Concentration of Measure for the Analysis of Randomised Algorithms

Devdatt P. Dubhashi   Alessandro Panconesi

September 7, 2006

Contents

1 Chernoff–Hoeffding Bounds 17

1.1 What is “Concentration of Measure”? 17

1.2 The Binomial Distribution 19

1.3 The Chernoff Bound 19

1.4 Heterogeneous Variables 21

1.5 The Hoeffding Extension 22

1.6 Useful Forms of the Bound 22

1.7 A Variance Bound 24

1.8 Bibliographic Notes 26

1.9 Problems 27

2 Interlude: Probabilistic Recurrences 33

3 Applying the CH-bounds 39

3.1 Probabilistic Amplification 39

3.2 Load Balancing 40

3.3 Data Structures: Skip Lists 41

3.3.1 Skip Lists: The Data Structure 41

3.3.2 Skip Lists: Randomization makes it easy 42

3.3.3 Quicksort 46

3.4 Packet Routing 47

3.5 Randomized Rounding 50

3.6 Bibliographic Notes 50

3.7 Problems 50

4 Chernoff-Hoeffding Bounds in Dependent Settings 55

4.1 Negative Dependence 55

4.2 Local Dependence 59

4.3 Janson's Inequality 61

4.4 Limited Independence 64

4.5 Markov Dependence 67

4.5.1 Definitions 67

4.5.2 Statement of the Bound 68

4.5.3 Application: Probability Amplification 68

4.6 Bibliographic Notes 70

4.7 Problems 71

5 Martingales and Azuma's Inequality 73

5.1 Review of Conditional Probabilities and Expectations 74

5.2 Martingales and Azuma's Inequality. 76

5.3 Generalizing Martingales and Azuma's Inequality. 81

5.4 The Method of Bounded Differences 83

5.5 Bibliographic Notes 87

5.6 Problems 88

6 The Method of Bounded Differences **91**

6.1 Chernoff–Hoeffding Revisited 91

6.2 Stochastic Optimization: Bin Packing 92

6.3 Game Theory and Blackwell's Approachability Theorem 92

6.4 Balls and Bins 93

6.5 Distributed Edge Colouring 94

6.5.1 Top vertices 96

6.5.2 The Difficulty with the Other Vertices 97

6.6 Problems 98

7 Averaged Bounded Differences **101**

7.1 Hypergeometric Distribution 102

7.2 Occupancy in Balls and Bins 103

7.3 Stochastic Optimization: TSP 104

7.4 Coupling 105

7.5 Distributed Edge Colouring 108

7.5.1 Preliminary Analysis 108

7.5.2 General Vertex in algorithm I 109

7.5.3 Bottom Vertex in Algorithm P 110

7.6 Handling Rare Bad Events 112

7.7 Quicksort 114

7.8 Problems 116

8 The Method of Bounded Variances **119**

8.1 A Variance Bound for Martingale Sequences 120

8.2 Applications 123
8.2.1 Bounding the Variance 125
8.2.2 Dealing with Unlikely Circumstances 127
8.3 Bibliographic Notes 130
8.4 Problems 130

9 The Infamous Upper Tail 135

9.1 Motivation: Non-Lipschitz Functions 135
9.2 Concentration of Multivariate Polynomials 136
9.3 The Deletion Method 137
9.4 Problems 139

10 Isoperimetric Inequalities and Concentration 141

10.1 Isoperimetric inequalities 141
10.2 Isoperimetry and Concentration 142
10.2.1 Concentration of Lipschitz functions 142
10.3 Examples: Classical and Discrete 144
10.3.1 Euclidean Space with Lebesgue Measure 144
10.3.2 The Euclidean Sphere 144
10.3.3 Euclidean Space with Gaussian Measure 145
10.3.4 The Hamming Cube 145
10.4 Martingales and Isoperimetric inequalities 146
10.5 Bibliographic Notes 147
10.6 Problems 148

11 Talagrand's Isoperimetric Inequality 151

11.1 Statement of the inequality 151

11.2 Hereditary Functions of of Index Sets 153

11.2.1 A General Framework 153

11.2.2 Increasing subsequences 154

11.2.3 Balls and Bins 155

11.2.4 Discrete Isoperimetric Inequalities 155

11.3 Certifiable Functions 155

11.3.1 Edge Colouring 158

11.4 The Method of Non–uniformly Bounded Differences 159

11.4.1 Chernoff–Hoeffding Bounds 160

11.4.2 Balls and Bins 161

11.4.3 Stochastic TSP 161

11.4.4 First Birth Times in Branching Processes 162

11.5 Bibliographic Notes 163

11.6 Problems 163

12 Isoperimetric Inequalities and Concentration via Transportation Cost Inequalities 167

12.1 Distance Between Probability Distributions 167

12.1.1 Distance in Product Spaces with Hamming Metric 168

12.2 TC Inequalities Imply Isoperimetric Inequalities and Concentration 169

12.3 TC Inequality in Product Spaces with Hamming Distance 171

12.3.1 One dimension 171

12.3.2 Higher dimensions 172

12.4 An Extension to Non-Product Measures 174

12.5 Problems 176

12.6 Bibliographic Notes 178

13 Quadratic Transportation Cost and Talagrand's Inequality 179

13.1 Introduction 179

13.2 Review and Roadmap 180

13.3 A L_2 (Pseudo)Metric on Distributions 182

13.3.1 One Dimension 182

13.3.2 Tensorization to Higher Dimensions 183

13.4 Quadratic Transportation Cost 183

13.4.1 Base case: One Dimension 184

13.4.2 Tensorization to Higher Dimensions 185

13.5 Talagrand's Inequality via Quadratic Transportation Cost 186

13.6 Method of Bounded Differences Revisited 187

13.7 Extension to Dependent Processes 189

13.8 Bibliographic Notes 190

13.9 Problems 190

14 Log-Sobolev Inequalities and Concentration 191

14.1 Introduction 191

14.2 A Discrete Log-Sobolev Inequality on the Hamming Cube 192

14.3 Concentration: The Herbst Argument 192

14.4 Tensorization 194

14.5 Modified Log-Sobolev Inequalities in Product Spaces 195

14.6 The Method of Bounded Differences Revisited 197

14.7 Talagrand's Inequality Revisited 198

14.8 Problems 200

14.9 Bibliographic Notes 201

Preface

The aim of this monograph is to make a body of tools for establishing concentration of measure accessible to researchers working in the design and analysis of randomized algorithms.

Concentration of measure refers to the phenomenon that a function of a large number of random variables tends to concentrate its values in a relatively narrow range (under certain conditions of smoothness of the function and under certain conditions on the dependence amongst the set of random variables). Such a result is of obvious importance to the analysis of randomized algorithms: for instance, the running time of such an algorithm can then be guaranteed to be concentrated around a pre-computed value. More generally, various other parameters measuring the performance of randomized algorithms can be provided tight guarantees via such an analysis.

In a sense, the subject of concentration of measure lies at the core of modern probability theory as embodied in the laws of large numbers, the Central Limit Theorem and, in particular, the theory of Large Deviations [15]. However, these results are asymptotic – they refer to the limit as the number of variables n, goes to infinity, for example. In the analysis of algorithms, we typically require quantitative estimates that are valid for finite (though large) values of n. The earliest such results can be traced back to the work of Azuma, Chernoff and Hoeffding in the 1950s. Subsequently there have been steady advances, particularly in the classical setting of martingales. In the last couple of decades, these methods have taken on renewed interest driven by applications in algorithms and optimization. Also several new techniques have been developed.

Unfortunately, much of this material is scattered in the literature, and also rather forbidding for someone entering the field from a Computer Science/Algorithms background. Often this is because the methods are couched in the technical language of analysis and/or measure theory. While this may be strictly necessary to develop results in their full generality, it is not needed when the method is used in computer science applications (where the probability spaces are often finite and discrete), and indeed may only serve as a distraction or barrier.

Our main goal here is to give an exposition of the basic methods for measure concentration in a manner which makes it accessible to the researcher in randomized algorithms and enables him/her to quickly start putting them to work in his/her application. Our approach is as follows:

1. Motivate the need for a concentration tool by picking an application in the form of the analysis of a randomized algorithm or probabilistic combinatorics.

2. Give only the main outline of the application, suppressing details and isolating the abstraction that relates to the concentration of measure analysis. The reader can go back to the details of the specific application following the references or the complementary book used (see below for suggestions).

3. State and prove the results not necessarily in the most general or sharpest form possible, but rather in the form that is clearest to understand and convenient as well as sufficient to use for the application at hand.

4. Return to the same example or abstraction several times using different tools to illustrate their relative strengths and weaknesses and ease of use – a particular tool works better than another in some situations, worse in others.

Other significant benefits of our exposition are: we collect and systematize the results previously scattered in the literature, explain them in a manner accessible to someone familiar with the type of discrete probability used in the analysis of algorithms and we relate different approaches to one another

Here is an outline of the book. It falls naturally into two parts. The first part contains the core "bread-and-butter" methods that we believe belong as an absolutely essential ingredient in the toolkit of a researcher in randomized algorithms today. Chapter 1 starts with the basic Chernoff–Hoeffding (CH) bound on the sum of bounded independent random variables. Many simple randomized algorithms can be analysed using this bound and we give some typical examples in Chapter 3. Since this topic is now covered in other recent books, we give only a few examples here and refer the reader to these books which can be read profitably together with this one (see suggestions below). Chapter 2 is a small interlude on probabilistic recurrences which can often give very quick estimates of tail probabilities based only on expectations. In Chapter 4, we give four versions of the CH bound in situations where the random variables are not independent – this often is the case in the analysis of algorithms and we show examples from the recent literature where such extensions simplify the analysis considerably.

The next series of chapters is devoted to a powerful extension of the CH bound to arbitrary functions of random variables (rather than just the sum) and where the assumption of independence can be relaxed somewhat. This is achieved via the concept of a *martingale*. These methods are by now rightly perceived as being fundamental in algorithmic applications and have begun to appear in introductory books such as [55], albeit very scantily, and, more thoroughly, in the more recent [**?**]. Our treatment here is far more comprehensive and nuanced, while at the same time also very accessible to the beginner. We also offer a host of relevant examples where the various methods are seen in action. As discussed below, ours can serve as a companion book for a course based on [**?**] and [**?**] and the tools developed here can be applied to give a more complete analysis of many of the examples in these books.

Chapter 5 gives an introduction to the basic definition and theory of martingales leading to the Azuma inequality. The concept of martingales, as found in probability textbooks, poses quite a barrier to the Computer Scientist who is unfamiliar with the language of filters, partitions and measurable sets from measure theory. The survey by McDiarmid [50] is the authoritative reference for martingale methods, but though directed towards discrete mathematicians interested in algorithms, is still, we feel, quite a formidable prospect for the entrant to navigate. Here we give a self-contained introduction to a special case which is sufficient for all the applications we treat and to those found in all analyses of algorithms we know of. So we are able to dispense with the measure-theoretic baggage entirely and keep to very elementary discrete probability. Chapter 6 is devoted to an inequality that is especially packaged nicely for applications, the so called Method of Bounded Differences (MOBD). This form is very easy to apply and yields surprisingly powerful results in many different settings where the function to be analyzed is ”smooth” in the sense of satisfying a Lipschitz condition. Chapter 7 progresses to a stronger version of the inequality, which we have dubbed the Method of Average Bounded Differences (MOABD) and applies in situations where the function to be analysed, while not smooth in the worst case Lipschitz sense, nevertheless satisfies some kind of ”average” smoothness property under the given distribution. This version of the method is somewhat more complicated to apply, but is essential to obtain meaningful results in many algorithms. One of the special features of our exposition is to introduce a very useful concept in probability called *coupling* and to show how it can be used to great advantage in working with the MOABD. In Chapter 8 , we give another version of the martingale method we call the Method of Bounded Variances (MOBV) which can often be used with great efficacy in certain situations.

Chapter 9 is a short interlude containing an introduction to aome recent specialized methods that were very successful in analyzing certain key problems in random graphs.

We end Part I with Chapter 10, which is an introduction to isoperimetric inequalities that are a common setting for results on the concentration of measure. We show how the MOBD is essentially equivalent to an isoperimetric inequality and this forms a natural bridge to a more powerful isoperimetric inequality in the following chapter. It is also an introduction to a method that is to come in Part II.

Part II of the book contains some more advanced techniques and recent developments. Here we systematize and make accessible some very useful tools that appear scattered in the literature and are couched in terms quite unfamiliar to computer scientists. From this (for a computer scientist) arcane body of work we distill out what is relevant and useful for algorithmic applications, using many non-trivial examples showing how these methods can be put to good use.

Chapter 11 is an introduction to Talagrand's isoperimetric theory, a theory developed in his 1995 epic that proved a major landmark in the subject and led to the resolution of some outstanding open problems. We give a statement of the inequality that is simpler, at least conceptually, than the ones usually found in the literature. Once again, the simpler statement is sufficient for all the known applications. We defer the proof of the inequality to after the methods in Part II have been developed. Instead, we focus once again on applications. We highlight two nicely packaged forms of the inequality that can be put to immediate use. Two problems whose concentration status was resolved by the Talagrand inequality are the Traveling Salesman Problem (TSP) and the Increasing subsequences problem. We give an exposition of both. We also go back to some of the algorithms analysed earlier with martingale techniques and reanalyse them with the new techniques, comparing the results for ease of applicability and strength of the conclusion.

In Chapter 12, we give an introduction to an approach from information theory via the so-called *Transportation Cost* inequalities. This approach yields very elegant proofs of isoperimetric inequalities in Chapter 10. This approach is particularly useful since it can handle certain controlled dependence between the variables. Also Kati Marton has shown how it can be adapted to prove inequalities that imply the Talagrand isoperimetric inequality, and we give an account of this in Chapter 13. In Chapter 14, we give an introduction to another approach from information theory that leads to concentration inequalities – the so-called *Entropy* method or *Log-Sobolev* inequalities. This approach also yields short proofs of Talagrand's inequality, and we also revisit the method of bounded differences in a different light.

How to use the book

The book is, we hope, a self-contained, comprehensive and quite accessible resource for any person with a typical computer science or mathematics background who is interested in applying concentration of measure methods in the design and analysis of randomized algorithms.

This book can also be used as a textbook in an advanced course in randomized algorithms (or related courses) as a supplement and complement with some well established textbooks. For instance, we recommend using it for a course in

Randomized Algorithms together with the books

- R. Motwani and P. Raghavan, *Randomized Algorithms*, Cambridge University Press 1995.
- M. Mitzenmacher and E. Upfal, *Probability and Computing*, Cambridge University Press, 2005.

Probabilistic Combinatorics together with the classic

- N. Alon and J. Spencer, *The Probabilistic Method*, Second edition, John Wiley 2000.

Graph Coloring together with the book

- M. Molloy and B. Reed, *Graph Coloring and the Probabilistic Method*, Springer 2002.

Random Graphs together with the book:

- S. Janson, T. Luczak and A. Rucinski, *Random Graphs*, Wiley 2000.

Large Deviation Theory together with the book

- F. den Hollander, *Large Deviations*, Fields Institute Monograph, American Mathematical Society 2000.

Acknowledgements

Fred Kochman, Bernd Gartner, Johan Hastad and student, Karonski, HAMMER!

Chapter 1

Chernoff–Hoeffding Bounds

[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 1.8 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

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

This explains the name as the function $\mathtt{E}[e^{\lambda X}]$ 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 $\boldsymbol{p} := (p_1, \ldots, p_n)$ and $\boldsymbol{q} := (q_1, \ldots, q_n)$ on a space of cardinality n, the *relative entropy distance* between them, $H(\boldsymbol{p}, \boldsymbol{q})$ is defined by [1]:

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

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 $\mathtt{Pr}[X < (p-t)n]$.

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 1.8 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 $\mathtt{E}[e^{\lambda X}]$ exactly as in (1.3) except for the last line to get:

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

Recall that the arithmetic–geometric mean inequality states that

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

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

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

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

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.

1.5 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

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

It would be useful to write the final bound

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 1.12).

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 1.12 ∎

Often we would like to apply the bopunds above to a sum $\sum_i X_i$ where we do not know the exact values of the expectations $\mathtt{E}[X_i]$ but only upper or lower bounds on it. In such situations, one can neverthelesss apply the CH bounds with the known bounds instead as you should verify in the following exercise.

Exercise 1.2 *Suppose* $X := \sum_{i=1}^n X_i$ *as in Theorem 1.1 above, and suppose* $\mu_L \leq \mu \leq \mu_H$. *Show that*

(a) For any $t > 0$,

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

(b) For $0 < \epsilon < 1$,

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

say which chapter

You may need to use the following useful and intuitively obvious fact that we will prove in a later chapter. Let $X_1, \cdots, X_n$ *be independent random variables distributed in* $[0,1]$ *with* $\mathtt{E}[X_i] = p_i$ *for each* $i \in [n]$. *Let* $Y_1, \cdots, Y_n$ *and* $Z_1, \cdots, Z_n$ *be independent random variables with* $\mathtt{E}[Y_i] = q_i$ *and* $\mathtt{E}[Z_i] = r_i$ *for each* $i \in [n]$. *Now suppose* $q_i \leq p_i \leq r_i$ *for each* $i \in [n]$. *Then, if* $X := \sum_i X_i, Y := \sum_i Y_i$ *and* $Z := \sum_i Z_i$, *for any* t,

$$\mathtt{Pr}[X > t] \leq \mathtt{Pr}[Z > t], \quad \textit{and} \quad \mathtt{Pr}[X < t] \leq \mathtt{Pr}[Y < t].$$

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:

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

for $t < 2\sigma^2 / \max_i \max(\mu_i, 1 - \mu_i)$.

Exercise 1.3 *Check that for random variables distributed in* $[0, 1]$, *this is of the same form as the CH bound derived in the previous section upto constant factors in the exponent. You may need to use the fact that for a random variable distributed in the interval* $[a, b]$, *the variance is bounded by* $(b - a)^2/4$.

The following bound is often referred to as Bernstein's inequality:

Theorem 1.4 (Bernstein's inequality) *Let the random variables* $X_1, \cdots, X_n$ *be independent with* $X_i - \mathtt{E}[X_i] \leq b$ *for ach* $i \in [n]$. *Let* $X := \sum_i X_i$ *and let* $\sigma^2 := \sum_i \sigma_i^2$ *be the variance of* X. *Then, for any* $t > 0$,

$$\mathtt{Pr}[X > \mathtt{E}[X] + t] \leq \exp\left(-\frac{t^2}{2\sigma^2(1 + bt/3\sigma^2)}\right).$$

Exercise 1.5 *Check that for random variables in* $[0, 1]$ *and* $t < 2\sigma^2/b$, *this is roughly the same order bound as we derived above.*

In typical applications, the "error" term $bt/3\sigma^2$ will be negligible. Suppose the random variables $X_1, \cdots, X_n$ have the same bounded distribution with positive variance c^2, so $\sigma^2 = nc^2$. Then for $t = o(n)$, this bound is $\exp\left(-(1+o(1))\frac{t^2}{2\sigma^2}\right)$ which is consistent with the Central Limit Theorem assertion that in the asymptotic limit, $X - \mathtt{E}[X]$ is normal with mean 0 and variance σ^2.

Exercise 1.6 *Let $X := \sum_i X_i$ where the $X_i, i \in [n]$ are i.i.d with $\mathtt{Pr}[X_i = 1] = p$ for each $i \in [n]$ for some $p \in [0,1]$. Compute the variance of X and apply and compare the two bounds above as well as the basic CH bound. Check that when $p = 1/2$, all these bounds are roughly the same.*

1.8 Bibliographic Notes

The original technique is from Chernoff [11] 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 [27]. Our presentation was much influenced by [48]. The quick derivation in Problems 1.8 and 1.9 are due to V. Chvátal [12].

1.9 Problems

Problem 1.7 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 1.8 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 1.9 [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 1.10 Show that for $0 < \alpha \leq 1/2$,

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

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

Problem 1.11 [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 $p_n(x) := \sum_{0 \leq i \leq n} \binom{n}{i} x^i (1-x)^{n-i} f(i/n)$.) ▽

Problem 1.12 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 $\epsilon > 0$,

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

(a) Compare this bound to the one obtained by setting $t := \epsilon \mathtt{E}[X]/n$ in the relative entropy bound derived in (1.4).
(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

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

▽

Problem 1.13 Let $X_1, \cdots, X_n$ be random variables bounded in $[0,1]$ such that for each $i \in [n]$,

$$\mathtt{E}[X_i \mid X_1, \cdots, X_{i-1}] \leq p_i.$$

Show that in this case, the upper tail for $\sum_i X_i$ can be upper bounded by the upper-tail CH-estiamte for an independent set of variables $X'_1, \cdots, X'_n$ with $\mathtt{E}[X'_i] = p_i$. Formulate and prove a symmetric condition for the lower tail. ▽

begin new

Problem 1.14 Let $X_1, \ldots, X_n$ be a set of binary random variables satisfying the condition

$$\Pr\left[\bigwedge_{i \in S} X_i = 1\right] \leq \prod_{i \in S} \Pr\left[X_i = 1\right]$$

for all subsets S. Prove that under this condition the Chernoff bound holds for $X = \sum_i X_i$. ▽

Problem 1.15 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:

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

▽

Problem 1.16 [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 $i \in [n]$,

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

Hence

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

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 $\texttt{E}[Z'] \leq \texttt{E}[Z]$. Second, that since for each $i \in [n]$, $\texttt{Pr}[Z_i > n^\beta] \leq 1 - F(n^\beta) = e^{-n^\beta}$,

$$\texttt{Pr}[\bigwedge_i Z_i = Z'_i] \geq 1 - ne^{-n^\beta}.$$

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 1.15. Hence,

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

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

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

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

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

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

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

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

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

Hence,

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

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

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

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

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

This is a much better bound than that achieved by truncation. △

Problem 1.17 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 1.18 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 1.19 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 $\boldsymbol{p}$. Let $B(n, \boldsymbol{p})$ denote the sumber of successes in n independent trials where the probability of success at the ith trial is p_i. Let

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

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

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

Let

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

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 $\boldsymbol{p}^*(\lambda), \hat{\boldsymbol{p}}(\lambda) \in D(\lambda)$.
(a) Show that for any $\boldsymbol{p} \in D(\lambda)$,

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

and

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

(b) More generally, let $\boldsymbol{p}, \boldsymbol{p}' \in D_\lambda$ be such that there is a doubly stochastic matrix Π with $\boldsymbol{p}' = \boldsymbol{p}\Pi$. 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 $\boldsymbol{p}$ is said to **majorise** the vector $\boldsymbol{p}'$. Show that

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

and

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

Verify that this generalises part (a). ▽

Chapter 2

Interlude: Probabilistic Recurrences

Karp[33] developed an attractive framework for the analysis of randomized algorithms. Suppose we have a randomized algorithm that on input x, performs "work" $a(x)$ and then produces a subproblem of size $H(x)$ which is then solved by recursion. One can analyze the performance of the algorithm by writing down a "recurrence":

$$T(x) = a(x) + T(H(x)). \tag{2.1}$$

Superficially this looks just the same as the usual analysis of algorithms via recurrence relations. However, the crucial difference is that in contrast with deterministic algorithms, the size of the subproblem produced here, $H(x)$ is a random variable, and so (2.1) is a *probabilistic recurrence* equation.

