mathtt{Pr}[A(X_n)]}{\mathtt{Pr}[\pi(A)]}\right] \\
&= \frac{1}{\mathtt{Pr}[\pi(A)]}\left(2-\frac{\mathtt{Pr}[A]}{\mathtt{Pr}[\pi(A)]}\right) \\
&= \frac{1}{\mathtt{Pr}[A]}\frac{\mathtt{Pr}[A]}{\mathtt{Pr}[\pi(A)]}\left(2-\frac{\mathtt{Pr}[A]}{\mathtt{Pr}[\pi(A)]}\right) \\
&\leq \frac{1}{\mathtt{Pr}[A]},
\end{aligned}$$

since $x(2-x) \leq 1$ for all reals x. ∎

We can apply this to derive an analogue of Theorem 6.3 in a negative dependence situation. Call a hereditary property $\phi(x, \mathrm{J})$ of index sets *bi–hereditary* if it is non–decreasing in x i.e. if $x_{\mathrm{J}} \leq y_{\mathrm{J}}$, then $\phi(x,\mathrm{J}) \leq \phi(y,\mathrm{J})$.

Theorem 6.8 *Let $\Omega := \{0,1\}^n$ be a symmetric space satisfying $(-R)$ and let f be a function defined by a bi-hereditary property ϕ and weights $w(i,J), i \in I, J \subseteq I$. Then for all $t > 0$,*

$$\mathtt{Pr}[f > \mathtt{M}[f]+t] \leq 2\exp\left(\frac{-t^2}{4wk(M[f]+t)}\right),$$

and

$$\mathtt{Pr}[f < \mathtt{M}[f]-t] \leq 2\exp\left(\frac{-t^2}{4wkM[f]}\right),$$

where $\mathtt{M}[f]$ is a median of f and

$$k := \max_{x\in\Omega}\max_{J\in\mathcal{J}(x)} |J|,$$

and

$$w := \max_{x\in\Omega}\max_{i\in I}\sum_{\substack{J\in\mathcal{J}(x)\\ i\in J}}\sum_{J} w(J)$$

.

Proof. The proof is exactly analogous to that of Theorem 6.3 noting that if ϕ is bi–hereditary then the set $A(a) := \{y \mid f(y) \leq a\}$ is non–increasing, and hence Theorem 6.5 is applicable. ∎

6.4.2 Examples

We give examples to illustrate how to obtain concentration of measure results by applying Theorem 6.8 in settings where the underlying space is not a product space but is symmetric and satisfies $(-R)$.

Hypergeometric Distribution

Consider a sample of size n drawn from an urn containing $N \geq n$ balls, $M \leq N$ of which are red. We are interested in the number of red balls drawn in the sample, $H(N, M, n)$ – this is the hypergeometric distribution, .

The underlying product space is $\{0,1\}^n$ where we identify 1 with a red ball and 0 with a non–red ball. The measure here is *not* the product measure but is easily shown to satisfy $(-R)$. To apply Theorem 6.8, take the hereditary property $\phi(x, \mathrm{J})$ to be

- $|J| = 1$ for all $J \in \mathrm{J}$.
- For $j \in \mathrm{J}$, $x(j) = 1$.

This is easily verified to be a bi–hereditary property and we get the following concentration of measure result on the hypergeometric distribution:

$$\mathtt{Pr}[H(N,M,n) > m+t] \leq 2\exp(\frac{-t^2}{m+t}), \quad \mathtt{Pr}[H(N,M,n) < m-t] \leq 2\exp(\frac{-t^2}{m}),$$

which is comparable to the Chernoff bound. Taking arbitrary weights $w_i, i \in [n]$ and setting $w(i, J) := [i \in J]w_i$ gives a weighted version of the result.

Balls and Bins

Let us return to the balls ad bins experiment, considered this time as the product space $\{0,1\}^{[n]\times[m]}$ where a 1 in co–ordinate (i,k) indicates that ball k is put into bin i. This measure (for any probabilities $p_{i,k}$ of ball k going into bin i) is easily shown to satisfy $(-R)$. Take $\phi(x, \mathrm{J})$ to be the same as in the previous example: $|J| = 1, J \in \mathrm{J}$ and $x(i,k) = 1$ for $(i,k) \in \mathrm{J}$. This gives the same concentration result for the number of non–empty bins as derived in the previous section.