What does one mean by the solution of such a probabilistic recurrence? The solution $T(x)$ is itself a random variable and we would like as much information about its distribution as possible. While a complete description of the exact distribution is usually neither possible nor really necessary, the "correct" useful analogue to the deterministic solution is a concentration of measure result for $T(x)$. Of course, to do this, one needs some information on the distribution of the subproblem $H(x)$ generated by the algorithm. Karp gives a very easy–to–apply framework that requires only the bare minimum of information on the distribution of $H(x)$, namely (a bound on) the expectation, and yields a concentration result for $T(x)$. Suppose that in (2.1), we have $\mathtt{E}[H(x)] \leq m(x)$ for some function $0 \leq m(x) \leq x$. Consider the "deterministic" version of (2.1) obtained by replacing the random variable $H(x)$ by the deterministic bound $m(x)$:

$$u(x) = a(x) + u(m(x)). \tag{2.2}$$

The solution to this equation is $u(x) = \sum_{i \geq 0} a(m^i(x))$, where $m^0(x) := 0$ and $m^{i+1}(x) = m(m^i(x))$. Karp gives a concentration result around this value $u(x)$:

Theorem 2.1 (Karp's First Theorem) *Suppose that in (2.1), we have* $\mathtt{E}[H(x)] \leq m(x)$ *for some function* $0 \leq m(x) \leq x$ *and such that* $a(x), m(x), \frac{m(x)}{x}$ *are all non–decreasing. Then*

$$\mathtt{Pr}[T(x) > u(x) + ta(x)] \leq \left(\frac{m(x)}{x}\right)^t.$$

We have stated the result in the simplest memorable form that captures the essence and is essentially correct. However, technically the statement of the theorem above is actually not quite accurate and we have omitted some continuity conditions on the functions involved. These conditions usually hold in all cases where we'd like to apply the theorem. Moreover, as shown in [10], some of these conditions can be discarded at the cost of only slightly weakening the bound. For instance, we can discard the condition that $\frac{m(x)}{x}$ is non–decreasing; in that case, the bound on the right hand side can be essentially replaced by $\left(\max_{0 \leq y \leq x} \frac{m(y)}{y}\right)^t$

Also, in the formulation above, we assumed that the distribution of $H(x)$, the size of the derived subproblem depends only on the input size x. Karp[33] gives a more general formulation where the subproblem is allowed to depend on the actual input instance. Suppose we have a "size" function s on inputs, and on processing an input z, we expend work $a(s(z))$ and get a subproblem $H(z)$ such that $\mathtt{E}[s(H(z))] \leq m(s(z))$. The probabilistic recurrence is now

$$T(z) = a(s(z)) + T(H(z)).$$

By considering $T'(x) := \max_{s(z)=x} T(z)$, one can bound this by a recurrence of the earlier form and apply the Theorem to give exactly the same solution. Thus we can apply the Theorem *per se* even in this more general situation.

We illustrate the ease of applicability of this cook–book style recipe by some examples (taken from Karp's paper).

Example 2.2 [Selection] Hoare's classic algorithm for finding the kth smallest element in a n–element set S, proceeds as follows: pick a random element $r \in S$ and by comparing each element in $S \setminus r$ with r, partition $S \setminus r$ into two subsets $L := \{y \in S \mid y < r\}$ and $U := \{y \in S \mid y > r\}$. Then,

- If $|L| \geq k$, recursively find the kth smallest element in L.
- If $|L| = k-1$, then return r.
- If $|L| < k-1$, then recursively find the $k-1-|L|$th smallest element in U.

The partitioning step requires $n-1$ comparisons. It can be shown that the expected size of the subproblem, namely the size of L or U is at most $3n/4$, for all k. Thus Karp's Theorem can be applied with $m(x) = 3x/4$. We compute $u(x) \leq 4x$. Thus, if $T(n,k)$ denotes the number of comparisons performed by the algorithm, we have the following concentration result: for all $t \geq 0$,

$$\texttt{Pr}[T(n,k) > 4n + t(n-1)] \leq \left(\frac{3}{4}\right)^t .$$

This bound is nearly tight as showed by the following simple argument. Define a *bad* splitter to be one where $\frac{n}{|U|} \geq \log\log n$ or $\frac{n}{|L|} \geq \log\log n$. The probability of this is greater than $\frac{2}{\log\log n}$. The probability of picking $\log\log n$ consecutive bad splitters is $\Omega(\frac{1}{(\log n)^{\log\log\log n}})$. The work done for $\log\log n$ consecutive bad splitters is

$$n + n\left(1 - \frac{1}{\log\log n}\right) + n\left(1 - \frac{1}{\log\log n}\right)^2 + \ldots n\left(1 - \frac{1}{\log\log n}\right)^{\log\log n}$$

which is $\Omega(n \log\log n)$. Compare this with the previous bound using $t = \log\log n$. $\bigtriangledown$

Example 2.3 [Luby's Maximal Independent Set Algorithm] Luby[41] gives a randomized parallel algorithm for constructing a maximal independent set in a graph. The algorithm works in stages: at each stage, the current independent set is augmented and some edges are deleted form the graph. The algorithm terminates when we arrive at the empty graph. The work performed at each iteration is equal to the number of edges in the current graph. Luby showed that at each stage, the expected number of edges deleted is at least one–eighth of the number of edges in the complete graph. If $T(G)$ is the number of stages the algorithm runs and $T'(G)$ is the total amount of work done, then we get the concentration results:

$$\texttt{Pr}[T(G) > \log_{8/7} n + t] \leq \left(\frac{7}{8}\right)^t ,$$

$$\texttt{Pr}[T'(G) > (8+t)n] \leq \left(\frac{7}{8}\right)^t .$$

$\bigtriangledown$

Example 2.4 [Tree Contraction] Miller and Reif [52] give a randomized *tree contraction* algorithm that starts with a n node tree representing an arithmetic expression and repeatedly applies a randomized contraction operation that provides a new tree representing a modified arithmetic expression.The process eventually reaches a one node tree and terminates. The work performed in the contraction step can be taken to be proportional to the number of nodes in the tree. Miller and Reif show that when applied to a tree with n nodes, the contraction step results in a tree of size at most $4n/5$. However the distribution of the size may depend on the original tree, not just the original size. Define the size function here to be the number of nodes in the tree in order to apply the more general framework. Let $T(z), T'(z)$ denote the number of iterations and the total work respectively when the contraction algorithm is applied to tree z. Then, Karp's Theorem gives the measure concentration results:

$$\texttt{Pr}[T(z) > \log_{5/4} n + t] \leq (4/5)^t,$$

and

$$\texttt{Pr}[T'(z) > (5+t)n] \leq (4/5)^t.$$

▽

Under the weak assumptions on the distribution of the input, Karp's First Theorem is essentially tight. However, if one has additional information on the distribution of the subproblem, say some higher moments, then one can get sharper results which will be explored below in § **??**.

Karp also gives an extension of the framework for the very useful case when the algorithm might generate more than one subproblem. Suppose we have an algorithm that on input x performs work $a(x)$ and then generates a fixed number $k \geq 1$ sub–problems $H_1(x), \ldots, H_k(x)$ each a random variable. This corresponds to the probabilistic recurrence:

$$T(x) = a(x) + T(H_1(x)) + \cdots + T(H_k(x)). \tag{2.3}$$

To obtain a concentration result in this case, Karp uses a different method which requires a certain condition:

Theorem 2.5 (Karp's Second Theorem) *Suppose that in (2.3), we have that for all possible values $(x_1, \ldots, x_k)$ of the tuple $(H_1(x), \ldots, H_k(x))$, we have*

$$\texttt{E}[T(x)] \geq \sum_i \texttt{E}[T(x_i)]. \tag{2.4}$$

Then, we have the concentration result: for all x and all $t > 0$,

$$\texttt{Pr}[T(x) > (t+1)\texttt{E}[T(x)]] < e^{-t}.$$

The condition (2.4) says that the expected work in processing *any* sub–problems that can result from the original one can never exceed the expected cost of the processing the original instance. This is a very strong assumption and unfortunately, in many cases of interest, for example in computational geometry, it does not hold. Consequently the theorem is somewhat severely limited in its applicability. A rare case in which the condition is satisfied is for

Example 2.6 [Quicksort] Hoare's Quicksort algorithm is a true classic in Computer Science: to sort a set S of n items, we proceed as in the selection algorithm from above: select a random element $r \in S$ and by comparing it to every other element, partition S as into the sets L of elements less than x and U, the set of elements at least as big as r. Then, recursively, sort L and U. Let $Q(n)$ denote the number of comparisons performed by Quicksort on a set of n elements. Then $Q(n)$ satisfies the probabilistic recurrence:

$$T(n) = n - 1 + Q(H_1(n)) + Q(H_2(n)),$$

where $H_1(n) = |L|$ and $H_2(n) = |U|$. For Quicksort we have "closed-form" solutions for $q_n := \mathtt{E}[Q(n)]$which imply that $q_n \geq q_i + q_{n-i-1} + n - 1$ for any $0 \leq i < n$, which is just the condition needed to apply Karp's Second Theorem. Thus we get the concentration result:

$$\mathtt{Pr}[Q(n) > (t+1)q_n] \leq e^{-t}.$$

Actually one can get a much stronger bound by applying Karp's First Theorem suitably! Charge each comparison made in Quicksort to the non–pivot element, and let $T(\ell)$ denote the number of comparisons charged to a fixed element when Quicksort is applied to a list ℓ. Use the natural size function $s(\ell) := |\ell|$, which gives the number of elements in the list. Then we have the recurrence, $T(\ell) = 1 + T(H(\ell))$, where $s(H(\ell) = |\ell|/2$ since the sublist containing the fixed element (when it's not the pivot) has size uniformly distributed in $[0, |\ell|]$. So applying Karp's First Theorem, we have that for $t \geq 1$,

$$\mathtt{Pr}[T(\ell) > (t+1)\log|\ell|] \leq (1/2)^{t \log |\ell|} = |\ell|^{-t}.$$

Thus any fixed element in a list of n elements is charged at most $(t+1)\log n$ comparisons with probability at least $1 - n^{-t}$. The total number of comparisons is therefore at most $(t+1)n \log n$ with probability at least $1 - n^{t-1}$.

This is an inverse polynomial concentration bound. In a later section we shall get a somewhat stronger and provably optimal bound on the concentration. ▽

It would naturally be of great interest to extend the range of Karp's Second Theorem by eliminating the restrictive hypothesis. For instance, it would be of interest to extend the Theorem under the kind of assumptions in Karp's First Theorem.

Chapter 3

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 usefulness of these bounds.

3.1 Probabilistic Amplification

The following situation is quite common. We have a probabilistic algorithm that, on input x, computes the correct answer $f(x)$ with probability strictly greater than $\frac{1}{2}$. For concreteness, let us assume that the success probability is $p \geq \frac{3}{4}$ and that the algorithm has two possible outcomes, 0 and 1. To boost our confidence we run the algorithm n times and select the majority answer. What is the probability that this procedure is correct?

Let X be the number of occurrences of the majority value. Then, $\mathtt{E}[X] = pn > \frac{3}{4}n$. The majority answer is wrong if and only if $X < \frac{n}{2}$. Note that here we do not know the exact value of $\mathtt{E}[X]$, but only an upper bound. In our case we have n independent trials X_i, each of which succeeds with probability $p \geq \frac{3}{4}$. Using the fact noted in Exercise 1.2, one can apply the CH bound directly. Recalling (1.6), if we set $t := \frac{n}{4}$, we have that

$$\Pr\left[X < \frac{n}{2}\right] \leq e^{-n/8}.$$

The reader can check that (1.7) yields worse estimates. Problem 3.4 asks to generalize this to the case when the algorithm takes values in an infinite set.

3.2 Load Balancing

Suppose we have a system in which m jobs arrive in a stream and need to be processed immediately on one of a collection of n identical processors. We would like to assign the jobs to the processors in a manner that *balances* the workload evenly. Furthermore, we are in a typical *distributed* setting where centralized coordination and control is impossible. A natural "light-weight" solution in such a situation is to assign each incoming job to a processor chosen uniformly at random, indepndently of other jobs. We analyse how well this scheme performs.

Focus on a particular processor. let $X_i, i \in [m]$ be th eindicator variable for whether job number i is assigned to this processor. The total load of the processor is then $X := \sum_i X_i$. Note that $\texttt{Pr}[X_i = 1] = 1/n$ bacuse each job is assigned to a processor chosen uniformly at random. Also, $X_1, \cdots X_m$ are independent.

First let us consider the case when the $m = 6n \ln n$. Then $\texttt{E}[X] = \sum_i \texttt{E}[X_i] = m/n = 6 \ln n$. Applying (1.7), we see that the probability of the processor's load exceeding $12 \ln n$ is at most

$$\texttt{Pr}[X > 12 \ln n] \leq e^{-2 \ln n} \leq 1/n^2.$$

Applying the union bound, we see that the load of no processor exceeds $6 \ln n$ with probability at least $1 - 1/n$.

Next, let us consider the case when $m = n$. In this case, $\texttt{E}[X] = 1$. Applying (1.8), we see that

$$\texttt{Pr}[X > 2 \log n] \leq 2^{-2 \log n} \leq 1/n^2.$$

Applying the union bound, we see that the load of no processor exceeds $2 \log n$ with probability at least $1 - 1/n$.

However, in this case, we can tighten the analysis using the bound in (1.12):

$$\texttt{Pr}\left[X \geq (1+\epsilon)\texttt{E}[X]\right] \leq \left(\frac{e^{\epsilon}}{(1+\epsilon)^{(1+\epsilon)}} \right)^{\texttt{E}[X]}.$$

Set $(1+\epsilon) := c$, then

$$\texttt{Pr}[X > c] < \frac{e^{c-1}}{c^c} \tag{3.1}$$

To pick the appropriate c to use here, we focus on the function x^x. What is the solution to $x^x = n$? Let $\gamma(n)$ denote this number. There is no closed form expression for $\gamma(n)$ but one can approximate it well. If $x^x = n$, taking logs gives $x \log x = \log n$, and taking logs once more gives $\log x + \log\log x = \log\log n$. Thus,

$$2 \log x > \log x + \log\log x = \log\log n >> 766 \log x.$$

Using this to divide throughout the equation $xlogx = \log n$ gives

$$\frac{1}{2}x \leq \frac{\log n}{\log\log n} \leq x = \gamma(n).$$

Thus $\gamma(n) = \Theta(\frac{\log n}{\log\log n})$.

Setting $c := e\gamma(n)$ in (3.1), we have:

$$\texttt{Pr}[X > c] < \frac{e^{c-1}}{c^c} < \left(\frac{e}{c}\right)^c = \left(\frac{1}{\gamma(n)}\right)^{e\gamma(n)} < \left(\frac{1}{\gamma(n)}\right)^{2\gamma(n)} = 1/n^2.$$

Thus the load of any one processor does not exceed $e\gamma(n) = \Theta(\log n/\log\log n)$ with probability at least $1 - 1/n^2$. Applying the Union bound, we conclude that with probability at least $1 - 1/n$, the load of no processor exceeds this value. It can be shown that this analysis is tight – with high probability some processor does receive $\Theta(\log n/\log\log n)$ jobs.

3.3 Data Structures: Skip Lists

The second example concerns the design and analysis of data structures. We shall discuss a useful data structure known as Skip List.

3.3.1 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 0th 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

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

one in Figure 0.1. This structure resembles a binary tree and likewise allows for 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.

3.3.2 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

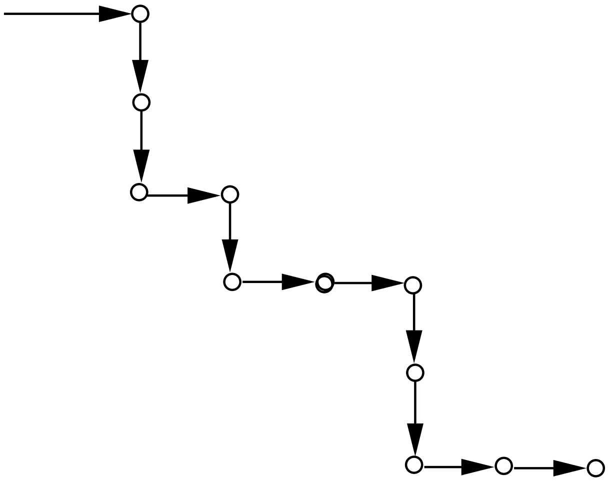

Figure 0.2: A zig-zag path through the data structure generated by a search.

which x_i belongs, then

$$\Pr[H_i = k] = p^k(1-p). \tag{3.2}$$

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 upper-bounded by the *height* of the data structure, defined as

$$H := \max_i H_i. \tag{3.3}$$

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 sizable large deviations are unlikely.

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

Proposition 3.1 $\Pr[H > a \log n] \leq n^{-a+1}$, *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:

$$\texttt{Pr}[H_i = k] = p^k q, \quad k \geq 0. \tag{3.4}$$

Hence for $\ell \geq 1$,

$$\begin{aligned} \texttt{Pr}[H_i > \ell] &= \sum_{k>\ell} \texttt{Pr}[H_i = k] \\ &= \sum_{k>\ell} p^k q \\ &= p^{\ell+1}. \end{aligned} \tag{3.5}$$

The height of the skip list, H is given by

$$H = \max_i H_i. \tag{3.6}$$

Hence,

$$\begin{aligned} \texttt{Pr}[H > \ell] &= \texttt{Pr}[\bigvee_i H_i > \ell] \\ &\leq \sum_i \texttt{Pr}[H_i > \ell] \\ &= np^{\ell+1}. \end{aligned} \tag{3.7}$$

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

$$\texttt{Pr}[H > a \log_p n] \leq n^{-a+1}. \tag{3.8}$$

■

Refer now to Figure 0.2. The cost of a traversal is equal to the number of ↓'s plus the number of →'s. If an ↓-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 → it means that when x flipped its coin at level L_k the outcome was TAIL. We label each ↓ with p– denoting *success*– and each → 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 3.2 $\Pr(W(k,p) \leq m) = \Pr(B(m,p) \geq k)$.

Proof. See Exercise 3.6. ∎

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 ↓'s and →'s of a traversal. By Proposition 3.2

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

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 3.1, and setting $a = 2$, $b = 8$, and $p = 1/2$

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

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.

3.3.3 Quicksort

The randomized version of well-known algorithm quicksort is one of, if not "the" most effective sorting algorithm. The input of the algorithm is an array

$$X := [x_1, \ldots, x_n]$$

of n numbers. The algorithm selects an element at random, the so-called *pivot*, denoted here as p, and partitions the array as follows,

$$[y_1, \ldots, y_i, p, z_1, \ldots, z_j]$$

where the ys are less than or equal to p and the zs are strictly greater (one of these two regions could be empty). The algorithm continues with two recursive calls, one on the y-region and the other on the z-region. The end of the recursion is when the input array has less than two elements.

We want to show that the running time of the algorithm is $O(n \log n)$ with probability at least $1 - \frac{1}{n}$. The overall running time is given by the tree of recursive calls. The tree is binary, with each node having at most two children corresponding to the y- and the z-region obtained by partitioning. Since partitioning requires linear time, if we start with an array of n elements, the total work done at every level of the tree is $O(n)$. Therefore to bound the running time it suffices to compute the height of the tree. We will show that, for any leaf, the length of the path from the root to the leaf is at most $4 \log_2 n$, with probability at least

$1 - \frac{1}{n^2}$. The claim will then follow from the union bound, since in the tree there are at most n nodes. Denoting with P a generic path from the root to a leaf,

$$\Pr[\exists P, |P| > 4\log_2 n] \leq n \; Pr[|P| > 4\log_2 n] \leq \frac{1}{n}.$$

We now bound the probability that a path has more than $4\log_2 n$ nodes. Call a node *good* if the corresponding pivot partitions the array into two regions, each of size at least $\frac{1}{3}$ of the array. The node is *bad* otherwise. If a path contains t good nodes the size of the array decreases as

$$s_t \leq \frac{2}{3} s_{t-1} \leq \left(\frac{2}{3}\right)^t n.$$

It follows that there can be at most

$$t = \frac{\log_2 n}{\log_2 \frac{3}{2}} < 2\log_2 n$$

good nodes in any path. We now use the Chernoff-Hoeffding bounds to show that

$$\Pr[|P| > 2\log_2 n] < \frac{1}{n^2}.$$

Let $\ell := |P|$ and let X_i be a binary random variable taking the value 1 if node i is bad, and 0 if it is good. The X_is are independent and such that $\Pr[X_i = 1] = \frac{1}{3}$. Thus $X := \sum_{i \in P} X_i$ is the number of bad nodes in the path P and $\mathtt{E}[X] = \ell/3$. Recalling (1.8), for $t > 2e\ell/3$,

$$\Pr[X > t] \leq 2^{-t} \leq \frac{1}{n^2}$$

provided that

$$\ell \geq \frac{3}{e}\log_2 n.$$

Therefore the total number of good and bad nodes along any path does not exceed $4\log_2 n$ with probability at least $1 - \frac{1}{n}$. By fiddling with the constant it is possible to show that the running time of randomized quicksort is $O(n \log n)$ with probability at least $1 - \frac{1}{n^k}$, for any fixed k. In Chapter 7 we will derive a stronger bound by using martingale methods.

3.4 Packet Routing

Packet routing is a fundamental problem in the context of parallel and distributed computation. The following set up, following [36], captures many of its combinatorial intricacies. The underlying communication network is modelled as a simple graph $G = (V, E)$ with $|V| = n$ and $|E| = m$. In the network there are N packets $p_1, \ldots, p_N$. Each packet p is to follow its route r_p from its source to its destination. The goal is that of finding a routing algorithm, or *schedule*, that minimizes the time to deliver all packets. The key constraint is that if two or more packets are to traverse an edge e at time t, only one packet can traverse it (packets queue up at edge endpoints if necessary, and the algorithm must decide which goes on next).

It is instructive to look for a lower bound for completion time. Since for every time unit each packet can traverse at most one edge, the following quantity, called the *dilation*, is a lower bound:

$$d := \max_p |r_p|.$$

A second trivial lower bound is the so-called *congestion*. Given an edge e, let P_e denote the set of packets that must traverse e to reach their destination. Then,

$$c := \max_e |P_e|$$

is a lower bound for completion time in the worst case. A trivial upper bound is then then $c \cdot d$ time units. Unless care is exercised in the routing policy the schedule *can* be as bad. A remarkable result states that, for every input, there is always a schedule that takes only $O(c + d)$ steps **??**. In what follows we exhibit a very simple schedule and show, using the Chernoff bounds, that it delivers all packets in $O(c + d \log(mN))$ steps with high probability. The basic idea is for each packet to pick a random delay $\rho \in [r]$ and then start moving. The maximum congestion possible at an edge e is bounded by $|P_e| \leq c$. If r were large enough, a random delay would ensure that, with high probability, for every edge e and any given time t, the queue for e at time t would consist of at most one packet. The resulting schedule would then deliver all packets within $O(r + d)$ steps. For this to work however, r must be too large.

schedule is computable?

Exercise 3.3 *How large must r be so that all packets are delivered within $r + d$ steps with probability at least $1 - \frac{1}{n}$?*

A way out is to accept to have more than one packet per edge at any given time, but to keep this number always below a certain maximum b. If congestion for any edge at any time is at most b, we can route all packets within $b(r + d)$ steps using the following simple algorithm,

- Pick a random delay $\rho \in [r]$ uniformly at random, independently of other packets.

- Traverse each edge of the path from source to destination using b time units for every edge.

Note that time is grouped in macro-steps of b time units each, and that every packet uses a full macro-step to traverse an edge. I.e. a packet p traverses teh kth edge of its route r_p at a time t in the interval $[(k-1)b+1, kb]$. Since queues never have more than b packets every packet will have a chance to traverse an edge within a macro-step. The time bound follows. We will show that the parameters r and b can be chosen so that $b(r+d) = O(c + d\log(mN))$.

For the analysis, fix an edge e and a macro-step s, and let X_{es} denote the number of packets that queue up at e at macro step s when following the above algorithm. We can decompose X_{es} as the sum of indicator random variables X_{pes} for every packet $p \in P_e$, where X_{pes} indicates if p queues up at e at macro step s or not. Thus,

$$X_{es} = \sum_{p \in P_e} X_{pes}$$

and

$$\mathtt{E}[X_{es}] = \sum_{p \in P_e} \mathtt{E}[X_{pes}] \leq \frac{c}{r}.$$

The bound on the expectation follows, since each packet picks a random delay $\rho \in [r]$ uniformly at random and $|P_e| \leq c$. Note that, for the same e and s, the X_{pes}'s are independent and we can therefore apply the Chernoff bounds. Since we do not know the expectation, but only an upper bound, we make use of the bound developed in Exercise (1.2) with $\mu_H = \frac{c}{r}$ and get,

$$\Pr(X > (1+\epsilon)c/r) \leq \exp\left\{-\epsilon^2 c/3r\right\}.$$

For definiteness, let $\epsilon = \frac{1}{2}$. We define

$$b := \frac{3}{4}\frac{c}{r}$$

and

$$r := \frac{c}{12\alpha\log(mN)}$$

so that

$$\Pr(X_{es} > b) \leq \frac{1}{(mN)^\alpha}.$$

Let E be the event that some edge has more than b queuing packets at some macro-step. Since there are m edges and $r \leq c \leq N$ macro-steps, we can bound $\Pr[E]$ using the union bound,

$$\Pr[E] \leq \sum_{e,s} \Pr(X_{es} > b) \leq mN\frac{1}{(mN)^\alpha} \leq \frac{1}{(mN)^{\alpha-1}}.$$

By choosing $\alpha = 2$ the probability that no edge ever has more than b queuing packets is at least $1 - \frac{1}{mN}$. Assuming this, the total time needed to deliver all packets is at most

$$b(r + d) = O(c + d \log(mN))$$

as claimed.

3.5 Randomized Rounding

Work in progress...

3.6 Bibliographic Notes

Skiplists are an invention of W. Pugh [59]

3.7 Problems

Problem 3.4 A randomized algorithm A, on input x, gives an answer $A(x)$ that is correct with probability $p > \frac{3}{4}$. $A(x)$ takes values in the set of natural numbers. Compute the probability that the majority outcome is correct when the algorithm is run n times. How large n must be to have a 0.99 confidence that the answer is correct? ▽

Problem 3.5 The following type of set systems is a crucial ingredient in the construction of pseudo-random generators [56]. Given a universe $\mathcal{U}$ of size $|\mathcal{U}| = cn$ a family $\mathcal{F}$ of subsets of $\mathcal{U}$ is a *good family* if (a) all sets in $\mathcal{F}$ have n elements; (b) given any two sets A and B in $\mathcal{F}$ their intersection has size at most $\frac{n}{2}$; and, (c) $|\mathcal{F}| = 2^{\Theta(n)}$.

Show that there is a value of c for which good families exists for every n (Hint: partition the universe into n blocks of size c and generate sets of n elements independently at random by choosing elements randomly in each block. Then compute the probability that the family generated in this fashion has the desired properties.) ▽

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

I'd like to insert a two lines comment here required to obtain a "heads" with a coin of bias p (i.e. $\texttt{Pr}(\text{heads}) = p, \texttt{Pr}(\text{tails}) = 1 - p =: q$). Note that $\texttt{Pr}[W(1,p) = \ell] = q^{\ell-1}p$, 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.*

Problem 3.6 Prove that $\Pr(W(k,p) \leq m) = \Pr(B(m,p) \geq k)$. ▽

This is actually an exercise...

Problem 3.7 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 $1/p$?

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

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

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

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

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

(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

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

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

Solution. Work in progress... △

Problem 3.9 Prove a sharp concentration result for the space used by a skip list. ▽

Solution. The space used by the data structure is

$$S_n = \sum_{i \in [n]} H_i \tag{3.9}$$

In view of (3.2), 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 3.2 (or, equivalently, Problem 3.6), 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 3.10 What is the best value of p in order to minimize the expected time of a search operation in a skip list? ▽

begin new

Problem 3.11 In this exercise we deal with a very elegant data structure called *treaps* (see for instance [38, 55]). A treap is a binary tree whose nodes contain two values, a *key* x and a *priority* p_x. The keys are drawn from a totally ordered set and the priorities are given by a random permutation of the keys. The tree is a heap according to the priorities and it is a search tree according to the keys (i.e. keys are ordered in in-order).

(a) Show that given a totally ordered set X of elements and a function p assigning unique priorities to elements in X, there always exists a unique treap with keys X and priorities p.

Treaps allow for fast insertion, deletion and search of an element. The cost of these operations is proportional to height of the treap. In what follows we will show that this quantity is $O(\log n)$ with high probability. Analyzing treaps boils down to the following problems on random permutations [38]. Given a permutation $p : [n] \to [n]$ of the n elements, an element is *checked* if it is larger than all elements appearing to its left in p. For instance, if

$$p = \mathbf{3}\ 1\ \mathbf{5}\ 4\ \mathbf{8}\ 6\ 2\ 7$$

the elements that are checked are in bold. It is convenient to generate the permutation by ranking n reals $r_i \in [0,1]$ chosen independently and uniformly at random for each element (the element i with the smallest r_i is the first element of the permutation, and so on. Ties occur with probability zero).

(b) Denoting with X_n the elements that are checked when p is random, prove that

$$\mathtt{E}[X_n] = 1 + \frac{1}{2} + \ldots + \frac{1}{n}.$$

(It is known that the quantity $H_n := \sum_{i=1}^{n} \frac{1}{i}$, the nth *harmonic number*, is $\Theta(\log n)$.

(c) Let Y_i be a binary random variable denoting whether element i is checked. Prove that

$$\Pr[Y_i = 1 \mid Y_n = y_k, \ldots, Y_{i+1} = y_{i+1}] = \frac{1}{k - i + 1}$$

for any choice of the ys.

(d) Is the following true?

$$\Pr[Y_i = 1 \mid Y_1 = y_k, \ldots, Y_{i+1} = y_{i-1}] = \frac{1}{i}$$

(e) Using the generalization of Problem 1.14 prove that X_n is $O(\log n)$ with high probability.

(f) Show that the number of nodes (x, p_x) such that $x < k$ that lie along the path from the root to (k, p_k) is given by X_k.

(g) Prove an analogous statement for the elements $x > k$ and conclude that the height of a treap is $O(\log n)$ with high probability.

$\bigtriangledown$

Problem 3.12 The following type of geometric random graphs arises in the study of power control for wireless networks. We are given n points distributed uniformly at random within the unit square. Each point connects to the k closest points. Let us denote the resulting (random) graph as G_k^n.

add refs?

- Show that there exists a constant α such that, if $k \geq \alpha \log n$, then G_k^n is connected with probability at least $1 - \frac{1}{n}$.
- Show that there exists a constant β such that, if $k \leq \beta \log n$, then G_k^n is not connected with positive probability.

$\bigtriangledown$

end new

Chapter 4

Chernoff-Hoeffding Bounds in Dependent Settings

[CH-bounds with Dependencies]

In this chapter, we consider the sum

$$X := \sum_{\alpha \in \mathcal{A}} X_\alpha, \tag{4.1}$$

where $\mathcal{A}$ is a index set and the variables $X_\alpha, \alpha, \in \mathcal{A}$ may not be independent. In some dependent situations, the Chernoff-Hoeffing bound can be salvaged to be applicable (as is, or with slight modifications) to X.

4.1 Negative Dependence

The results in this section are from D. Dubhashi and D. Ranjan, "Balls and Bins: A Study in Negative Dependence", *Random Structures and Algorithms*, 13 (1998), no. 2, 99–124.

We consider the sum (4.1) where $\mathcal{A} := [n]$. Random variables $X_1, \cdots, X_n$ are said to be *negatively dependent* if, intuitively, the conditioned on a subset $X_i, i \in I \subseteq [n]$ taking "high" values, a disjoint subset $X_j, j \in I \subseteq [n]$ with $I \cap J = \emptyset$ take "low" values. One way to formalize this intuitive notion is

Definition 4.1 (Negative Association) *The random variable $X_i, i \in [n]$ are* **negatively associated** *if for all disjoint subsets $I, J \subseteq [n]$ and all non-decreasing* *functions f and g,*

$$\mathtt{E}[f(X_i, i \in I)g(X_j, j \in J)] \leq \mathtt{E}[f(X_i, i \in I]\mathtt{E}[g(X_j, j \in J). \tag{4.2}$$

Exercise 4.2 *Show that if $X_1, \cdots X_n$ are negatively associated, then*

$$\mathtt{E}[X_i X_j] \leq \mathtt{E}[X_i]\mathtt{E}[X_j], \quad i \neq j.$$

More generally, show that if $f_i, i \in [n]$ are non-decreasing functions, then

$$\mathtt{E}[\prod_i f_i(X_i)] \leq \prod_i \mathtt{E}[f_i(X_i)].$$

In particular,

$$\mathtt{E}[e^{t(X_1+\cdots+X_n)}] \leq \prod_{i \in [n]} e^{tX_i}. \tag{4.3}$$

Theorem 4.3 (CH Bounds with Negative Dependence) *The Chernoff-Hoeffding bounds can be applied as is to $X := \sum_{i \in [n]} X_i$ if the random variables $X_i, \cdots, X_n$ are negatively associated.*

Proof. Use (4.3) at the relevant step in the proof of the CH bound. ■

Thus one needs techniques to establish the negative association condition. Although the defintion looks formidable, it is often easy to establish the condition *without any calculations* using only montonicity, symmetry and independence. The following two properties of negative association are very useful in these arguments.

Closure under Products If $X_1, \cdots, X_n$ and $Y_1. \cdots, Y_m$ are two independent families of random variables that are separetely negatively associated then, the family $X_1, \cdots, X_n, Y_1, \cdots, Y_m$ is also negatively associated.

Disjoint Monotone Aggregation If $X_i, i \in [n]$ are genatively associated, and $\mathcal{A}$ is a family of disjoint subsets of $[n]$, then the random variables

$$f_A(X_i, i \in A), A \in \mathcal{A},$$

is also negatively associated, where $f_A, A \in \mathcal{A}$ are arbitrary non-decreasing (or non-increasing) functions.

Exercise 4.4 *Show that these two properties follow directly from the definition of negative association.*

Example 4.5 [Balls and Bins] Consider the paradigm example of negative dependence: m balls are thrown independently into n bins. We do not assume the balls or bins are identical: ball k has probbaility $p_{i,k}$ of landing in bin i, for $i \in [n], k \in m$ (with $\sum_i p_{i,k} = 1$ for each $k \in [m]$). The *occupancy numbers* are $B_i := \sum_k B_{i,k}$. Intuitively it is clear that $B_1, \cdots, B_n$ are negatively dependent. To prove this, we first show that a simpler set of variables satisfies negative association, and then use the properties of disjoint monotone aggregation and closure under product.

Consider the indicator random variables:

$$B_{i,k} := \begin{cases} 1, & \text{ball } k \text{ falls in bin } i \\ 0, & \text{otherwise} \end{cases} \tag{4.4}$$

We have

Proposition 4.6 *For each k, the random variables $B_{i,k}, i \in [n]$ are negatively associated.*

Proof. Let I, J be disjoint subsest of $[n]$ and let f, g be non-decreasing functions. Translating by a constant, we may assume f and g are non-negative and $f(0, \cdots, 0) = 0 = g(0, \cdots, 0)$. Then,

$$\mathtt{E}[f(X_i, i \in I)g(X_j, j \in J)] = 0 \leq \mathtt{E}[f(X_i, i \in I)]\mathtt{E}[g(X_j, j \in J)].$$

∎

Now by closure under products, the full set $B_{i,k}, i \in [n], k \in [m]$ is negatively associated. Finally, by disjoint monotone aggregation, the variables $B_i = \sum_k B_{i,k} i \in [n]$ are negatively associated. ▽

Example 4.7 [Distributed Edge Colouring of Graphs] The application in this example is from A. Panconesi and A. Srinivasan, "Randomized Distributed Edge Colouring via an Extension of the Chernoff-Hoeffding Bounds", *SIAM J. Computing*, 26:2, pp. 350–368, 1997.

Consider the following simple distributed algorithm for edge colouring a bipartite graph $G = (B, T, E)$. (The bipartition is made up of the "bottom" vertices B and the "top" vertices T). For simplicity, assume $|B| = n = |T|$ and that the graph is Δ regular. At any stage of the algorithm,

1. In the first step, each "bottom" vertex makes a proposal by a tentative assigment of a random permutation of $[\Delta]$ to its incident edges.

2. In the second step, a “top” vertex chooses from among all incident edges that have the same tentative colour, a winner by using an arbitrary rule (lexicographically first or random for instance). The winner gets successfully coloured and the losers are decoloured and go to the next stage of the algorithm.

The basic question to analyse is: how many edges are successfully coloured in one stage of the colouring algorithm. The situation at a “top” vertex is exactly a balls and bins experiment: the incident edges are the balls falling into the bins which are the colours. Call a edge that receives a final colour successfully a “winner”, and otherwise a “loser”. Recalling that there are Δ edges and Δ colours, the number of losing edges is bounded as follows:

$$\begin{aligned}
\#\text{ losers} &= \#\text{ balls} - \#\text{ winners}\\
&\leq \#\text{ bins} - \#\text{ non-empty bins}\\
&= \#\text{ empty bins}.
\end{aligned}$$

Thus we need to analyse $Z := \sum_i Z_i$ where Z_i is the indicator random variable for whether bin i is empty $i \in [\Delta]$. These random variables are manifestly not independent. However, they are negatively associated because

$$Z_i = [B_i \leq 0], i \in [n].$$

are non-increasing functions of disjoint sets of the occupancy variables $B_1, \cdots B_\Delta$ whcih are negatively associated by the previous example.

The analysis of the “bottom” vertices is significantly more complicated and will require the use of more sophisticated techniques. $\bigtriangledown$

Example 4.8 [Glauber Dynamics and Graph Colouring] The application in this example is from Thomas P. Hayes [26]

Glauber dynamics is a stochastic process generating a sequence $f_0, f_1, \cdots, f_t, \cdots$ of random $[k]$-colourings of the vertices of a graph $G := (V, E)$. The colouring f_0 is arbitrary. Given f_{t-1}, the colouring f_t is determined as follows: select a vertex $v = \sigma(t)$ uniformly at random and a colour $c \in [k] \setminus f_{t-1}(\Gamma(v))$ unifromly at random. The colouring f_t is identcal to f_{t-1} except that $f_t(v) = c$.

In the analysis of the convergence of the process to stationarity, one needs concentration of the following random variable X. Fix a time t_0, and a vertex $v \in V$. Then, $X := \sum_{w \in \Gamma(v)} X_w$ where the indicator random variable X_w is 1 if w was selected by the colouring schedule σ in the time window $[t_0 - Cn, t_0 + Cn]$ for some constant $C > 0$. The random variables $X_w, w \in \Gamma(v)$ are not independent.

However, they are negatively associated. To see this, consider the indicator random variables $[\sigma(t) = v], v \in V, t \geq 1$. These are exactly like the Balls and Bins indicator variables: the "balls" are the time instants and the "bins" are the vertices. Hence $([\sigma(t) = v], v \in V, t \geq 1)$ are negatively associated. Now note that $X_w := \sum_{t_0 - Cn \leq t \leq t_0 + Cn} [\sigma(t) = w]$ are non-decreasing functions of disjoint index sets, and hence by the disjoint monotone aggregation property, the variables $X_w, w \in \Gamma(v)$ are also negatively associated. $\bigtriangledown$

Example 4.9 [Geometric Load Balancing] This application in this example is from [31].

Let n points be thrown uniformly at random on the unit circle. This splits the unit circle into n arcs which we can number $1 \cdots n$ in counterclockwise order starting from an arbitrary point.. Let $Z_i = 1$ if the i arc has length at least c/n and 0 otherwise. The variables $Z_i, i \in [n]$ are manifestly not independent. However they are negatively associated. To see this, let $L_i, i \in [n]$ denote the lengths of the arcs. Intuitively it is clear that $(L_i, i \in [n])$ are negatively dependent and indeed by Problem 4.26, $(L_i, i \in [n])$ are negatively associated. Then $Z_i = [L_i \geq c/n], i \in [n]$ are non-decreasing functions of disjoint sets of negatively associated variables, and hence, by the disjoint monotone aggregation property, are themselves negatively associated. $\bigtriangledown$

4.2 Local Dependence

The following results are from S. Janson [28].

Consider the sum (4.1) where there may be only *local dependence* in the following well known sense. Call a graph Γ on vertex set $\mathcal{A}$ a *dependency graph* for $(X_\alpha, \alpha \in \mathcal{A})$ if when there is no edge between $\alpha \in \mathcal{A}$ and $\mathcal{A}^{'} \subseteq \mathcal{A}$, then X_α is independent of $(X_{\alpha'}, \alpha^{'} \in \mathcal{A}^{'})$. Let $\chi^*(\Gamma)$ denote the *fractional chromatic number* of Γ.

The chromatic and fractional chromatic number $\chi^*(G)$ of a graph $G = (V, E)$ are defined as follows. Let B be the $|V| \times m$ matrix whose columns are characteristic vectors of independent sets in G. The chromatic number of G is the minimum number of colours needed in a proper colouring of G. Equivalently,

$$\chi(G) := \min \left(1^T x \mid Bx \geq 1, x \in \{0,1\}^m \right).$$

The fractional chromatic number $\chi^*(G)$ is the relaxation of this to non-negative vectors x:

$$\chi^*(G) := \min \left(1^T x \mid Bx \geq 1, x \geq 0 \right).$$

Clearly $\chi^*(G) \leq \chi(G)$.