Fermi–Dirac statistics

Consider an ensemble of particles occupying a set of n states. There are m types of particles, with $c_k \leq n$ particles of type $k \in [m]$. Each type is independent of any other type, but the particles of a given type satisfy the *Pauli exclusion principle*: no two particles of the same type may occupy the same state. In particular, the particles of a fixed type obey the Fermi–Dirac statistics: for each $k \in [m]$, all c_k subsets of the n–states are equally likely to be occupied. We are interested in the number of unoccupied states. Equivalently we acn focus on the number of occupied states. (This example generalises the balls and bins experiment above where each $c_k = 1$ and each ball is uniformly distributed in the bins.)

The underlying product space is $\{0,1\}^{[n]\times[m]}$ and the measure can be shown to satisfy $(-R)$. (the measure is not a product measure). Take the same bi–hereditary property $\phi(x, \mathrm{J})$ as in the previous balls and bins example to get a concentration result for the number of non–empty cells.

6.5 The Method of Non–uniformly Bounded Differences

One can extract out from Talagrand's inequality a nicely packaged lemma [1] that generalises the method of bounded differences.

Theorem 6.9 (The Method of Non–uniformly Bounded Differences) *Let f be a real–valued function on a product space Ω such that for each $x \in \Omega$, there exist non–negative reals $\alpha_i(x), i \in [n]$ with*

$$f(x) \leq f(y) + \sum_{x_i \neq y_i} \alpha_i(x), \quad \text{for all } y \in \Omega. \tag{6.11}$$

Furthermore, suppose that there exists a constant $c > 0$ such that uniformly for all $x \in \Omega$,

$$\sum_i \alpha_i^2(x) \leq c \tag{6.12}$$

[1] In Steele [14][Lemma 6.2.1], this is stated with some additional superfluous conditions.

(even though the $\alpha_i(x)$ may be different individually). Then

$$\mathtt{Pr}[|f - \mathtt{M}[f]| > t] \leq 2e^{-t^2/4c}. \tag{6.13}$$

Proof. Set $A = A(a) := \{y \mid f(y) \leq a\}$, where a is a parameter to be fixed later. By (6.11), we have,

$$f(x) \leq f(y) + \sum_{x_i \neq y_i} \alpha_i(x),$$

for any y. Hence minimising over $y \in A$, we have,

$$\begin{aligned} f(x) &\leq \min_{y \in A} f(y) + \sum_{x_i \neq y_i} \alpha_i(x) \\ &\leq a + c d_T(x, A), \end{aligned}$$

by the definition of the Talagrand distance and (6.12). Hence,

$$\begin{aligned} \mathtt{Pr}[f(X) \geq a + ct] &\leq \mathtt{Pr}[d_T(X, A)|geqt] \\ &\leq \frac{1}{\mathtt{Pr}[A]} e^{-t^2/4}, \end{aligned}$$

by applying Talagrand's inequality in the last step. Hence,

$$\mathtt{Pr}[f(X) \geq a + t] \leq \exp\left(\frac{-t^2}{4c}\right).$$

Remembering that $A := \{y \mid f(y) \leq a\}$, write this as

$$\mathtt{Pr}[f(X) \geq a + t]\mathtt{Pr}[f(X) \leq a] \leq \exp\left(\frac{-t^2}{4c}\right).$$

Setting $a := \mathtt{M}[f]$ and $a := \mathtt{M}[f] - t$ successively gives the result. ∎

Note that the condition of (6.11) is just like the bounded differences condition (4.10) except that the bounding parameters can be non–uniform i.e. a different set of parameters for each x. This is the crucial feature thatmakes this version substantially more powerful than the usual method of bounded differences as we illustrate with some examples below.