Exercise 4.10 *Compute $\chi(C_n)$ and $\chi^*(C_n)$ where C_n is the circle with n points.*

Theorem 4.11 *Suppose X is as in (4.1) with $a_\alpha \leq X_\alpha \leq b_\alpha$ for real numbers $a_\alpha \leq b_\alpha, \alpha \in \mathcal{A}$. Then, for $t > 0$,*

$$P[X \geq \mathtt{E}[X] + t] \quad P[X \leq \mathtt{E}[X] - t] \quad \leq \quad \exp\left(\frac{-2t^2}{\chi^*(\Gamma)\sum_{\alpha\in\mathcal{A}}(b_\alpha - a_\alpha)^2}\right).$$

Exercise 4.12 *Check that $\chi^*(\Gamma) = 1$ iff the variables X_α are independent, so Theorem 4.11 is a proper generalization of the Hoeffding inequality.*

Example 4.13 [U-Statistics] Let $\xi_1, \cdots, \xi_n$ be independent random variables, and let

$$X := \sum_{1\leq i_1<\cdots<i_d} f_{i_1,\cdots,i_d}(\xi_{i_1}, \cdots, \xi_{i_d}).$$

This is a special case of (4.1) with $\mathcal{A} := [n]^d_<$ and includes the so-called *U-Statistics*. The dependency graph Γ has vertex set $[n]^d_<$ and $(\alpha, \beta) \in E(\Gamma)$ iff $\alpha \cap \beta$ / $\emptyset$, when the tuples α, β are regarded as sets. One can check (see Problem 4.32 that

$$\chi^*(\Gamma) \leq \frac{\binom{n}{d}}{\lfloor n/d \rfloor}.$$

Hence, if $a \leq f_{i_1,\cdots,i_d}(\xi_{i_1}, \cdots, \xi_{i_d}) \leq b$ for every $i_1, \cdots, i_d$ for some reals $a \leq b$, we have the estimate of Hoeffding:

$$P\left[X \geq \mathtt{E}[X] + t\binom{n}{d}\right] \leq \exp\left(\frac{-2\lfloor n/d \rfloor t^2}{(b-a)^2}\right)$$

Since $d\lfloor n/d \rfloor \geq n - d + 1$, we have $\chi^*(\Gamma) \leq \binom{n}{d-1}$ and we have a bound that looks somewhat simpler:

$$P\left[X \geq \mathtt{E}[X] + tn^{d-1/2}\right] \leq \exp\left(\frac{-2d!(d-1)!t^2}{(b-a)^2}\right)$$

$\bigtriangledown$

Example 4.14 [Subgraph Counts] Let $G(n, p)$ be the random graph on vertex set $[n]$ with each possible edge (i, j) present independently with probability p. Let X denote the number of triangles in $G(n, p)$. This can be written in the form (4.1) with $\mathcal{A} := \binom{[n]}{3}$ and X_α is the indicator that the edges between the three vertices in α are all present. Note that X_α and X_β are independent even if $\alpha \cap \beta = 1$ (but not 2). The dependency graph Γ has vertex set $\binom{[n]}{3}$ and $(\alpha, \beta) \in E(\Gamma)$ iff $\alpha \cap \beta = 2$. Note that $\Delta(\Gamma) = 3(n-3)$ and hence

$$\chi^*(\Gamma) \leq \chi(\Gamma) \leq \Delta(\Gamma) + 1 \leq 3n.$$

We compute $\mathtt{E}[X] = \binom{n}{3}p^3$ and hence

$$P[X \geq (1+\epsilon)\mathtt{E}[X]] \leq \exp\left(\frac{-2\epsilon^2\binom{n}{3}^2 p^6}{3n\binom{n}{2}}\right) = \exp\left(-\Theta(\epsilon^2 n^3 p^6)\right).$$

This estimate can be improved taking into account the variance of the summands. $\bigtriangledown$

4.3 Janson's Inequality

Let $R = R_{p_1,\cdots,p_n}$ be a random subset of $[n]$ formed by including each $i \in [n]$ in R with probability p_i, independently. Let $\mathcal{S}$ be a family of subset of $[n]$, and for each $A \in \mathcal{S}$, introduce the indicators

$$X_A := [A \subseteq R] = \bigwedge_{i \in A} [i \in R].$$

Let $X := \sum_{A \in \int} X_A$. Clearly the summands are not independent. In the terminilogy of the last section, a natural dependency graph G for $(X_A, A \in \mathcal{S})$ has vertex set $\mathcal{S}$ and an edge $(A, B) \in G$ iff $A \cap B \neq \emptyset$: in this case, we write $A \sim B$.

Theorem 4.15 *Let $X := \sum_A X_A$ as above, and let $\mu := \mathtt{E}[X] = \sum_A \mathtt{Pr}[X_A = 1]$. Define*

$$\Delta := \sum_{A \sim B} \mathtt{E}[X_A X_B] = \sum_{A \sim B} \mathtt{Pr}[X_A = 1 = X_B], \tag{4.5}$$

where the sum is over **ordered** *pairs. Then, for any $0 \leq t \leq \mathtt{E}[X]$,*

$$\mathtt{Pr}[X \leq \mathtt{E}[X] - t] \leq \exp\left(-\frac{t^2}{2\mu + \Delta}\right).$$

Exercise 4.16 *Check that when the sets $A \in \mathcal{S}$ are disjoint, then this reduces to the CH-bound.*

In particular, taking $t := \mathtt{E}[X]$ gives a avrey useful estimate on the probability that no set in $\mathcal{S}$ occurs which is important enough to deserve a separate statement of its own:

Theorem 4.17 (Janson's Inequality)

$$\texttt{Pr}[X = 0] \leq e^{-\frac{\mu^2}{\mu+\Delta}}.$$

As verified in the exercise above, when the sets are disjoint, we are in the independent case, More importantly, when the dependence is "small" i.e. $\Delta = o(\mu)$, then, we get neraly the same bound as well.

Example 4.18 [Subgraph Counts] Consider again the random graph $G(n,p)$ with vertex set $[n]$ and each (undirected) edge (i,j) present with probability p independently and focus again on the number of triangles in the random graph. An interesting regime of the parameter p is $p := c/n$. The base set Γ here is $\binom{[n]}{2}$, the set of all possible edges and the random set of edges in G picked as above is object of study. Let S be a set of three edges forming a traingle, and let X_S be the indicator that this triangle is present in $G(n,p)$. Then $\texttt{Pr}[X_S = 1] = p^3$. The property that G is triangle-free is expressed as $X := \sum_S X_S = 0$ where the sum is over all such $\binom{n}{3}$ subsets of edges S. If the X_S were independent then, we would have

$$\texttt{Pr}[X = 0] = \texttt{Pr}\left[\bigwedge_S X_S = 0\right] = \prod_S \texttt{Pr}[X_S = 0] = (1-p^3)^{\binom{n}{3}} \sim e^{-\binom{n}{3}p^3} \to e^{-c^3/6}.$$

Of course the X_S are not independent. But if $\mathcal{A}$ and $\mathcal{B}$ are collectiosn of subsets such that each $S \in \mathcal{A}$ is disjoint from each $T \in \mathcal{B}$, then $(X_S, S \in \mathcal{A})$ is mutually independent of $(X_T, T \in \mathcal{B})$.

We can thus apply Janson's inequality, Theorem 4.17. here $\mu = \texttt{E}[X] = \binom{n}{3}p^3 \sim c^3/6$. To estimate Δ, we note that there are $nchoose3(n-3) = O(n^4)$ ordered pairs (S,T) with $S \cap T \neq \emptyset$, and for each such pair, $\texttt{Pr}[X_S = 1 = X_T] = p^5$. Thus, $\Delta = O(n^4)p^5 = n^{-1+o(1)} = o(1)$. Thus, we get the bound

$$\texttt{Pr}[X = 0] \leq \exp\left(-\frac{c^6}{36c^3 + o(1)}\right) \sim e^{c^3/36},$$

which is (asymptotically) almost the same (upto constants) as the estimate above assuming the variables were independent. In problem 4.30, you are asked to generalize this from traingles to arbitrary fixed graphs. $\bigtriangledown$

Example 4.19 [Randomized Rounding] The following example is taken from an analysis of approximation algorithms for the so-called *group and covering Steiner* problems [37, 19]. We are given a full binary tree T rooted at a special vertex r. In the group Steiner problem, we are also given groups $A_1, \cdots, A_n$ of subsets of the leaves of T. The objective is to select a subtree of minimum size rooted at r whoses leaves intersect each of the n groups.

The first step in the problem is to formulate a linear program which provides a lower bound on the size of any such tree. Solving this linear program gives a set of values $x_e \in [0,1], e \in T$. These values have the property that $\sum_{e \in E} x_e \geq 1$ for any set of edges that forma cut between r and a group g_i. Thus these values x_e can be used as a guide to constructing the required subtree.

This is done via the following variant of the *randomized rounding* methodology: for each edge $e \in T$, include e independently with probability x_e/x_f where f is the unique parent edge connecting e to the next vertex up the tree. If e is incident on the root, we include it with probability x_e (alternatively imagine a fictitious parent edge e_{-1} with $x_{e_{-1}} = 1$). Then pick the unique connected component rooted at r.

The rounding procedure has the property that any edge $e \in T$ is included with probability x_e. To see this, note that an edge is included iff all the edges $e = e_1, e_2, \cdots, e_p$ on the path up to the root from e are included, and this happens with probability

$$\frac{x_{e_1}}{x_{e_2}} \frac{x_{e_2}}{x_{e_3}} \cdots \frac{x_{e_{p-1}}}{x_{e_p}} \frac{x_{e_p}}{1} = x_e.$$

Let us focus attention on a particular group A and estimate the probability that this group is not "hit". We can identify the gropu A with the corresponding pendant edges. Let $X_e, e \in A$ be the indicator for whether the element $e \in A$ is selecetd, and let $X := \sum_{e \in A} X_e$. Then

$$\mathtt{E}[X] = \sum_{e \in A} \mathtt{E}[X_e] = \sum_{e \in A} x_e \geq 1,$$

where the last inequality is because of the cut-property of the x_e values.

Note however that the $X_e, e \in A$ are *not* independent : the dependencies arise because of shared edges on the path up the tree. Let us estimate Δ in this situation. To this end, first we note that the event $X_e = 1 = X_f$ for distinct $e, f \in A$ occurs iff (a) all edges up to and inclusing the common ancestor g of e aand f are picked, and (b) the remaining edges from g to e and f are all picked. Thus, $\mathtt{Pr}[X_e = 1 = X_f] = x_e x_f / x_g$.

Exercise 4.20 *Check this!*

Thus,

$$\Delta = \sum_e \sum_f x_e x_f / x_g.$$

To continue with the estimation, we make some simplifying assumptions (which are justified in the paper [19]: we assume that the group A is contained in a single subtree of height $d := \lceil |A| \rceil$, that $\sum_{e \in A} x_e = 1$ finally, that for any vertex v in the tree whose parent edge is e, we have

$$\sum_{f \in T'} x_f \leq x_e, \tag{4.6}$$

where T' is either the left or the right subtree rooted at v.

Now, to return to Δ, consider an edge e is the first summation. Number the path up from e to the root $r = v_0, v_1, \cdots v_{i-1}, v_i$ where $e = v_{i-1}v_i$. Let $T_j, 0 \leq j \leq i$ denote the subtree rooted at v_j which does not include e_i. Then,

$$\begin{aligned} \Delta &= \sum_e \sum_f x_e x_f / x_g \\ &= \sum_e \sum_{0 \leq j \leq i} \sum_{f \in T_j} f x_e x_f / x_g \\ &= \sum_e x_e \sum_{0 \leq j \leq i} \frac{\left(\sum_{f \in T_j} x_f \right)}{x_{e_{j-1}}} \\ &\leq \sum_e x_e \sum_{0 \leq j \leq i} 1, \quad \text{by (4.6)} \\ &= \sum_e (i+1) x_e \\ &\leq (d+2) \sum_e x_e \\ &= (d+2). \end{aligned}$$

Thus applying Janson's inequality, we get that the probability that the group A fails to be "hit" is at most $e^{-1/(3+\log |A|)} \approx 1 - \frac{1}{3 \log |A|}$. ▽

4.4 Limited Independence

One key objective in modern complexity theory has been to seek ways to reduce the amount of randomness used by probabilistic algorithms. The ultimate objective of course would be to remove the randomenss altogether leading to a deterministic algorithm via a complete *derandomization* of a randomized algorithm. In this quest, a reduction in randomization leads to some progress in the form of a partial derandomization.

One approach to reducing randomness comes from the observation that some algorithms do not need full independence of their source of random bits. We say that a set of random variables $X_1, \cdots X_n$ is *k-wise independent* if for every $I \subseteq [n]$ with $|I| \leq k$,

$$\mathtt{Pr}\left[\prod_{i \in I} X_i = x_i\right] = \prod_{i \in I} \mathtt{Pr}[X_i = x_i].$$

Fully independent varaiables correspond to n-wise independence.

In this section, we outline the approach of [64] to obtaining CH-like bounds for the case of random variables with limited dependence i.e. when they are only k-wise independent for some $k < n$.

Consider the *elementary symmetric functions*:

$$S_k(x_1, \cdots, x_n) := \sum_{I \subseteq [n], |I| = k} \prod_{i \in I} x_i.$$

Observe that for 0/1 variables $x_1, \cdots, x_n$, and an integer $m \geq 0$,

$$\sum_i x_i = m \quad \leftrightarrow \quad S_k(x_1, \cdots, x_n) = \binom{m}{k}.$$

Also, if $X_1, \cdots, X_n$ are k-wise independent, then:

$$\begin{aligned} \mathtt{E}\left[S_k(X_1, \cdots, X_n)\right] &= \mathtt{E}\left[\sum_{|I|=k} \prod_{i \in I} X_i\right] \\ &= \sum_{|I|=k} \mathtt{E}\left[\prod_{i \in I} X_i\right] \\ &= \sum_{|I|=k} \prod_{i \in I} \mathtt{E}[X_i] \end{aligned}$$

In the last line, we use the k-wise independence of the variables.

Hence, if $X := X_1 + \cdots + X_n$ for binary random variables $X_1, \cdots, X_n$ which are k-wise independent and $\mathtt{E}[X_i] = \mathtt{Pr}[X_i = 1] = p$ for each $i \in [n]$, then

$$\begin{aligned} \mathtt{Pr}[X > t] &= \mathtt{Pr}\left[S_k(X_1, \cdots, X_n) > \binom{t}{k}\right] \\ &\leq \mathtt{E}\left[S_k(X_1, \cdots, X_n)\right] / \binom{t}{k} \\ &= \frac{\binom{n}{k} p^k}{\binom{t}{k}} \end{aligned}$$

In Problem 4.29, you are asked to check that this bound holds also when the variable are not identically distributed and when they take values in the interval $[0, 1]$. This yields the following version of the CH-bound for variables with limited independence:

Theorem 4.21 *Let $X_1, \cdots, X_n$ be random variables with $0 \leq X_i \leq 1$ and $\mathtt{E}[X_i] = p_i$ for each $i \in [n]$. Let $X := \sum_i X_i$ and set $\mu := \mathtt{E}[X]$ and $p := \mu/n$. Then, for any $\delta > 0$, if $X_1, \cdots, X_n$ are k-wise independent for $k \geq k_* := \lceil \mu\delta/(1-p) \rceil$,*

$$\mathtt{Pr}\left[X \geq \mu(1+\delta)\right] \leq \binom{n}{k_*} p^{k_*} \Big/ \binom{\mu(1+\delta)}{k_*}$$

Exercise 4.22 *Check that this bound is better than the CH-bound $e^{-\frac{\delta^2}{3}\mu}$ derived in the previous chapter.*

Another approach due to Bellare and Rompel [4] goes via the k–th *moment inequality*:

$$\begin{aligned} \mathtt{Pr}[|X-\mu| > t] &= \mathtt{Pr}[(X-\mu)^k > t^k], \text{since } k \text{ is even} \\ &< \frac{\mathtt{E}[(X-\mu)^k]}{t^k}, \text{by Markov's inequality.} \end{aligned} \tag{4.7}$$

To estimate $\mathtt{E}[(X-\mu)^k]$, we observe that by expanding and using linearity of expectation, we only need to compute $\mathtt{E}[\prod_{i \in S}(X_i - \mu_i)]$ for multi–sets S of size k. By the k–wise independence property, this is the same as $\mathtt{E}[\prod_{i \in S}(\hat{X}_i - \mu_i)]$, where $\hat{X}_i, i \in [n]$ are fully independent random variables with the same marginals as $X_i, i \in [n]$. Turning the manipulation on its head, we now use Chernoff–Hoeffding bounds on $\hat{X} := \sum_i \hat{X}_i$:

$$\begin{aligned} \mathtt{E}[(\hat{X}-\mu)^k] &= \int_0^\infty \mathtt{Pr}[(\hat{X}-\mu)^k > t] dt \\ &= \int_0^\infty \mathtt{Pr}[|\hat{X}-\mu| > t^{1/k}] dt \\ &< \int_0^\infty e^{-2t^{2/k}/n} dt, \quad \text{using CH bounds} \\ &= (n/2)^{k/2} \frac{k}{2} \int_0^\infty e^{-y} y^{k/2-1} dy \\ &= (n/2)^{k/2} \frac{k}{2} \Gamma(k/2-1) \\ &= (n/2)^{k/2} (k/2)! \end{aligned}$$

Now using Stirling's approximation for $n!$ gives the estimate:

$$\mathtt{E}[(\hat{X}-\mu)^k] \leq 2e^{1/6k}\sqrt{\pi t}\left(\frac{nk}{e}\right)^{k/2},$$

which in turn, plugged into (4.7) gives the following version of a tail estimate valid under limited i.e. k–wise dependence:

$$\mathtt{Pr}[|X-\mu|>t] \leq C_k\left(\frac{nk}{t^2}\right)^{k/2},$$

where $C_k := 2\sqrt{\pi k}e^{1/6k} \leq 1.0004$.

4.5 Markov Dependence

4.5.1 Definitions

A Markov chain M is defined by a state space U and a stochastic transition matrix P (i.e. $\sum_x P(x,y)=1$). Starting with an initial distribution q on U, it determines a sequence of random variables $X_i, i \geq 1$ as follows: for $n \geq 1$ and any $x_1, \cdots, x_n, x_{n+1} \in U$,

$$\mathtt{Pr}[X_1 = x_1] = q(x_1),$$

. and,

$$\mathtt{Pr}[X_{n+1} = x_{n+1} \mid X_1 = x_1, \cdots, X_n = x_n] = \mathtt{Pr}[X_{n+1} = x_{n+1} \mid X_n = x_n] = P(x_{n+1}, x_n).$$

A distribution π on S is called *stationary* for M if $\pi P = P$. Under a technical condition called *aperiodicity*, a Markov chain whose state space is connected has a unique stationary distribution. The aperiodicity condition can usually be made to hold in all the applications we consider here. For more details on these conditions and a careful but friendly introduction to Markov chains, see [25].

The general theory of Markov chains [25] shows that under these conditions, the Markov chain, started at any point in the state space, eventually converges to the stationary distribution in the limit. The rate of convergence is determined by the so-called *eigenvalue* gap of the transition matrix P of the Markov chain. Since the matrix is stochastic, the largest eigenvalue is $\lambda_1 = 1$ and the general theory of non-negative matrices implies that the second eigenvalue λ_2 is strictly less than 1. The eigenvalue gap is $\epsilon := \lambda_1 - \lambda_2 = 1 - \lambda_2$.

4.5.2 Statement of the Bound

Let $X_1, X_2, \cdots, X_n$ be a sequence generated by a Markov chain with eigenvalue gap ϵ starting from an initial distribution q. Let f be a non-negative function on the state space of M, and let $F_n := \sum_{i\in[n]} f(X_n)$. By the convergence to stationarity of the Markov chain, we know that $\lim_{n\to\infty} F_n/n = \mathtt{E}[f]$. The following Theorem due independently to Gillman [20] and Kahale [32] gives a quantitative bound on this convergence.

Theorem 4.23 *Let $X_1, X_2, \cdots, X_n$ be a sequence generated by a Markov chain with eigenvalue gap ϵ starting from an initial distribution q. For a For a non-negative function f, on the state space of M let $F_n := \sum_{i\in[n]} f(X_n)$. Then,*

$$\mathtt{Pr}[|F_n - n\mathtt{E}[f]| > t] \leq C_{\gamma,\epsilon,n,q} \exp\left(-\epsilon\frac{t^2}{cn}\right).$$

where c is an absolute constant and $C_{\gamma,\epsilon,n,q}$ is a rational function. In particular, taking $f := \chi_S$, the characteristic function of a subset S of the state space, and letting $T_n := \sum_{i\in[n]} \chi_S(X_i)$ denote the number of times the chain is in state S,

$$\mathtt{Pr}[|T_n - n\pi(S)| > t] \leq C_{\gamma,\epsilon,n,q} \exp\left(-\epsilon\frac{t^2}{cn}\right).$$

Note that this is very similar to the usual Chernoff bound, except for the rational term and, more importantly, the appearence of the eigenvalue gap in the exponent.

4.5.3 Application: Probability Amplification

Let $f : \{0,1\}^n \to \{0,1\}$ be a function that is computed by a randomized algorithm A that takes as input the argument $x \in \{0,1\}$ at which f has to be evaluated and also a sequence r of n random bits. Suppose the algorithm A is guaranteed to compute f correctly with a constant probability bounded away from 1/2, say,

$$\mathtt{Pr}_r[A(x,r) = f(x)] \geq 3/4.$$

We would like to *amplify* the success probability i.e. provide an algorithm $\hat{A}$ that computes f correctly with probability arbitrarily close to 1.

The standard way to do this is by repetition: make k runs k of algorithm A and take the majority outcome. Each run of the algorithm is independent of the previous one and uses n fresh independent random bits. What is the success probability of the resulting algorithm? Recall the standard application of the Chernoff bound in the previous chapter: let $X_1, \cdots, X_n$ be indicator random variables with $X_i = 1$ iff algorithm A computes f correctly on the ith invocation, and set $X := \sum_i X_i$. The Chernoff bound yields

$$\texttt{Pr}[X < 1/2k] \leq e^{-k/8}.$$

Thus to achive an error probability of at most δ, we can take $k = O(\log \frac{1}{\delta})$.

We shall now describe an algorithm that achieves similar amplification of probability, but with the advantage that the algorithm will be significantly more effcient in its use of randomness as a resource. The algorithm above uses a total of nk random bits. The algorithm we describe next will use only $O(n+k)$ random bits to achieve very similar error probability.

To do this we start with an *expander* graph G on the vertex set $\{0,1\}$, the underlying probability space of the original algorithm A. Expander graphs are very useful in many different areas of algorithms and complexity. This example is tyoical and can be viewed as an introduction to their uses. Here, we will only state the properties we need. The expander graph G is regular of constant degree d. The expansion property is that any subset A of the vertices has a neighbourhood of size at least $\gamma|A|$ for some positive constant γ.

There is an equivalent algebraic characterization which is more directly of use to us here. Consider the simple random walk on the graph G: start at any vertex and choose the next vertex uniformly at random from all the neighbours. This defines a Markov chain $M(G)$ with state space the vertices of G whose unique stationary distribution is the uniform distribution. The expansion property of G translates equivalently into the property that the the Markov chain $M(G)$ has an eigenvalue gap $\epsilon > 0$ i.e. the first eigenvalue is 1 and the second is bounded from above by $1 - \epsilon$.

We are now in a position to state our algorithm and analyse it using the CH bound for Markov chains. The algorithm $\tilde{A}$ is as follows:

1. Pick a point $r_1 :=\in \{0,1\}$ at random. Then starting at r_1, execute a random walk on G: $r_1, r_2, \cdots, r_k$.

2. Run the algorithm k times, using these bits as the random source:

   $$A(x, r_1), A(x, r_2) \cdots, A(x, r_k),$$

   and take the majority outcome.

To analyse the success probability of the algorithm $\tilde{A}$, we introduce as before, the indicators $X'_i, i \in [k]$ with $X'_i = 1$ if the algorithm is correct on trial i and 0 otherwise. Now, since r_1 is picked according to the stationary distribution, the merginal distribution of each $r_i, i \in k$ separately is also the stationary distribution which is uniform. Hence $\mathtt{Pr}[X'_i = 1] \geq 3/4$ for each $i \in [k]$ and so, if $X' := \sum_{i \in [k]} X'_i$, then $\mathtt{E}[X'] \geq 3/4k$. So far the analysis is identical to what we saw before.

The hitch is in the fact that whereas the indicators $X_1, \cdots, X_k$ were independent before due to the fresh choice of random bits every time the algorithm A is rerun, this time, the indicators $X'_1, \cdots, X'_k$ are *not* independent because the sequence $r_1, \cdots, r_k$ is chosen by a random walk - thus each r_i depends heavily on its predecessor r_{i-1}. This is the place where Theorem 4.23 kicks in. Let $S := \{r \in \{0,1\}^n \mid A(x,r) = f(x)\}$. Note that since the stationary distribution π is uniform, $\pi(S) \geq 3/4$. Applying Theoremth:markov-ch, we get:

$$\mathtt{Pr}[X' < 1/2k] \leq e^{-c\epsilon k},$$

for some constant $c > 0$. This is essentially the same error probability as we had for algorithm $\hat{A}$ with the independent repetitions except for constant factors in the exponent. However, in this case, the number of random bits used by algorithm $\tilde{A}$ is $O(n+k)$ compared to nk bits needed by algorithm $\hat{A}$.

Exercise 4.24 *Work out the number of bits used by algorithm $\tilde{A}$. Note the fact that G is a constant degree graph is needed here.*

Exercise 4.25 *Work out the constant in the exponent of the error bound in terms of the constants in Theorem 4.23.*

4.6 Bibliographic Notes

Negative Dependence is tretated at greater length in [16]. A plethora of versions of CH bounds for limited independence are given in [64] with applications to reducing randomness requirements of algorithms. Stronger versions of Theorem 4.11 are given in [28] with more applications. Gillman [20] gives more applications of Theorem 4.23. Kahale [32] gives almost tight versions of the bound for Markov chians and compares to the bounds of Gillman.

4.7 Problems

Problem 4.26 Let $X_i, i \in [n]$ be random variables such that for any subset $I \subseteq [n]$, and any $t > 0$, the distribution of $X_i, i \in I$ conditioned on $\sum_{i \in I} X_i = t$ is

- conditionally independent of any other variables.
- stochastically increasing in t.

Further suppose the distribution of $X_i, i \in [n]$ is concentrated on the event $\sum_{i \in [n]} X_i = c$ for some constant c. Then $X_i, i \in [n]$ are negatively associated. Deduce that the arc variables $L_i, i \in [n]$ in Example 4.9 are negatively asociated.
▽

Problem 4.27 [Negative Regression] A set of random variables $X_1, \cdots, X_n$ satisfy the **negative regression** condition $(-R)$, if, for any two disjoint index ests $I, J \subseteq [n]$, and any non-decresing function f,

$$\mathtt{E}[f(X_i, i \in I) \mid X_j = a_j, j \in J] \tag{4.8}$$

is non-incresing in each $a_j, j \in J$.

1. Show that if $(-R)$ holds, then $\mathtt{E}\left[\prod_i f_i(X_i)\right] \leq \prod_i \mathtt{E}[f_i(X_i)]$ for any non-decresing functions $f_i, i \in [n]$.
2. Deduce that the CH bound applies to variables satisfying $(-R)$.

▽

Problem 4.28 [Permutations] Recall the following problem on permutations encountered int he analysis of Treaps: a position i in a permutation σ of $[n]$ is "checked" if $\sigma(j) < \sigma(i)$ for all $j < i$. Let σ be a permutation chosen uniformly at random, and let $X_i, i \in [n]$ be indicator variables for whether a position is checked. Shwo that these variables satisfy $(-R)$. ▽

Problem 4.29 Prove Theorem 4.21. Also derive a bound on the lower tail. ▽

Problem 4.30 [Subgraph Counts] Consider the random graph $G(n,p)$ and let us consider the number of occurences $X(H)$ of the number of occurences of H in G. Define

$$\phi_H = \phi_H(n,p) := \min\{\mathtt{E}[X_{H'}] \mid H' \subseteq H, e_{H'} > 0\}.$$

Note that $\phi_H \approx \min_{H' \subseteq H, e_{H'} > 0} n^{v'_H} p^{e'_H}$. where v_H is the number of vertices and e_H the number of edges of a graph H. Show that for any fixed graph H (with at least one edge), $\mathtt{Pr}[X_H = 0] \leq \exp\left(-\Theta(\phi_H)\right)$. $\bigtriangledown$

Problem 4.31 [Sampling with reduced randomness [64]] Recall the problem of estimating the fraction $f^* := |W|/|U|$ of elements of a special subset W of a large universal set U. The approach is to take a random sample S from U and estimate f^* by $\hat{f} := |W \cap S|/|S$. Investigate the possibility of reducing the randomness requirements of this algorithm using Theorem 4.21 or Theorem 4.23. $\bigtriangledown$

Problem 4.32 [Fractional Chromatic Number of Kneser Graphs] Consider the **Kneser graphs** $K(n,d)$ whose vertex set is $\binom{[n]}{d}$ and whose edge set is $\{(A,B) \mid A \cap B = \emptyset\}$. Compute bounds on $\Delta(K(n,d))$, $\chi(K(n,d))$ and $\chi^*(K(n,d))$. $\bigtriangledown$

Chapter 5

Martingales and Azuma's Inequality

[Martingales and Azuma's Inequality]

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.

5.1 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 $\boldsymbol{X}$ and a sequence of real values $a_1, a_2, \ldots$ by $\boldsymbol{a}$. When given such a sequence, we shall use the subscript n to denote the prefix of length n; thus $\boldsymbol{X}_n$ will denote $X_1, \ldots, X_n$ and $\boldsymbol{a}_n$ 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 $\boldsymbol{X}_n = \boldsymbol{a}_n$. 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

$$\mathtt{E}[Y \mid \mathcal{E}] := \sum_b b \cdot \mathtt{Pr}[Y = b \mid \mathcal{E}]. \tag{5.1}$$

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 (5.1) is $\boldsymbol{X} = \boldsymbol{a}$, we have a multivariate function

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

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

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 5.1 *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 5.1 generalises smoothly to the multivariate case. Once again we leave the verification as an exercise.

Proposition 5.2 (Fundamental Facts about Conditional Expectation) *Let $\boldsymbol{X}$, $\boldsymbol{Y}$ and Z be random variables defined on a probability space. For arbitrary* *functions f and g,*

$$\mathtt{E}[\mathtt{E}[f(\boldsymbol{X})g(\boldsymbol{X},\boldsymbol{Y}) \mid \boldsymbol{X}]] = \mathtt{E}[f(\boldsymbol{X})\mathtt{E}[g(\boldsymbol{X},\boldsymbol{Y}) \mid \boldsymbol{X}]]. \tag{5.2}$$

Also,

$$\mathtt{E}[X] = \mathtt{E}[\mathtt{E}[X \mid \boldsymbol{Y}]], \tag{5.3}$$

and

$$\mathtt{E}[X \mid \boldsymbol{Z}] = \mathtt{E}[\mathtt{E}[X \mid \boldsymbol{Y},\boldsymbol{Z}] \mid \boldsymbol{Z}]. \tag{5.4}$$

These facts will be heavily used in this chapter.

5.2 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 5.3 *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

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

The next examples and exercises should help clarify the definition.

Example 5.4 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. ▽

Example 5.5 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. ▽

Exercise 5.6 *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 5.4?*

The following definition is central.

Definition 5.7 (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 5.8 (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 $\boldsymbol{X}$ replaced by $-\boldsymbol{X}$. 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:

$$\begin{aligned} \Pr(X_n > t) &= \Pr(e^{\lambda X_n} > e^{\lambda t}) \\ &\leq \frac{\mathtt{E}[e^{\lambda X_n}]}{e^{\lambda t}} \end{aligned} \tag{5.5}$$

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*:

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

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

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

Note that the martingale condition can be rephrased as follows:

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

Applying the basic equality (5.3), we get

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

The last equality follows from (5.2).

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

$$\begin{aligned}
\mathtt{E}[e^{\lambda X_n}] &= \mathtt{E}[\exp(\lambda X_{n-1})\mathtt{E}[e^{\lambda Y_n}|\boldsymbol{X}_{n-1}]] \\
&\leq \mathtt{E}[\exp(\lambda X_{n-1})] \cdot U_n(\lambda) \qquad (5.6) \\
&\leq \prod_{i=1}^{n} U_i(\lambda) =: U(\lambda).
\end{aligned}$$

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

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

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

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

$$e^{\lambda Z} \leq mZ + q \qquad (5.7)$$

where

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

Taking expectations on both sides of Equation (5.7) we have

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

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. $\blacksquare$

We apply the lemma to the random variable

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

Z satisfies the hypotheses of the lemma since

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

by the martingale property, and

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

by the Bounded Difference Condition. Therefore

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

which, after substituting into Equation (5.6), 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 (5.5),

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

which ends the proof of Theorem 5.8.

Exercise 5.10 *Derive a similar inequality for the case when the martingale condition is replaced by the following:*

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

for some non–negative real number m.

Exercise 5.11 *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 $c := \sum_{i=1}^{n} c_i^2$.

5.3 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 5.12 *A sequence of random variables* $\boldsymbol{Y} := Y_0, Y_1, \ldots$ *is a* martingale with respect to another sequence $\boldsymbol{X} := X_0, X_1, \ldots$ *if for each* $i \geq 0$,

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

for some function g_i *and*

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

Example 5.13 Let us consider again example 5.4, 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 $n \geq 0$. $\bigtriangledown$

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

Definition 5.14 *The* **Doob sequence** *of a function f with respect to a sequence of random variables $X_1, \ldots, X_n$ is defined by*

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

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

Proposition 5.15 *The Doob sequence of a function defines a martingale. That is,*

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

The proof is an immediate consequence of (5.4).

Example 5.16 [Edge exposure martingale] An important special case of definition 5.14 occurs in the context of the *random graph $G_{n,p}$*. This is the graph with vertex set $[n]$ and each edge $\{i, j\}, i \neq j$ present with probability p independently of all other edges. Let $f : \binom{[n]}{2} \rightarrow \mathbb{R}$ 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. $\bigtriangledown$

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

Theorem 5.17 (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 $\boldsymbol{Y}$ satisfies the Bounded Difference Condition with parameters $c_i, i \geq 1$. Then,*

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

and

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

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

Proof. The proof is almost identical to that of Theorem 5.8. 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 $\mathtt{E}[e^{\lambda Y_n}]$.

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

The last line, the only place where the proof differs from that of Theorem 5.8, follows form (5.2) because $Y_{n-1} = g_{n-1}(\boldsymbol{X}_{n-1})$. The proof now proceeds identical to that of Theorem 5.8 provided Lemma 5.9 is invoked for the variables $Z := (D_n | \boldsymbol{X}_{n-1})$. The verification that Z satisfies the hypotheses of Lemma 5.9 is straightforward and is left as an exercise. ∎

5.4 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 5.18 *[The Method of Averaged Bounded 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

$$|\mathtt{E}[f|\boldsymbol{X}_i] - \mathtt{E}[f|\boldsymbol{X}_{i-1}]| \leq c_i. \tag{5.8}$$

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 5.17 for the special case of the Doob sequence $Y_i := \mathtt{E}[f|\boldsymbol{X}_i], 0 \leq i \leq n$. *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 5.19 (Averaged Lipschitz Condition) *A function f satisfies the* Averaged Lipschitz Condition *(henceforth* ALC*) 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',*

$$|\mathtt{E}[f|\boldsymbol{X}_{i-1}, X_i = a_i] - \mathtt{E}[f|\boldsymbol{X}_{i-1}, X_i = a_i']| \leq c_i \tag{5.9}$$

for $1 \leq i \leq n$.

In words, the condition ALC in (5.9) 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 5.20 (The Method of Averaged Bounded Differences: Alternate Take) *Let f satisfy the* ALC *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 (5.9) holds then so does (8.6). To see this, expand using total conditional probability:

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

and write

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

Hence,

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

■

Exercise 5.21 *Show that if for each $i \in [n]$,*

$$|\mathtt{E}[f|\boldsymbol{X}_i] - \mathtt{E}[f|\boldsymbol{X}_{i-1}]| \leq c_i,$$

. *then for any a_i, a'_i,*

$$|\mathtt{E}[f|\boldsymbol{X}_{i-1}, X_i = a_i] - \mathtt{E}[f|\boldsymbol{X}_{i-1}, X_i = a'_i]| \leq 2c_i.$$

That is, the two alternate takes of the Method of Averaged Bounded Differences are virtually the same but for a factor of 2.

A further significant simplification obtains from the following definition.

Definition 5.22 *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*

$$|f(\boldsymbol{a}) - f(\boldsymbol{a'})| \leq d_i, \tag{5.10}$$

whenever $\boldsymbol{a}$ and $\boldsymbol{a'}$ 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 5.23 (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 $\boldsymbol{X}^{i+1}$ be shorthand notation for the sequence $X_{i+1}, \ldots, X_n$. And let $\boldsymbol{X}^{i+1} = \boldsymbol{a}^{i+1}$ 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 (5.9) holds. To see this, expand using total conditional probability to get

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

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

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

■

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

Corollary 5.23 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 ALC condition. The parameters c_i required for the ALC 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 5.20, The Method Of Average Bounded Differences will be significantly better than that from Corollary 5.23, the Method of Bounded Differences.

Theorem 5.18 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 ALC condition, and hence the probability bound is always stronger.

The price to be paid however, is that both the martingale differences condition and the ALC 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 5.18 and in Corollay 5.20, *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.

5.5 Bibliographic Notes

Martingales ae a classic subject treated in many standard texts on Probability such as Grimmett and Stirzaker [24][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 [48]. Both these are couched in measure–theoretic terminology. A more elementary account can be found in Alon and Spencer [1].

5.6 Problems

Problem 5.24 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 $\boldsymbol{X}$. Show that $\mathtt{E}[X_i X_j] = 0$ if $i \neq j$. ▽

Problem 5.25 (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 5.26 Let $X_0, X_1, \ldots$ be a sequence of random variables with finite means satisfying

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

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 $\boldsymbol{X}$. ▽

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

Problem 5.28 [Vertex Exposure Martingale] Use Definition 5.27 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 5.29 [Azuma generalised further] Show that Azuma's inequality can be generalised to apply to the Definition 5.27 of a martingale. ▽

Problem 5.30 [Azuma and Centering Sequences [49]] A sequence of random variables $X_i, i \geq 0$ is called a *centering sequence* if $\mathtt{E}[X_{i+1} - X_i \mid X_i = t]$ is a non-increasing function of t.

(a) Show that Azuma's inequality applies to a centering sequence with bounded differences.

(b) Let $X_i, i \geq 0$ be the number of red balls in a random sample of size i picked without replacement from n objects r of which are red. Show that the $X_i, i \geq 0$ form a centering sequence and derive a concentration result on X_k for any $k \leq n$.

▽

Problem 5.31 [Negative Regression and MOBD [16]]

(a) Show that the MOBD applies when the underlying variables satisfy the negative regression condition (**??**).

(b) Consider a random sample of size k drawn from n objects r of which are red, and let $X_i, i \leq k$ be the indicator for whether the ith draw was red. Show that $X_1, \cdots X_k$ satisfy $(-R)$ and deduce a sharp concentration on the number of red balls int he sample.

▽

Chapter 6

The Method of Bounded Differences

[The Method of Bounded Differences]

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.

6.1 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 (5.10) 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 6.1 *Remove the factor of* 4 *by applying the method of bounded martingale differences.*

6.2 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 $B_n = B_n(x_1, \dots, x_n)$ 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:

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

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 6.2 *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.)*

6.3 Game Theory and Blackwell's Approachability Theorem

Consider anon-collaborative two-player game given by a matrix $\boldsymbol{M}$ with n rows and m columns. There are two players, the row player and the column player. The row player chooses a row i and simultaneously, the column player chooses a column j. The selected entry $M(i, j)$ is the*loss* suffered by the row player. We asuume for simpliciry that the entries in $\boldsymbol{M}$ are bounded in the range $[0, 1]$.

By standard game theoretic terminology, the choice of a specific row or column is called a *pure* strategy and a distribution over the rows or columns is called a *mixed* strategy. We will use $\boldsymbol{P}$ to denote the strategy of the row player and $\boldsymbol{Q}$ to denote the strategy of the *column* player. $P(i)$ denotes the probability with which row i is selected and similarly $Q(j)$ the probability with which column j is selected. We write $\boldsymbol{M}(\boldsymbol{P}, \boldsymbol{Q}) := \boldsymbol{P}^T \boldsymbol{M} \boldsymbol{Q}$ to denote the expected loss of the row player when the two players use the strategies $\boldsymbol{P}$ and $\boldsymbol{Q}$ respectively.

Consider now a *repeated* play of the game. That is, the two players play a series of *rounds* of interactions. At round $t \ge 0$, the row player picks a row I_t

using strategy P independently of the earlier rounds, and simultaneously, the column player picks a column J_t using startegy Q independently. The total loss suffered by the row player after T rounds is $\sum_{0\leq t\leq T} M(I_t, J_t)$.whose expectation is $T\boldsymbol{M}(\boldsymbol{P}, \boldsymbol{Q})$. Since each netry of M is bounded in $[0, 1]$, changing any one of the underlying choice changes the total loss by at most 1. Hence, applying the MOBD,

$$\mathtt{Pr}[|\sum_{0\leq t\leq T} M(I_t, J_t) - T\boldsymbol{M}(\boldsymbol{P}, \boldsymbol{Q})| > \epsilon T] \leq 2e^{-2\epsilon^2 T},$$

for any $\epsilon > 0$.

A powerful generalization of this setting is that in Blackwell's Approachability Theorem [5], see also [18, 17]. In this case, the payoff $M(i, j)$ is a vector in some compact space. In this space, there is a convex set G called the *target* set. The goal of the row player is to force the average payoff $A_T := \sum_{0\leq t\leq T} M(I_t, J_t)/T$ to approach G arbitrarily closely.

Let $d(A_T, G)$ denote the distance from the average playoff to the closest point in the set G. If $d(A_T, G) \to 0$ almost surely as $T \to \infty$, then the set G is said to be *approachable*. Blackwell gave a necessary and sufficient condition for a convex set to be approchable: *a convex set G is approachable iff every tangent hyperplane to G is approachable* Assuming that any tangent hyperplane is approcahble, it can be shown that $\mathtt{E}[d(A_T, G)] \to 0$. To get the conclusion of Blackwell's Theorem, we then use the MOBD to show that at each time T, $d(A_T, G)$ is sharply concentrated around its expectation. This follows effortlessly from the MOBD: at each stage T, changing any one choice of the strategies so far can change the value of A_T by at most D/T where D is the maximum distance between any two points in the payoff space (finite because of compactness). Thus, $\mathtt{Pr}[|d(A_T, G) - \mathtt{E}[d(A_T, G)]| > t] \leq 2e^{-t^2 T/D}$.