6.5.1 Chernoff–Hoeffding Bounds

Let $f(x_1, \ldots, x_n) := \sum_i x_i$ with $x_1, \ldots, x_n \in [0, 1]$. Take $\alpha_i(x) := x_i$; then clearly (6.11) is satisfied. Moreover $\sum_i \alpha_i^2 \leq n$. Hence,

$$\mathtt{Pr}[|f - \mathtt{M}[f]| > t] \leq 2e^{-t^2/n},$$

which is just the Chernoff–Hoeffding bound upto constant factors.

6.5.2 Stochastic TSP

Let $X_1, \ldots, X_n$ be points selected uniformly and independently at random in the unit square and let $TSP(X)$ denote the length of the minimum TSP tour through these points. In this subsection, we shall show a sharp concentration result for the TSP tour. This was a notable success of Talagrand's inequality over the previous approcahes using Martingales.

In order to apply Proposition 6.9, we need to find suitable $\alpha_i, i \in [n]$. These turn out to emerge from considering the *space filling curve* heuristic for the TSP. A space–filling curve is a map $\psi : [0,1] \to [0,1]^2$ which is surjective. One often requires some other smoothness properties, for instance that it is Lipschitz of order $frac12$ that is,

$$||\psi(s) - \psi(t)|| \leq c_\psi |s-t|^{1/2}, \tag{6.14}$$

for some constant c_ψ. In addition, to be computationally useful, we require that it is easy to compute, for each $x \in [0,1]^2$, some $t \in [0,1]$ such that $\psi(t) = x$. Many space–filling curves satisfy these properties.

The space-filling curve heuristic is very simple. Fix a space filling curve (with the above properties). Then

- For each point x_i, find a preimage t_i, $i \in [n]$.
- Sort the preimages $t_i, i \in [n]$ in increasing order $t_{\sigma(1)}, \ldots, t_{\sigma(n)}$.
- Follow the points in this order, i.e. define the permutation let the tour be $\psi(t_\sigma(i)), i \in [n]$.

It can be shown that this heursitic gives a $O(\log n)$ approximation algorithm for the TSP.

Notice that

$$\begin{aligned}
\sum_{i<n} |x_{\sigma(i)} - x_{\sigma(i+1)}|^2 &= \sum_{i<n} |\psi(t_{\sigma(i)}) - \psi(t_{\sigma(i+1)})|^2 \\
&\leq \sum_{i<n} c_\psi^2 |t_{\sigma(i)} - t_{\sigma(i+1)}| \\
&\leq c_\psi^2.
\end{aligned}$$

(This shows that surprisingly, that there is a constant c such that for any $x_1, \ldots, x_n \in [0,1]^2$, there is a permutation $\sigma : [n] \to [n]$ such that $\sum_{i<n} |x_{\sigma(i)} - x_{\sigma(i+1)}|^2 \leq c$!)

Returning to the TSP in the unit square, choose $\alpha(x)_i$ to be twice the sum of the lengths of the two edges incident on x_i in the space-filling curve heuristic applied to $x_i, i \in [n]$. Thus, by the previous calculation, we automatically have the upper bound (6.12).

It only remains to verify (6.11).

This gives the truly Gaussian tail bound:

$$\texttt{Pr}[|TSP(X) - \texttt{M}[TSP(X)]| > t] \leq e^{-t^2/4c_\psi^2}.$$

6.5.3 First Birth Times in Branching Processes

Branching processes are a very attractive model of population growth dynamics and have also proven very useful in the study of properties of trees grown by incremental random processes. We consider here, branching processes of the *Galton–Watson* type: there is a single ancestor at time 0. This ancestor produces a number m of children at a random time Z (distributed with the exponential distribution with parameter 1) and dies. Subsequently each child independently of the others reproduces in exactly the same manner.

We can represent the process by an infinite m–ary tree whose vertices represent the memebers of the population produced and are labelled as follows: the root is given the empty label. If a vertex has label v, then its children are labelled $v1, \ldots, vm$. The root, representing the initia ancestor is the unique member of the 0th generation. The children of a memeber of the kth generation fall in the $k + 1$st generation.