6.4 Balls and Bins

RSA paper improve writeup?

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

Exercise 6.4 *Show that the* Z_i*'s are not independent.*

In view of Exercise 6.4 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:

$$\mathtt{Pr}[|Z - \mathtt{E}[Z]| > t] \leq 2\exp\left(\frac{-t^2}{2m}\right). \tag{6.1}$$

6.5 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 [58, 13]).

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, [7]). 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 [40].

In this section we shall present and analyze two classes of simple distributed algorithms that compute near-optimal edge colourings.Both algorithms proceed in a sequence of rounds. In each round, a simple randomized heuristic is invoked to colour a significant fraction of the edges successfully. This continues until the number of edges is small enough to employ a brute-force method at the final step.

One class of algorithms involves a reduction to bipartite graphs: the graph is split into two parts, T ("top") and B ("bottom") . The induced bipartite graph $G[T, B]$ is coloured using the algorithm P below. then the algorithm is recursively applied to the induced graphs $G[T]$ and $G[B]$ using a fresh set of colours (the same for both). Thus it suffices to describe this algorithm for bipartite graphs.

We describe the action carried out by both algorithms in a single round. For the second class of algorithms, we describe its action only on bipartite graphs (additionally we assume each vertex "knows" if it is "bottom" or "top").

At the start of each round, there is a palette of available coloure $[\Delta]$ where Δ is the maximum degree of the graph at that stage. For simplicity we will assume the graph is Δ-regular.

Algorithm I (Independent): Each edge *independently* picks a *tentative* colour. This tentative colour becomes permanent if there are no conflicting edges that pick the same tentative colour at either endpoint.

Algorithm P(Permutation): This is a two-step protocol:

1. Each bottom vertex. in parallel independently of the others, makes a *proposal* by assigning a *permutation* of the colours to its incident edges chosen uniformly at random.
2. Each top vertex, in parallel, then picks a *winner* out of each set of incident edges that have the same colour. The tentative colour of the winner becomes final. The *losers* i,e, the edges which are not winners are decoloured and passed on to the next round.

For the purposes of the high probbaility analysis, the exact rule used to select winners is not relevant - any rule (deterministic or randomized) that picks one winner out of the degs of a particular colour may be used. This illustrates agian, the power of the martingale method.

It is apparent that both classes of algorithms are *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.

We focus our attention on one round of both algorithms. Let Δ denote the maximum degree of the graph at the start of a round and Δ' the maximum degree of the remaining graph pased on to the next round.. It is easy to show that for both algorithms, $\mathtt{E}[\Delta' \mid \Delta] \leq \beta\Delta$ for some constant $\beta < 1$. For algorithm I, $\beta = 1 - 1/e^2$ whereas for algorithm P, $\beta = 1/e$. The goal is to show that in fact Δ' is sharply concentrated around this value.

For completeness, we sketch a calculation of the total number of colours $\textsc{bc}(\Delta)$ used by algorithm P on a graph with maximum degree Δ: is, with high probability,

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

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 $1.59\Delta + O(\log n)$, [58].

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 [58]. The ease with which the algorithm can be analyses with the method of bounded average differences, as shown below, testifies to its power.

6.5.1 Top vertices

The analysis is particularly easy when v is a top vertex in algorithm P. 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: 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 6.5 *Show that this function has the Lipschitz property with constant* 2. *(Note that this is true regardless of the rule used to select winners.)*

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,

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

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.

6.5.2 The Difficulty with the Other Vertices

The analysis for the "bottom" vertices in algorithm P 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 malicious.

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, which applies also to algorithm I, 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 5.23) would give the bound

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

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

In the next chapter, we shall see how the Method of Averaged Bounded Differences can be applied to get a good concentration bound for the "bottom" vertices.

6.6 Problems

begin new

Problem 6.6 Consider Algorithm I of § 6.5 acting on d-regular graphs with girth at least 4 (the girth of a graph is the length of its smallest cycles). Use the MOBD in simplest form to show that the new vertex degree after one round is sharply concentrated around its expected value (the new vertex degree is given by the edges that do not colour themsleves). $\triangledown$

end new

Problem 6.7 (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,

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

$\triangledown$

Problem 6.8 (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? $\triangledown$

Problem 6.9 [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 $t > 0$
$$\mathtt{Pr}[|f - \mathtt{M}[f]| > t] \le k_3 e^{-\delta_3 t^2}.$$
(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. $\bigtriangledown$

Problem 6.10 [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 $x \le \sqrt{2}$.
(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

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

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

Problem 6.11 [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 $d_H(B_0, B_1) \ge d_H(X, Y)$. 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 $|A_t| \ge (1 - e^{-t^2/2n})2^n$. $\bigtriangledown$

Problem 6.12 [Sampling without replacement] Show that the sequence $\frac{M_i}{N_i}, i \ge 0$ 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. $\bigtriangledown$

Chapter 7

The Method of Averaged Bounded Differences

Sometimes, the function f for which we are trying to show a concentration result does not satisfy the conditions needed to apply the simple MOBD: the Lipschitz coefficients are simply too large in the *worst case*. The function is not "smooth" in this ense in the worst case. We saw this for example in the analysis of the "top" vertices in the distributed edge colouring example. In such cases, the Method of Average Bounded Differences can be deployed needing only an *averged* smoothness condiiton. That is, we need a bound on the follwing averaged smoothness coeffciients:

$$|\mathtt{E}[f \mid \boldsymbol{X}_{i-1}, X_i = a_i] - \mathtt{E}[f \mid \boldsymbol{X}_{i-1}, X_i = a'_i]| \, , \tag{7.1}$$

or, the similar

$$|\mathtt{E}[f \mid \boldsymbol{X}_{i-1}, X_i = a_i] - \mathtt{E}[f \mid \boldsymbol{X}_{i-1}, X_i = a'_i]| \, , \tag{7.2}$$

At first glance, getting a handle on this appears formidable, and indeed it is often non-trivial. We illustrate three main approaches to this:

1. Direct computation is sometimes possible (using linearity of expectation for example).

2. *Coupling* which is a very versatile tool for comparing two closely related distributions such as in (7.1) or (7.2)

3. Bounding the difference by conditioning on the non-occurence of some rare "bad" events.

7.1 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 $|\mathtt{E}[f \mid \boldsymbol{X}_i] - \mathtt{E}[f \mid \boldsymbol{X}_{i-1}]|$, 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

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

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

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

Furthermore

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

Thus we get the bound:

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

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

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

7.2 Occupancy in Balls and Bins

recall the bound:

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

on the concentration of the number of empty bins when we throw m balls independently and uniformly at random into n bins. A better bound can be obtained by applying the method of bounded average differences. Now we need to compute

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

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

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

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

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

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

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

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

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

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

  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 $\texttt{E}[Z_i \mid \boldsymbol{X}_{k-1} = b_{k-1}, X_k = b_k] = 0$ but if $b'_k \notin S$

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

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

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

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

This gives the bound:

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

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

How does it compare to the previous bound?

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

7.3 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 ?

$$\texttt{Pr}[|T_n - \texttt{E}[T_n]| > t] \leq \exp\left(\frac{-t^2}{2\sqrt{2}n}\right). \tag{7.3}$$

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

$$T_n(i) \leq T_n \leq T_n(i) + 2Z_i, \quad i < n, \tag{7.4}$$

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 (7.4), we get:

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

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

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

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

Computing $\mathtt{E}[Z_i \mid \boldsymbol{P}_i]$ 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 $\mathtt{E}[Z_i \mid \boldsymbol{P}_{i-1}]$ is the same, except the point Q is also picked at random. This exercise is relegated to Problem 6.10. The answer is that $\mathtt{E}[Z_i \mid \boldsymbol{P}_i], \mathtt{E}[Z_i \mid \boldsymbol{P}_{i-1}] \leq c/\sqrt{n-i}$ for some constant $c > 0$. Finally, taking the trivial bound $|\mathtt{E}[T_n \mid \boldsymbol{P}_n] - \mathtt{E}[T_n \mid \boldsymbol{P}_{n-1}]| \leq 2\sqrt{2}$ we get

$$\begin{aligned} \mathtt{Pr}[|T_n - \mathtt{E}[T_n]| > t) &\leq 2\exp\left(\frac{-t^2}{2(8 + \sum_{i<n} 4c^2/(n-i))}\right) \\ &\leq 2\exp\left(\frac{-at^2}{\log n}\right), \end{aligned} \tag{7.5}$$

for some $a > 0$. Compare (7.5) to (7.3); 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!

7.4 Coupling

An elegant and effective device to do this is a *coupling*: suppose we can find a joint distribution $\pi(\boldsymbol{Z}, \boldsymbol{Z}')$ such that

1. The marginal distribution $\pi(Z)$ is identical to the distribution of $(\boldsymbol{X} \mid \boldsymbol{X}_{i-1}, X_i = a_i]$, and
2. The marginal distribution $\pi(Z')$ is identical the distribution of $(\boldsymbol{X} \mid \boldsymbol{X}_{i-1}, X_i = a'_i]$

Such a joint distribution is called a *coupling of the two distributions* $(\boldsymbol{X} \mid \boldsymbol{X}_{i-1}, X_i = a_i]$ and $(\boldsymbol{X} \mid \boldsymbol{X}_{i-1}, X_i = a'_i]$.

Then,

$$\begin{aligned}|\mathtt{E}[f \mid \boldsymbol{X}_{i-1}, X_i = a_i] - \mathtt{E}[f \mid \boldsymbol{X}_{i-1}, X_i = a'_i]| =&\\ = \;& |\mathtt{E}_\pi[f(Z)] - \mathtt{E}_\pi[f(Z')]| \\ = \;& |\mathtt{E}_\pi[f(Z) - f(Z')]| \qquad (7.6)\end{aligned}$$

Suppose further that the coupling is chosen well so that $|f(Z) - f(Z')|$ is usually very small. Then we can get a good bound on (7.1). For example, suppose that

1. For any sample point $(\boldsymbol{z}, \boldsymbol{z'})$ which has positive probability in the joint space, $|f(\boldsymbol{z}) - f(\boldsymbol{z'})| \leq d$, and
2. $\mathtt{Pr}[f(Z) \neq f(Z')] \leq p$,

with both d and p "small". Then, from (7.6), we get:

$$|\mathtt{E}[f \mid \boldsymbol{X}_{i-1}, X_i = a_i] - \mathtt{E}[f \mid \boldsymbol{X}_{i-1}, X_i = a'_i]| \leq pd.$$

We will construct such good couplings below. However, first we give some simple examples to get used to the concept of a coupling.

Example 7.1 Suppose we perform two independent trials of tossing a coin n times, the first with a coin of bias p of turning up heads and the second with bias $p' \geq p$. Intuitively it is clear that we expect to get more heads in the second case. To make this rigorous, let $X_1, \cdots, X_n$ be the indicator variables corresponding to getting a heads with the first coin and $X'_1, \cdots, X'_n$ the corresponding ones for the second coin. We would like to assert that for any $t \geq 0$,

$$\mathtt{Pr}[X_1 + \cdots + X_n > t] \leq \mathtt{Pr}[X'_1 + \cdots + X'_n > t].$$

To do this, we will introduce a coupling of the two distributions i.e. we will define a joint distribution $\pi(Z_1, \cdots, Z_n, Z'_1, \cdots, Z'_n)$ such that $\pi(Z_1, \cdots, Z_n)$ has the same distribution as $(X_1, \cdots, X_n)$ and $\pi(Z'_1, \cdots, Z'_n)$ has the same distribution as $(X'_1, \cdots, X'_n)$, and moreover, at each point of the sample space, $Z_i \leq Z'_i, i \in [n]$. Then,

$$\begin{aligned}
\mathtt{Pr}[X_1 + \cdots + x_n > t] &= \pi[Z_1 + \cdots + Z_n > t] \\
&\leq \pi[Z'_1 + \cdots + Z'_n > t] \\
&= \mathtt{Pr}[X'_1 + \cdots + X'_n > t]
\end{aligned}$$

Now for the construction of the coupling. Recall that $\mathtt{Pr}[X_i = 1] = p \leq p' = \mathtt{Pr}[X'_i = 1]$ for each $i \in [n]$. We define the joint distribution $\pi(Z_1, \cdots, Z_n, Z'_1, \cdots, Z'_n)$ be specifying the distribution of each pair (Z_i, Z'_i) independently for each $i \in [n]$. The distribution π is the product of these marginal distributions. For each $i \in [n]$, first toss a coin with bias p of turning up heads. If it shows heads, set $Z_i = 1 = Z'_i$. Otherwise, toss set $Z_i = 0$ and toss another coin with bias $p' - p$ of showing up heads. If this turns up heads, set $Z'_i = 1$, otherwise set $Z'_i = 0$.

It is easy to see that in the distribution π, $Z_i \leq Z'_i$ for each $i \in [n]$. Also, the marginal distributions are as claimed above. ▽

Exercise 7.2 *Generalize the example above in two steps:*

(a) Suppose the probabilities $\mathtt{Pr}[X_i] = p_i \leq p'_i = \mathtt{Pr}[X'_i]$ *are not necessarily all equal. Give the required modification in the above coupling to prove the same result.*

(b) Suppose $X_1, \cdots, X_n$ *and* $X'_1, \cdots, X'_n$ *are distributed in* $[0, 1]$ *and not necessarily identically. However* $\mathtt{E}[X_i] \leq \mathtt{E}[X'_i]$ *for each* $i \in [n]$. *What further modifications are needed now?*

Example 7.3 [Load Balancing] Suppose we throw m balls into n bins in the first experiment and $m' \geq m$ balls in the second. In both cases, a ball is thrown uniformly at random into the n bins and independently of the other balls. Obviously we expect the maximum load to be larger in the second experiment.

To make this rigorous, we construct a coupling π of the two distributions. We may visualize the experiment underlying the coupling as consisting of n bins coloured blue and n bins coloured green both sets labelled $1 \cdots n$ and m balls coloured blue labelled $1 \cdots m$ and m' balls coloured green labelled $1 \cdots m'$. The blue balls will be thrown into the blue bins and the green balls into the green bins. The marginal distribution of the configuration in the blue bins will correspond to our first experiment and the marginal distribution in the green bins to the second experiment. The joint distribution will ensure that a green bin will have at least as many balls as the corresponding blue bin with the same number. Then, if L_m and L'_{m+1} are the maximum loads in the original two experiments respectively and L_b and L_g are the maximum blue and green loads,

$$\begin{aligned} \mathtt{Pr}[L_m > t] &= \pi[L_b > t] \\ &\leq \pi[L_g > t] \\ &= \mathtt{Pr}[L_{m'} > t]. \end{aligned}$$

The coupling itself is easy to describe. First we throw the m blue balls uniformly at random into the n blue bins. Next we place the first m green balls in the green bins as follows: a green ball goes into the green bin with the same number as the blue bin in which the corresponding blue ball went. The remaining $m' - m$ green balls are thrown uniformly at random into the n green bins.

Verfify that the coupling has the two properties claimed. ▽

Exercise 7.4 *Suppose the balls are not identical; ball number k has a probability $p_{k,i}$ of falling into bin number i. Extend the argument to this situation.*

7.5 Distributed Edge Colouring

Recall the distributed edge colouring problem and algorithms from the previous chapter. We applied the simple MOBD successfully to get a srong concentration result for the "top" vertices, but reached an *impasse* with the "bottom" vertices.

We will use the method of bounded average differences to get a strong concentration bound for the "top" vertices as well. 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.

7.5.1 Preliminary Analysis

Let's now move on to the analysis. In what follows, we shall focus on a generic bottom vertex v in algorithm P or an arbitrary vertex in algorithm I. 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

$$\lambda_k := |\mathtt{E}[f \mid \boldsymbol{T}_{k-1}, T_k = c_k] - \mathtt{E}[f \mid \boldsymbol{T}_{k-1}, T_k = c'_k]|. \tag{7.7}$$

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 $e \in N^1(v)$,

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

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. In algorithm I, all edges have their tentative colours assigned independently.

7.5.2 General Vertex in algorithm I

To get a good bound on (7.7), we shall construct a suitable coupling $(\boldsymbol{Y}, \boldsymbol{Y'})$ of the two conditional distributions.

$$(\boldsymbol{T} \mid \boldsymbol{T}_{k-1}, T_k = c_k), (\boldsymbol{T} \mid \boldsymbol{T}_{k-1}, T_k = c'_k)$$

The coupling $(\boldsymbol{Y}, \boldsymbol{Y'})$ is almost trivial: $\boldsymbol{Y}$ is distributed as $\boldsymbol{T}$ conditioned on $(\boldsymbol{T}_{k-1}, T_k = c_k)$ i.e. these settings are fixed as given and the other edges are coloured independently. $\boldsymbol{Y'}$ is distributed identically as $\boldsymbol{Y}$ except that $\boldsymbol{Y'}_k = c'_k$. It is easy to see that the marginal distributions of $\boldsymbol{Y}$ and $\boldsymbol{Y'}$ are exactly the same as the two conditioned distributions $(\boldsymbol{T}_{k-1}, T_k = c_k)$ and $(\boldsymbol{T}_{k-1}, T_k = c'_k)$ respectively.

Now, let us compute $|\mathtt{E}[f(\boldsymbol{Y}) - f(\boldsymbol{Y'})]|$ under this joint distribution. Recall that f was decomposed as a sum $\sum_{v \in e} f_e$ and hence by linearity of expectation, we only need to bound each $|\mathtt{E}[f_e(\boldsymbol{Y}) - f_e(\boldsymbol{Y'})]|$ separately.

First, consider the case when $e_1, \ldots, e_k \in N^2(v)$ i.e. only the choices of the indirect edges have been exposed. Let $e_k = (w, z)$ where w is a neighbour of v. Then, for a direct edge $e \neq vw$, $f_e(\boldsymbol{y}) = f_e(\boldsymbol{y'})$ because under the coupling, $\boldsymbol{y}$ and $\boldsymbol{y'}$ agree on all edges incident on e. So, we only need to compute $|\mathtt{E}[f_{vw}(\boldsymbol{Y}) - f_{vw}(\boldsymbol{Y'})]|$. To bound this simply, note that $f_{vw}(\boldsymbol{y}) - f_{vw}(\boldsymbol{y}) \in \{-1, 0, +1\}$ and that $f_{vw}(\boldsymbol{y}) = f_{vw}(\boldsymbol{y'})$ unless $y_{vw} = c_k$ or $y_{vw} = c'_k$. Thus, we conclude that

$$|\mathtt{E}[f_{vw}(\boldsymbol{Y}) - f_{vw}(\boldsymbol{Y'})]| \leq \mathtt{Pr}[Y_e = c_k \vee Y_e = c'_k] \leq \frac{2}{\Delta}.$$

In fact, one can show by a tighter analysis (see Problem 7.13) that

$$|\mathtt{E}[f_{vw}(\boldsymbol{Y}) - f_{vw}(\boldsymbol{Y'})]| \leq \frac{1}{\Delta}.$$

Now, let us consider the case when $e_k \in N^1(v)$ i.e. the choices of all indirect edges have been exposed and possibly some direct edges as well. In this case, we merely observe that f is Lipschitz with constant 2, and hence, trivially, $|\mathtt{E}[f(\boldsymbol{Y}) - f(\boldsymbol{Y'})]| \leq 2$.

Thus,

$$\lambda_k \leq \begin{cases} \frac{1}{\Delta}; & \text{if } e_k \in N^2(v) \\ 0; & \text{otherwise,} \end{cases}$$

and so,

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

Plugging into the MOABD, we get the following sharp concentration result for the new degee f of an arbitrary vertex in algorithm I:

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

Exercise 7.5 *By regarding f as a function of 2Δ (vector-valued) variables $\boldsymbol{T}(w)$ (which records the colours of all edges incident on w, obtain a similar (but slighly weaker) result using the simple MOBD.*

7.5.3 Bottom Vertex in Algorithm P

Again, to get a good bound on (7.7), we shall construct a suitable coupling $(\boldsymbol{Y}, \boldsymbol{Y'})$ of the two conditional distributions.

$$(\boldsymbol{T} \mid \boldsymbol{T}_{k-1}, T_k = c_k), (\boldsymbol{T} \mid \boldsymbol{T}_{k-1}, T_k = c'_k)$$

This time, the coupling is somewhat more involved.

Suppose e_k is an edge zy where z is a "bottom" vertex. The coupling $(\boldsymbol{Y}, \boldsymbol{Y'})$ is the following: $\boldsymbol{Y}$ is distributed as $(\boldsymbol{T} \mid \boldsymbol{T}_{k-1}, T_k = c_k)$ i.e. it is the product of the permutation distribution resulting from the (possible) conditionings around every bottom vertex. The varaible $\boldsymbol{Y'}$ is *identical* to $\boldsymbol{Y}$ *except on the edges inciodent on z where the colours c_k and c'_k are switched.*

We can think of the joint distribution as divided into two classes: on the degs incident on a vertex other than z, the two variables $\boldsymbol{Y}$ and $\boldsymbol{Y'}$ are identical. So if $v \neq z$, the incident edges have indentical colours under $\boldsymbol{Y}$ and $\boldsymbol{Y'}$ uniformly distributed over all permutations. However, on edges incident on z, the two variables $\boldsymbol{Y}$ and $\boldsymbol{Y'}$ differ on exactly two edges where the two colours c_k and c'_k are switched.

Exercise 7.6 *Verify that the marginal distributions of $\boldsymbol{Y}$ and $\boldsymbol{Y'}$ are $(\boldsymbol{T} \mid \boldsymbol{T}_{k-1}, T_k = c_k)$ and $(\boldsymbol{T} \mid \boldsymbol{T}_{k-1}, T_k = c'_k)$ respectively.*

Now, let us compute $|\mathtt{E}[f(\boldsymbol{Y}) - f(\boldsymbol{Y'})]|$ under this joint distribution. Recall as before that f was decomposed as a sum $\sum_{v \in e} f_e$ and hence by linearity of expectation, we only need to bound each $|\mathtt{E}[f_e(\boldsymbol{Y}) - f_e(\boldsymbol{Y'})]|$ separately.

First, consider the case when $e_1, \ldots, e_k \in N^2(v)$ i.e. only the choices of the indirect edges have been exposed. Let $e_k = (w, z)$ for a neighbour w of v. Note that since

$$\mathtt{E}[f(\boldsymbol{Y}) - f(\boldsymbol{Y'})] = \mathtt{E}[\mathtt{E}[f(\boldsymbol{Y}) - f(\boldsymbol{Y'}) \mid \boldsymbol{Y_e}, \boldsymbol{Y'_e}, z \in e]],$$

it suffices to bound $\left|\mathtt{E}[f(\boldsymbol{Y}) - f(\boldsymbol{Y'}) \mid \boldsymbol{Y_e}, \boldsymbol{Y'_e}, z \in e]\right|$. Recall that $\boldsymbol{Y}_{wz} = c_k$ and $\boldsymbol{Y'}_{wz} = c'_k$. Fix some distribution of the other colours around z. Suppose that $\boldsymbol{Y}_{z,w'} = c'_k$ for some other neighbour w' of z. Then, by our coupling construction, $\boldsymbol{Y'}_{z,w'} = c_k$, and on the remaining edges, $\boldsymbol{Y}$ and $\boldsymbol{Y'}$ agree.. Morover, by independence of the bottom vertices, the distribution on the remaining edges conditioned on the distribution around z is unaffected. Let us denote the joint distribution conditioned on the settings around z by $[(\boldsymbol{Y}, \boldsymbol{Y'}) \mid z]$. Thus, we need to bound $|\mathtt{E}[f(\boldsymbol{Y}) - f(\boldsymbol{Y'}) \mid z]|$

For a direct edge $e \neq vw, vw'$, $f_e(\boldsymbol{y}) = f_e(\boldsymbol{y'})$ becuase in the joint distribution space (even conditioned), $\boldsymbol{y}$ and $\boldsymbol{y'}$ agree on all edges incident on e. So we can concentrate only on $|\mathtt{E}[f_e(\boldsymbol{Y}) - f_e(\boldsymbol{Y'}) \mid z]|$ for $e = vw, vw'$. To bound this simply, observe that for either $e = vw$ or $e = vw'$, first, $f_e(\boldsymbol{y}) - f_e(\boldsymbol{y'}) \in \{-1, 0, 1\}$ and second, that $f_e(\boldsymbol{y}) = f_e(\boldsymbol{y'})$ unless $y_e = c_k$ or $y_e = c'_k$. Thus, we can conclude that

$$\mathtt{E}[f_e(\boldsymbol{Y}) - f_e(\boldsymbol{Y'}) \mid z] \le \mathtt{Pr}[Y_e = c_k \vee Y_e = c'_k \mid z] \le \frac{2}{\Delta}.$$

Taking the two contributions for $e = vw$ and $e = vw'$ together, we finally conclude that

$$|\mathtt{E}[f(\boldsymbol{Y}) - f(\boldsymbol{Y'})]| \le \frac{4}{\Delta}.$$

In fact, one can show by a tighter analysis (see Problem 7.15) that

$$|\mathtt{E}[f_{vw}(\boldsymbol{Y}) - f_{vw}(\boldsymbol{Y'})]| \le \frac{2}{\Delta}.$$

Let us now consider the case when $e_k \in N^1(v)$ i.e. the choices of all indirect edges and possibly some direct edgeshas been exposed. In this case we observe merely that $|f(\boldsymbol{y}) - f(\boldsymbol{y'})| \le 2$ since $\boldsymbol{y}$ and $\boldsymbol{y'}$) differe on exactly two edges. Hence also, $|\mathtt{E}[f(\boldsymbol{Y}) - f(\boldsymbol{Y'})]| \le 2$.

Thus, overall

$$\lambda_k \le \begin{cases} \frac{2}{\Delta}; & \text{if } e_k \in N^2(v) \\ 0; & \text{otherwise,} \end{cases}$$

and,

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

Plugging into the MOABD, we get the following sharp concentration result for the new degee f of an bottom vertex in algorithm P:

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

We observe that the failure probabilities in algorithm I and algorithm P are nearly identical. In particular, for $t := \epsilon\Delta$, both decresae exponentially in Δ.

7.6 Handling Rare Bad Events

In some situations, one can apply the MOABD successfully by bounding the “maximum effect” coefficients but for certain pathological circumstances. Such rare “bad events” can be handled using the following version of the MOABD:

Theorem 7.7 *Let f be a function of n random variables $X_1, \ldots, X_n$, each X_i taking values in a set A_i, such that $\mathtt{E}f$ is bounded. Assume that*

$$m \le f(X_1, \ldots, X_n) \le M.$$

Let $\mathcal{B}$ any event, and let c_i be the maximum effect of f assuming $\mathcal{B}^c$:

$$|\mathtt{E}[f|\boldsymbol{X}_{i-1}, X_i = a_i, \mathcal{B}^c] - \mathtt{E}[f|\boldsymbol{X}_{i-1}, X_i = a_i', \mathcal{B}^c]| \leq c_i.$$

Then,

$$\Pr[f > \mathtt{E}[f] + t + (M - m)\Pr(\mathcal{B})] \leq \exp\left(-\frac{2t^2}{\sum_i c_i^2}\right) + \Pr[\mathcal{B}]$$

and

$$\Pr[f < \mathtt{E}[f] - t - (M - m)\Pr(\mathcal{B})] \leq \exp\left(-\frac{t^2}{\sum_i c_i^2}\right) + \Pr[\mathcal{B}].$$

Proof. We prove the statement for the upper tail. The proof for the lower tail is analogous. For any value $t > 0$,

$$\Pr(f > \mathtt{E}[f] + t) \leq \Pr(f > \mathtt{E}[f] + t \mid \mathcal{B}^c) + \Pr[\mathcal{B}]. \tag{7.8}$$

To bound $\Pr(f > \mathtt{E}[f] + t \mid \mathcal{B}^c)$ we apply Theorem **??** to $(f \mid \mathcal{B}^c)$ and get

$$\Pr(f > \mathtt{E}[f] + t \mid \mathcal{B}^c) \leq \exp\left(-\frac{2t^2}{\sum_i c_i^2}\right). \tag{7.9}$$

Note that all c_is are computed in the subspace obtained by conditioning on $\mathcal{B}^c$. To conclude the proof we show that $\mathtt{E}[f]$ and $\mathtt{E}[f \mid \mathcal{B}^c]$ are very close. Now, since

$$\mathtt{E}[f] = \mathtt{E}[f|\mathcal{B}]\Pr[\mathtt{E}[\mathcal{B}] + \mathtt{E}[f|\mathcal{B}^c]\Pr[\mathcal{B}^c]$$

and $m \leq f \leq M$, we have that

$$\mathtt{E}[f \mid \mathcal{B}^c] - (\mathtt{E}[f \mid \mathcal{B}^c] - m)\Pr[\mathcal{B}] \leq \mathtt{E}[f] \leq \mathtt{E}[f \mid \mathcal{B}^c] + (M - \mathtt{E}[f \mid \mathcal{B}^c])\Pr[\mathcal{B}]$$

so that

$$|\mathtt{E}[f] - \mathtt{E}[f \mid \mathcal{B}^c]| \leq (M - m)\Pr[\mathcal{B}].$$

The claim follows. ∎

The error term $(M - m)\Pr[\mathcal{B}^c]$ in practice is going to be small and easy to estimate, as the next example will make clear. However, using some tricky technical arguments, one can prove [50][Theorem 3.7] the following cleaner statement. For any $t \geq 0$,

Theorem 7.8 *Let f be a function of n random variables $X_1, \ldots, X_n$, each X_i taking values in a set A_i, such that $\mathtt{E}f$ is bounded. Let $\mathcal{B}$ any event, and let c_i be the maximum effect of f assuming $\mathcal{B}^c$:*

$$|\mathtt{E}[f \mid \boldsymbol{X}_{i-1}, X_i = a_i, \mathcal{B}^c] - \mathtt{E}[f \mid \boldsymbol{X}_{i-1}, X_i = a_i', \mathcal{B}^c]| \leq c_i.$$

Then,

$$\Pr(f > \mathtt{E}[f] + t) \leq \exp\left(-\frac{2t^2}{\sum_i c_i^2}\right) + \Pr[\mathcal{B}] \tag{7.10}$$

where again, the maximum effects c_i are those obtained conditioned on $\mathcal{B}^c$.

7.7 Quicksort

We shall sketch the application of the MOABD to Quicksort. This application is interesting because it is a very natural application of the method and yields a provably optimal tail bound. While conceptually simple, the details required to obtain the tighest bound are messy, so we shall confine ourselves to indicating the basic method.

Recall that Quicksort can be modeled as a binary tree T, corresponding to the partition around the pivot element performed at each stage. With each node v of the binary tree, we associate the list L_v that needs to be sorted there. At the outset, the root r is associated with $L_r = L$, the input list,and if the the pivot element chosen at node v is X_v, the lists associated with the left and right children of v are the sublists of L_v consisting of, respectively, all elements less than X_v and all elements greater than X_v (for simplicity, we assume that the input list contains all distinct elements). Now, the number of comparisons performed by Quicksort on the input list L, Q_L is a random variable given by some function f of the random choices made for the pivot elements, $X_v, v \in T$:

$$Q_L = f(X_v, v \in T).$$

We shall now expose the variables $X_v, v \in T$ in the natural top–down fashion: level–by–level and left to right within a level, starting with the root. Let us denote this (inorder) ordering of the nodes of T by $<$. Thus, to apply the Method of Martingale Differences, we merely need to estimate for each node $v \in T$,

$$|\mathtt{E}[Q_L \mid X_w, w < v] - \mathtt{E}[Q_L \mid X_w, w \leq v]|.$$

A moment's reflection shows that this difference is simply

$$|\mathtt{E}[Q_{L_v}] - \mathtt{E}[Q_{L_v} \mid X_v]|,$$

where L_v is the list associated with v as a result of the previous choices of the partitions given by $X_w, w < v$. That is, the problem reduces to estimating the difference between the expected number of comparisons performed on a given list when the first partition is specified and when it is not. Such an estimate is readily available for Quicksort via the recurrence satisfied by the expected value $q_n := \mathtt{E}[Q_n]$, the expected number of comparisons performed on a input list of length n. If the first partition (which by itself requires $n-1$ comparisons) splits the list into a left part of size $k, 0 \leq k < n$ and a right part of size $n-1-k$, the expected number of comparisons is $n-1+q_k+q_{n-1-k}$ and the estimate is:

$$|q_n - (n-1+q_k+q_{n-k-1})| \leq n-1.$$

We shall plug this estimate into the Method of Bounded Differences: thus, if $\ell_v := |L_v|$ is the length of the list associated with node v, then we need to estimate $\sum_v \ell_v^2$. This is potentially problematical, since these lengths are themselves random variables! Suppose, that we restrict attention to levels $k \geq k_1$ for which we can show that

1. $\ell_v \leq \alpha n$ for some parameter α to be chosen later, and
2. k_1 is small enough that the difference between the real process and the one obtained by fixing the values upto level k_1 arbitrarily is negligibly small.

Then summing over all levels $\geq k_1$, level by level,

$$\begin{aligned}
\sum_v \ell_v^2 &= \sum_{k \geq k_1} \sum_{h(v)=k} \ell_v^2 \\
&\leq \sum_{k \geq k_1} \sum_{h(v)=k} \alpha n \ell_v \\
&= \sum_{k \geq k_1} \alpha n \sum_{h(v)=k} \ell_v \\
&\leq \sum_{k \geq k_1} \alpha n^2.
\end{aligned}$$

Next we are faced with yet another problem: the number of levels, which itself is again a random variable! Suppose we can show for some $k_2 > k_1$, that the tree has height no more than k_2 with high probability. Then the previously computed sum reduces to $(k_2 - k_1)\alpha n^2$.

Finally we can apply Theorem 7.8. Here the “bad events” we want to exclude are the event that after k_1 levels, the list sizes exceed α, and that the height of the tree exceeds k_2. all that remains is to choose the parameters careSuppose the maximum size of the list associated with a node at height at least k_1 exceeds αn with probability at most p_1 and that the overall height of the tree exceeds k_2 with probability at most p_2. (One can estimate these probabilities in an elementary way by using the fact that the size of the list at a node at depth $k \geq 0$ is explicitly given by $n \prod_{1 \leq i \leq k} Z_i$, where each Z_i is uniformly distributed in $[0, 1]$.) Then the final result, applying Theorem 7.8 will be:

$$\mathtt{Pr}[Q_n > q_n + t] < p_1 + p_2 + \exp\left(\frac{-2t^2}{(k_2 - k_1)\alpha n^2}\right).$$

(If we applied Theorem 7.7, we would have an additional error term: if we use pessimistic estimates of the maximum and minimum values of the number of comparsions as n^2 and 0 respectively, then the error term is $n^2(p_1 + p_2)$ which is $o(1)$.)

We choose the parameters to optimize this sum of three terms. The result whose details are messy (see [47]) is:

Theorem 7.9 *Let $\epsilon = \epsilon(n)$ satisfy $1/\ln n < \epsilon \leq 1$. Then as $n \to \infty$,*

$$\mathtt{Pr}[|\frac{Q_n}{q_n} - 1| > \epsilon] < n^{-2\epsilon(\ln\ln n - \ln(1/\epsilon) + O(\ln\ln\ln n))}.$$

This bound is slightly better than an inverse polynomial bound and can be shown to be essentially tight [**?**].

7.8 Problems

Problem 7.10 [FKG/Chebyshev Correlation Inequality] Show that for any non-decreasing functions f and g and for any random variable X,

$$\mathtt{E}[f(X)g(X)] \geq \mathtt{E}[f(X)]\mathtt{E}[g(X)].$$

(Hint: Let Y be distributed identical to X but independent of it. Consider $\mathtt{E}[(f(X) - f(Y))\,(g(X) - g(Y))]$. Argue this is non-negative and simplify it using linearity of expectation.) ▽

Problem 7.11 Use coupling to give a simple proof that if a function satisfies the Lipschitz condition with coefficients $c_i, i \in [n]$ then the same bounds can be used with the MOABD i.e. the latter are stronger. Show that the two versions of the latter method differ at most by a factor of 2. ▽

Problem 7.12 [Empty Bins revisited] Rework the concentration of the number of empty bins using a coupling in the method of average bounded differences. ▽

Problem 7.13 Show by a tighter analysis of an arbitrary vertex in algorithm I that

$$|\mathtt{E}[f_{vw}(\boldsymbol{Y}) - f_{vw}(\boldsymbol{Y'})]| \leq \frac{1}{\Delta}.$$

▽

begin new

Problem 7.14 [Kryptographs] The following graph model arises in the context of cryptographically secure sensor networks [60, 57]. We are given a *pool* of cryptographic keys that can be identified with the finite set $P := [m]$, and a set of n vertices. Each vertex i is given a *key ring* S_i generated by sampling P with replacement k times. Two vertices i and j are joined by an edge if and only if $S_i \cap S_j \neq \emptyset$. In the following we assume that $k = \Theta(\log n)$ and $m = \Theta(n \log n)$.

(a) Show that the graph is connected with probability at least $1 - \frac{1}{n^2}$. (Hint: show that, given a set S of vertices, the size of the union of the key rings of vertices in S is not far from its expectation. Using this, show that it is unlikley that G has a cut.)

(b) Using coupling show that the graph is connected with at least the same probability when the key rings are generated without replacement.

$\bigtriangledown$

end new

Problem 7.15 Show by a tighter analysis of a bottom vertex in algorithm P that

$$|\mathtt{E}[f_{vw}(\boldsymbol{Y}) - f_{vw}(\boldsymbol{Y'})]| \leq \frac{2}{\Delta}.$$

$\bigtriangledown$

Problem 7.16 [Concentration for Permutations] Let $f(x_1, \cdots, x_n)$ be a Lipschitz function with constant c i.e. changing any coordinate changes the value of f by at most c. Let σ be a be permutation of $[n]$ chosen uniformly at random. Show a strong concentration for $f(\sigma(1), \cdots, \sigma(n))$. (Hint: Use a natural coupling to bound

$$|\mathtt{E}[f \mid \boldsymbol{X}_{i-1}, X_i = a_i] - \mathtt{E}[f \mid \boldsymbol{X}_{i-1}, X_i = a'_i]|.$$

) $\bigtriangledown$

Chapter 8

The Method of Bounded Variances

In this chapter we describe a tail bound similar in flavour to the Method of Bounded Differences (MOBD). The new bound too rests on a martingale inequality similar to Azuma's. In the previous chapters we saw how, given a function $f(X_1, \ldots, X_n)$, the strength of the MOBD depends on our ability to bound the absolute increments of the Doob martingale sequence $Z_i := \mathtt{E}[f|X_1, \ldots, X_i]$. In doing this, we would expose the variables $X_1, \ldots, X_n$ one at a time and consider the expected change of f when X_i is revealed, conditioned on the values of the $X_1, \ldots, X_{i-1}$ exposed so far, taking the maximum value among all assignments to the first $i-1$ variable. That is, we would look for a bound c_i as small as possible such that,

$$|\mathtt{E}[f|X_1, \ldots, X_i] - \mathtt{E}[f|X_1, \ldots, X_{i-1}]| \leq c_i$$

for all possible assignments to $X_1, \ldots, X_{i-1}$. The resulting bound is

$$\Pr[|X - \mathtt{E}[X]| > t] \leq 2 \exp \left\{ -t^2/2 \sum_i c_i^2 \right\}.$$

We will see in this chapter that basically the same result obtains if we consider the sum of variances of the increments, conditioned on the variables exposed so far:

$$v_i := var(\mathtt{E}[f|X_1, \ldots, X_i] - \mathtt{E}[f|X_1, \ldots, X_{i-1}]).$$

The resulting bound will be,

$$\Pr[|X - \mathtt{E}[X]| > t] \leq 2 \exp \left\{ -t^2/4 \sum_i v_i^2 \right\}$$

assuming some mild conditions on t. Since the variance factors in the probability with which jumps occur, this estimate is often quite sharp. What we will see in this chapter resembles quite closely what we saw in Chapter 1, where we derived the variance bound

$$\Pr[|X - \mathtt{E}[X]| > t] \le 2\exp\left\{-t^2/4\sigma^2\right\}.$$

This bound can be much stronger than the original Chernoff bound and in fact it essentially subsumes it. In practice we will see that good estimates of the variance are not hard to compute. In a sense, the method can be viewed as a quick-and-dirty version of the MOBD. We begin by proving the basic underlying martingale inequality.

8.1 A Variance Bound for Martingale Sequences

We make use of the basic definitions of martingales and their properties developed in Chapter 5. Recall that given a vector $\boldsymbol{X}$ the notation $\boldsymbol{X}_i$ denotes to the truncated vector consisting of the first i coordinates.

Theorem 8.1 *Let $Z_0, Z_1, \ldots, Z_n$ be a martingale w.r.t. the sequence $X_0, X_1, \ldots, X_n$ satisfying the bounded difference condition,*

$$|Z_i - Z_{i-1}| \le c_i$$

$D_i := (Z_i - Z_{i-1}|\boldsymbol{X}_{i-1})$?

for some set of non-negative values c_i. Let

$$V := \sum_{i \le n} v_i$$

where

$$v_i = \sup \operatorname{var}(D_i|\boldsymbol{X}_{i-1})$$

where the sup *is taken over all possible assignments to $\boldsymbol{X}_{i-1}$. Then,*

$$\Pr(Z_n > Z_0 + t) \le \exp\left(-\frac{t^2}{4V}\right)$$

and

$$\Pr(Z_n < Z_0 - t) \le \exp\left(-\frac{t^2}{4V}\right)$$

provided that

$$t \le 2V/\max_i c_i. \tag{8.1}$$

Proof. The initial part of the proof is identical to that of Theorem 5.17 but we reproduce it here for ease of exposition. It suffices to prove the statement for the upper tail. The proof for the lower tail is symmetrical with the martingale $\boldsymbol{Z}$ replaced by $-\boldsymbol{Z}$.

Assume without loss of generality that $Z_0 := 0$ and define the martingale difference sequence $D_i := Z_i - Z_{i-1}, i \geq 1$. Then $Z_n = Z_{n-1} + D_n$. Note that $\mathtt{E}[D_i] = 0$, for all i. By Markov's inequality,

$$\Pr(Z_n > t) \leq \min_{\lambda > 0} \frac{\mathtt{E}[e^{\lambda Z_n}]}{e^{\lambda t}}. \tag{8.2}$$

With foresight we set,

$$\lambda := \frac{t}{2V}. \tag{8.3}$$

As usual we focus on the numerator $\mathtt{E}[e^{\lambda Z_n}]$ and seek a good upper bound for it.