A very important random variable associated with a branching process is the *first birth time* in the kth generation, B_k: this is the time at which the first member of the kth generation is born. A powerful theorem of Kingman from the theory of sub–additive stochastic processes shows that $\frac{B_k}{k} \to \gamma$ almost surely for some constant γ.

Here we would like to find the rate of convergence by giving high probability estimates on the deviation of B_k from its mean. Let us label the edges of the branching process tree as follows: The unique edge leading into a vertex labelled v is labelled with Z_v, an independent and identical copy of the random variable Z. Then, with P representing a path in the tree,

$$B_k = \min_{|P|=k} \sum_{v \in P} Z_v.$$

Thus B_k is this function of the labels attached to the edges in the binary tree on paths of length at most k from the root.

For a labelling x of the dedges, let $P^*(x)$ denote the minimising path determining B_k and set $\alpha(x)_v := x_v$ for $v \in P^*$ and 0 otherwise. Then clearly,

$$B_k(x) \leq B_k(y) + \sum_{x_v \neq y_v} \alpha(x)_v.$$

Moreover, by the result of Problem 1.9,

$$\texttt{Pr}[\texttt{Pr}[\sum_v Z_v^2 > (1+\epsilon)2k] \leq \exp\left(-4(\frac{\epsilon}{\alpha})^2 k^{1/3}\right) + ne^{-k^{1/3}}.$$

Thus, applying Theorem 6.9,

$$\texttt{Pr}[|B_k - \texttt{M}[B_k]| > t] \leq \exp\left(\frac{-t^2}{(1+\epsilon)2k}\right) + \exp\left(-4(\frac{\epsilon}{\alpha})^2 k^{1/3}\right) + ne^{-k^{1/3}}.$$

For $t := 2\epsilon k$, this gives a probability that decreases exponentially in $k^{1/3}$.

Bibliography

[1] N. Alon and J. Spencer. *The Probabilistic Method.* John Wiley and Sons, Inc., 1992.

[2] B. Bollobás. *Graph theory, an introductory course.* Springer-Verlag, 1979.

[3] B. Bollobás and I. Leader. Compression and isoperimetric inequalities. *J. Comb. Theory A*, 56:47–62, 1991.

[4] H. Chernoff. A measure of asymptotic efficiency for tests of hypothesis based on the sum of observations. *Annals of Mathematical Statistics*, 23:493–509, 1952.

[5] P-B. Miltersen D. Dubhashi and D. Ranjan. On two notions of negative dependence. 1996.

[6] D. Grable D. P. Dubhashi and A. Panconesi. Nearly-optimal, distributed edge-colouring via the nibble method. *to appear in Theoretical Computer Science*, 1997.

[7] G. Grimmett and Stirzaker D. *Probability and Random Processes.* Clarendon Press, Oxford, second edition edition, 1993.

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

[9] M. Luby. Removing randomness without a processor penalty. *Journal Comput. and Syst. Sciences*, (47):250–286, 1993.

[10] C.J.H. McDiarmid. On the method of bounded differences. In J. Siemons, editor, *Surveys in Combinatorics: Invited Papers at the 12th British Combinatorial Conference*, number 141 in London Mathematical Society Lecture Notes Series, pages 148–188. Cambridge University Press, 1989.

[11] A. Panconesi and A. Srinivasan. Randomized distributed edge coloring via an extension of the chernoff–hoeffding bounds. *SIAM J. Comput.*, 26(2):350–368, 1997.

[12] W. Pugh. Skip lists: A probabilistic alternative to balanced trees. *Communications of the ACM*, 33(6):668–676, 1990.

[13] J. Spencer. Applications of talagrand's inequality. 1996.

[14] J. M. Steele. *Probability Theory and Combinatorial Optimization*. SIAM, 1997.

[15] M. Talagrand. Concentration of measure and isoperimetric inequalities in product spaces. *Publ. Math. IHES*, 81:73–205, 1995.

[16] v. Chvátal. The tail of the hypergeometric distribution. *Discrete Mathematics*, 25(3):285–287, 1979.