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

We now show that, for all i,

$$\mathtt{E}[e^{\lambda D_i} | \boldsymbol{X}_{i-1}] \leq e^{\lambda^2 v_i}. \tag{8.4}$$

Assuming this, it follows by induction that,

$$\begin{aligned} \mathtt{E}[e^{\lambda Z_n}] &= \mathtt{E}[e^{\lambda Z_{n-1}} \mathtt{E}[e^{\lambda D_n} | \boldsymbol{X}_{n-1}]] \\ &\leq \mathtt{E}[e^{\lambda Z_{n-1}}] e^{\lambda^2 v_n} \\ &\leq e^{\lambda^2 V}. \end{aligned}$$

The claim then follows by induction. The base case is the trivial case $Z_0 = 0$. Using our choice for λ and the bound on the numerator it follows that,

$$\begin{aligned} \Pr(Z_n > t) &\leq \min_{\lambda > 0} \frac{\mathtt{E}[e^{\lambda Z_n}]}{e^{\lambda t}} \\ &\leq \frac{e^{\lambda^2 V}}{e^{\lambda t}} \\ &= e^{t^2/2V}. \end{aligned}$$

The crux then is to establish (8.4). This follows from the well-known inequalities $1 + x \leq e^x$, valid for all x, and $e^x \leq 1 + x + x^2$, valid for $|x| \leq 1$. Since $\boldsymbol{Z}$ is a martingale with respect to $\boldsymbol{X}$, $\mathtt{E}[D_i|\boldsymbol{X}_{i-1}] = 0$. Now, if $\lambda|D_i| \leq 1$ then,

$$\begin{aligned}
\mathtt{E}[e^{\lambda D_i}|\boldsymbol{X}_{i-1}] &\leq \mathtt{E}[1 + \lambda D_i + (\lambda D_i)^2|\boldsymbol{X}_{i-1}] \\
&= 1 + \lambda^2 \mathtt{E}[D_i^2|\boldsymbol{X}_{i-1}] \\
&= 1 + \lambda^2 v_i \\
&\leq e^{\lambda^2 v_i}.
\end{aligned}$$

The condition $\lambda|D_i| \leq 1$ follows, for all i, from the hypothesis 8.1 and Equation 8.3. The claim follows. $\blacksquare$

Should check this

A couple of observations are in order. First, the term V is related to the variance of Z_n in the following way: $E[V] = var(Z_n)$ (see Problem section). Second, the condition on t roughly says that this inequality is a bound for deviations that are "not too large". By using Bernstein's estimate (1.4) it is possible to obtain the following slightly sharper bound without making any assumptions on t.

$$\Pr(Z_n > Z_0 + t) \leq \exp\left(-\frac{t^2}{2V(1 + bt/3V)}\right). \tag{8.5}$$

check dev mess

The term b is defined as the $\max_k dev_k$, where $dev_k := \sup\{(Z_k - Z_{k-1}|X_1 = x_1, \ldots, X_{k-1} = x_{k-1}\}$. In some situations the error term bt/V is negligible, and (8.5) yields a slightly sharper bound than that of Theorem 8.2. The interested reader can refer for example to [50].

The next step is to package this inequality in a form suitable for the applications. Note that the ground variables X_is need not be independent for the next theorem to hold.

Correct?

Theorem 8.2 *[The Method of Bounded Variances] Let $X_1, \ldots, X_n$ be an arbitrary set of random variables and let $f := f(X_1, \ldots, X_n)$ be such that $\mathtt{E}f$ is finite. Let*

$$D_i := \mathtt{E}[f|\boldsymbol{X}_i] - \mathtt{E}[f|\boldsymbol{X}_{i-1}]$$

and let $c_1, \ldots, c_n$ be such that

$$|D_i| \leq c_i. \tag{8.6}$$

And let

$$V := \sum_{i=1}^{n} v_i$$

where

$$v_i := \sup \operatorname{var}(D_i|\boldsymbol{X}_{i-1})$$

with the sup *taken over all possible assignment to* $\boldsymbol{X}_{i-1}$*. Then,*

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

and

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

provided that $t \leq 2V/\max_i c_i$*.*

Proof. Apply Theorem 8.1 to the Doob martingale sequence $Z_i := \mathtt{E}[f|\boldsymbol{X}_i]$. Problem 8.8. ∎

Intuitively, when applying this inequality we will expose, or *query*, the values of the variables X_i one by one, starting from $\mathtt{E}[f]$ and ending with $f(X_1, \ldots, X_n)$. As we shall see, the power of this inequality derives from the fact that it is possible, and sometimes easy, to give good estimates of the v_is. Note that one has the freedom to decide the sequence according to which the variables are exposed. This will be put to good effect in the applications to follow.

Notation 8.3 *In the sequel we shall refer to* V *as the* variance *of* f *and to* c_i *as the* maximum effect *of the* i*th query.*

8.2 Applications

As usual, the best approach to understand the method is by means of examples of increasing sophistication. For the method to be useful one needs simple ways to bound the variance. A simple but useful bound is the following. Assume that a random variable X is such that $\mathtt{E}[X] = 0$ and $|X| \leq r$. Then,

$$var(X) \leq \frac{r^2}{4}. \tag{8.7}$$

(See Problem 8.7).

With this we can essentially recover the basic version of the MOBD (Theorem 5.18). We are given a function $f(X_1, \ldots, X_n)$ satisfying the conditions

$$|f(X) - f(X')| \leq c_i$$

for each i, whenever X and X' differ only in the ith coordinate. We apply 8.7 to the random variable $D_i := \mathtt{E}[f|X_1, \ldots, X_i] - \mathtt{E}[f|X_1, \ldots, X_{i-1}]$ which has zero mean. Thus

$$var(D_i) \leq \frac{c_i^2}{4}.$$

Therefore $V \leq \frac{1}{4}\sum_i c_i^2$ and by Theorem 8.2,

$$\Pr[f > \mathtt{E}f + t] \leq \exp\left(-\frac{t^2}{4V}\right) \leq \exp\left(-\frac{t^2}{\sum_i c_i^2}\right)$$

provided that $t \leq 2V/\max_i c_i$.

Exercise 8.4 *Establish the basic Chernoff-Hoeffding bounds by using the Method of Bounded Variances.*

The next example is to derive the variance bound of the basic Chernoff-Hoeffding bounds that we developed in § 1.7. We are given n independent random variables $X_i \in [0,1]$ and we want to prove that,

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

where $X := \sum_i X_i$, $\sigma^2 := \sum_i \sigma_i^2$ and $\sigma_i^2 := var(X_i)$. We apply the method to $f(X_1, \ldots, X_n) := \sum_i X_i$. Now, if we set

$$Z_i := \mathtt{E}[f|X_1, \ldots, X_i]$$

we have that

$$|Z_i - Z_{i-1}| \leq 1.$$

Furthermore, by independence we get

$$D_i := Z_i - Z_{i-1} = X_i - \mathtt{E}[X_i]$$

and

$$var(D_i) = var(X_i)$$

and thus $V = \sum_i var(X_i) = \sigma^2$. Therefore, by invoking Theorem 8.2,

$$\Pr[f > \mathtt{E}f + t] \leq \exp\left(-\frac{t^2}{4V}\right) = \exp\left(-\frac{t^2}{4\sigma^2}\right)$$

if $t \leq 2V$. In the case when $\Pr[X_i = 1] = p$, for all i, we have by independence that $\sigma^2 = np(1-p)$ and the bound becomes,

$$\Pr[f > \mathtt{E}f + t] \leq e^{-t^2/4np(1-p)}.$$

The variance is maximized when $p = \frac{1}{2}$ so that $\sigma^2 \leq \frac{n}{4}$, which gives

$$\Pr[f > \mathtt{E}f + t] \leq \exp^{-t^2/n}.$$

This bound loses a factor of two in the exponent. By applying the slightly sharper bound 8.5 one essentially recovers 1.6.

8.2.1 Bounding the Variance

We saw that (8.7) gives a simple but useful bound for the variance that always applies. A common situation that arises when studying a function $f(X_1, \ldots, X_n)$ is when each X_i takes on two values, a "good" value with probability p_i, and a "bad" value with probability $1 - p_i$. In this case, assuming that X_i takes values in a finite set A_i, we get the following useful bound:

$$v_i \leq p_i(1 - p_i)c_i^2 \tag{8.8}$$

so that

$$\Pr[f > \mathtt{E}f + t] \leq \exp^{-t^2/\sum_i p_i(1-p_i)c_i^2} \tag{8.9}$$

and

$$\Pr[f < \mathtt{E}f - t] \leq \exp^{-t^2/\sum_i p_i(1-p_i)c_i^2} . \tag{8.10}$$

Does it hold also for finite intervals? Exercise?

Bound (8.8) follows from elementary, but non-trivial computations (see Problem 8.16).

Let us apply this bounding technique to the following problem. We are given a d-regular, undirected graph $G = (V, E)$. Consider again Algorithm I from § 7.5. Each edge is given a list of c colors and the following simple, distributed algorithm is executed. Each edge picks a tentative color at random, uniformly from its list. If there is a conflict– two neighbouring edges make the same tentative choice– the color is dropped, and the edge will try to color itself later. Otherwise, the color becomes the final color of the edge. At the end of the round, the lists are updated in the natural way, by removing colors succesfully used by neighbouring edges. Edges that succesfully color themselves are removed from the graph. The process is repeated with the left-over graph and left-over color lists until all edges color or the algorithm gets stuck because some list runs out of colors.

It is possible to show that, for any $\epsilon > 0$, if $d \gg \log n$ and $c = (1 + \epsilon)d$ this simple algorithm will, with high probability, color all edges of the graph within $O(\log n)$ many rounds [13, 22]. Since clearly d colors are needed to edge color the graph, this shows that one can obtain nearly-optimal edge-colorings by means of this very simple and inexpensive distributed procedure. Here we analize the first round of the algorithm to show that the degree of each vertex is sharply concentrated around its expectation. The discussion exemplifies some of the important points of the full analysis. For simplicity we assume $c = d$.

Fix a vertex u. For this vertex we want to show that its new degree is sharply concentrated around its expected degree. The probability that an edge e is colored is the probability that no neighbouring edge picks the same tentative color,

$$\Pr[e \text{ colors}] = \left(1 - \frac{1}{c}\right)^{2d-1} \sim \frac{1}{e^2}.$$

Therefore, if we denote by Z the number of edges that succesfully color,

$$\mathtt{E}[Z] \sim \frac{d}{e^2} = \Theta(d).$$

We are interested in the variable $d-Z$ and if we show that Z is concentrated, so is $d-Z$. The variable Z is a function of the random choices made not only by the edges incident on u, but also those of the edges incident on the neighbours of u. Let $E(u)$ denote this set, and let X_e denote the random color chosen by an edge e. Before applying the new method it is worth asking why we cannot use the inequalities that we know already. Let us introduce an indicator random variable Z_e for each edge incident on u denoting whether e successfully color. Thus,

$$X = \sum_{e \ni u} X_e.$$

These indicator variables are clearly not independent, so one cannot apply the Chernoff-Hoeffding bounds. Can we apply the MOBD in its simplest form? With our notation, we have

$$Z = f(X_e : \ e \in E(u)).$$

Clearly, for each $e \in E(u)$ the best we can say is,

$$|f(X_1, \ldots, X_{e-1}, X_e, X_{e+1}, \ldots, X_D) - |f(X_1, \ldots, X_{e-1}, X'_e, X_{e+1}, \ldots, X_D)| \leq 1$$

where $D := |E(u)| = d(d-1)$. This gives a very weak bound,

$$\Pr[|Z - \mathtt{E}[Z]| > t) \leq 2\exp(t^2/2d^2). \tag{8.11}$$

An alternative is to use the MOBD in expected form. As we saw in § 7.5, this works but it requires somewhat lengthy calculations. An easy way out is given by the Method of Bounded Variances. We query the edges in this order. First, we expose the choices of every edge incident on u. Each such edge can affect the final degree by at most 1. Then we expose the random choices of the remaining edges. They key observation is the following. Let $e = wv$ be the edge we are considering and let $f = vu$ denote an edge incident on u that touches e. Note that when we expose e's tentative choice, f has already been queried. The choice of e can affect f, but only if e picks the same color chosen by f and this happens with probability $1/c = 1/d$. Therefore, the variance of this choice is at most $1/c = 1/d$ (e can touch two edges incident on u, but its effect can be at most 1. Why?). Therefore, we can bound the total variance as follows,

$$d \cdot 1 + d(d-1)\frac{1}{c} \leq 2d$$

which gives, for $t \leq 2V$,

$$\Pr[|Z - \mathtt{E}[Z]| > t] \leq 2\exp(t^2/8d) \tag{8.12}$$

a much stronger bound than Equation 8.11.

8.2.2 Dealing with Unlikely Circumstances

Sometimes the effect of a random variable X_i on the value of a function $f(X_1, \ldots, X_n)$ can be quite large, but only with very low probability. In other words, it might be the case that for most outcomes of $X_1, \ldots, X_n$ the variance of f is very small. In dealing with such situations the following result comes handy. In the statement of the next theorem the event $\mathcal{B}$ is to be understood as a set of exceptional outcomes of very low probability.

Theorem 8.5 *Let f be a function of n random variables $X_1, \ldots, X_n$, each X_i taking values in a set A_i, such that $\mathtt{E}f$ is bounded. Assume that*

$$m \leq f(X_1, \ldots, X_n) \leq M.$$

Let $\mathcal{B}$ any event, and let V and c_i be, respectively, the variance and the maximum effects of f assuming $\mathcal{B}^c$. Then,

What is c_i doing here??

$$\Pr[Z_n > Z_0 + t + (M - m)\Pr(\mathcal{B})] \leq \exp\left(-\frac{t^2}{4V}\right) + \Pr[\mathcal{B}]$$

What is Z_n wrt f??

and

$$\Pr[Z_n < Z_0 - t - (M - m)\Pr(\mathcal{B})] \leq \exp\left(-\frac{t^2}{4V}\right) + \Pr[\mathcal{B}].$$

Proof. We prove the statement for the upper tail. The proof for the lower tail is analogous. For any value T,

$$\Pr(f > \mathtt{E}[f] + T) \leq \Pr(Z_n > Z_0 + T|\mathcal{B}^c) + \Pr[\mathcal{B}]. \tag{8.13}$$

To bound $\Pr(f > \mathtt{E}[f] + T|\mathcal{B}^c)$ we apply Theorem 8.2 to $(f|\mathcal{B}^c)$ and get

t and T??

$$\Pr(f > \mathtt{E}[f|\mathcal{B}^c] + T|\mathcal{B}^c) \leq \exp\left(-\frac{t^2}{4V}\right) \tag{8.14}$$

for every $t \leq 2V/\max c_i$, provided that V and all c_is are computed in the subspace obtained by conditioning on $\mathcal{B}^c$. To conclude the proof we show that $\mathtt{E}[f]$ and $\mathtt{E}[f|\mathcal{B}^c]$ are very close. Now, since

$$\mathtt{E}[f] = \mathtt{E}[f|\mathcal{B}]\Pr[\mathtt{E}[\mathcal{B}] + \mathtt{E}[f|\mathcal{B}^c]\Pr[\mathcal{B}^c]$$

and $m \leq f \leq M$, we have that

$$\mathtt{E}[f|\mathcal{B}^c] - (\mathtt{E}[f|\mathcal{B}^c] - m)\Pr[\mathcal{B}] \leq \mathtt{E}[f] \leq \mathtt{E}[f|\mathcal{B}^c] + (M - \mathtt{E}[f|\mathcal{B}^c])\Pr[\mathcal{B}]$$

so that

$$|\mathtt{E}[f] - \mathtt{E}[f|\mathcal{B}^c]| \leq (M - m)\Pr[\mathcal{B}].$$

The claim follows. ∎

The error term $(M - m)\Pr[\mathcal{B}^c]$ in practice is going to be a $o(1)$ and easy to estimate, as the next example will make clear. It is possible to prove the following cleaner statement. For any $t \leq 2V/\max c_i$,

lower tail also?

$$\Pr(f > \mathtt{E}[f] + t) \leq \exp\left(-\frac{t^2}{4V}\right) + \Pr[\mathcal{B}] \tag{8.15}$$

where V is the variance and the maximum effects c_i are those obtained conditioned on $\mathcal{B}^c$. The proof however is not as simple as that of Theorem 8.5, while this formulation in practice is not any stronger, at least for the kind of applications that one normally encounters in the analysis of algorithms.

Let us see a non trivial application of Theorem 8.5. We have a d-regular graph G in which each vertex is given a list of c colors. We consider the same simple distributed algorithm of the previous section, this time applied to the vertices instead. It is possible to prove that this algorithm computes a vertex coloring with high probability in $O(\log n)$ many rounds, provided that G has no triangles, $d \gg \log n$ and $c = \Omega(d/\ln d)$ [23]. Note that c can be much smaller than d. More generally, it is known that such good colorings exist for triangle-free graphs, and that this is the best that one can hope for, since there are infinite families of triangle-free graphs whose chromatic number is $\Omega(d/\ln d)$ [35, 9]. In what follows we assume for simplicity that $c = d/\log_2 d$.

Here we analize what happens to the degree of a vertex after one round and show that it is sharply concentrated around its expectation. As with the previous example this will exemplify some of the difficulties of the full analysis. Let us fix a vertex u and let Z be the number of neighbours of u which color themselves succesfully. We first compute $\mathtt{E}[Z]$. The probability that a vertex colors itself is,

$$\left(1 - \frac{1}{c}\right)^d \sim e^{-d/c}$$

so that

$$\mathtt{E}[Z] \sim de^{-d/c}$$

We are interested in a bound on the probability that the new degree $d' := d - Z$ is far from its expectation. We will show that this happens only with inverse polynomial probability in d. As with the edge coloring example the value of Z depends not only on the tentative color choices of the neighbours of u but also on those of the neighbours of the neighbours– $\Theta(d^2)$ choices in total. To compund the problem, vertices at distance two can now have very large effects. Assume for instance that all neighbours of u pick color a tentatively. If a vertex w at distance 2 from u also picks a the effect can be as large as $|N_u \cap N_w|$ which, in general, can be as large as d. We can get around this problem using the fact that it is unlikely that "many" neighbours of u will pick the same color.

Let $N_i(u)$ denote the set of vertices at distance i from u. Z depends on the choices of vertices in $\{u\} \cup N_1(u) \cup N_2(u)$. We expose the color choices in this order. First u, then the vertices in $N_1(u)$ (in any order) and finally those in $N_2(u)$ (in any order). The first query does not affect the variance. The next d queries can each affect the final outcome of Z by one, but note that this is only if the vertex selects the same tentative color of u, an event that occurs with probability $1/c$. The total variance after these queries is then at most,

$$0 + \sum_{x \in N_1(u)} v_x \leq \frac{d}{c}.$$

So far so good, but we now need to estimate the total variance of vertices in $N_2(u)$ and we know that this can be extremely large in the worst case. We exploit the fact that the tentative color choices of the neighbours of u are binomially distributed. Fix $w \in N_2(u)$ and let

$$x := |N_u \cap N_w|.$$

Moreover let x_a denote the number of vertices in $N_u \cap N_w$ that choose color a tentatively. For each color a, the expected number of u-neighbours that pick a is d/c. The set of "bad" events $\mathcal{B}$ that we are going to consider is when there exists a color that is chosen more than $r_\delta := (1+\delta)d/c$ times. By the Chernoff and the union bounds, for any $\delta > 2e - 1$, the probability of $\mathcal{B}$ is at most $c2^{-r_\delta} = c2^{-(1+\delta)d/c}$. Note that this gives an "error term" of at most $d\Pr[\mathcal{B}] = dc2^{-(1+\delta)d/c} = o(1)$.

We now want to estimate the variance assuming the "good" event $E\mathcal{B}^c$. Let

$$d_a := |\mathtt{E}[Z|X_1, \ldots, X_w = a] - \mathtt{E}[Z|X_1, \ldots, X_{w-1}]| \leq x_a$$

and thus

$$v_w = \sum_a \Pr[X_w = a] d_a^2 \leq \frac{1}{c} \sum_a x_a^2.$$

This sum of squares is subject to

$$x = \sum_a x_a$$

and, by the previous assumption,

$$0 \leq x_a \leq r_\delta,$$

The maximum is therefore attained at the extreme point, when there are x/r_δ terms, each equal to r_δ^2 (see problem section). Therefore,

$$v_w \leq \frac{1}{c}\frac{x}{r_\delta}r_\delta^2 = \frac{(1+\delta)xd}{c^2}.$$

The total variance of vertices in $N_2(u)$ is $\sum_{w\in N_2(u)} v_w$. If we assign a weight of v_w/x on each edge between w and $N_1(u)$, we then have

$$\sum_{w\in N_2(u)} v_w = \sum_{wv:w\in N_2(u),v\in N_1(u)} \frac{v_w}{x} \leq d(d-1)\frac{(1+\delta)d}{c^2}$$

for a total variance of

$$V \leq 0 + \frac{d}{c} + d(d-1)\frac{(1+\delta)d}{c^2} \leq \frac{(1+\delta)d^3}{c^2}.$$

Therefore, if $\delta \geq 2e-1$,

$$\Pr[Z - \mathtt{E}[Z] > t + o(1)] \leq e^{-t^2/4V} + c2^{-(1+\delta)d/c}$$

provided that $t \leq 2V$. If $c = \Theta(d/\ln d)$ and $t = \Theta(V\ln d)$, this says that Z deviates from its expectation by more than $\Theta(\sqrt{d\ln^3 d})$ with inverse polynomial probability. An analogous derivation establishes the result for the lower tail.

8.3 Bibliographic Notes

A good source for coloring problems of the type discussed here is the book of Molloy and Reed [53]. McDiarmid's survey presents a treatment of some of the inequalities that we discussed, with several useful variations on the theme [50]. The basic result was established in [3] for 0/1-random variables and was later extended to the case of multi-way choices in [21]. The edge coloring application can be found in [22].

8.4 Problems

Problem 8.6 With reference to the statement of Thereom 8.1, show that

$$E[V] = var(Z_n).$$

▽

Solution. Page 224 of McDiarmid's survey △

Problem 8.7 Prove that if a random variable X satisfies $\mathtt{E}X = 0$ and $a \leq X \leq b$, then

$$var(X) \leq (b-a)^2/4.$$

▽

Solution. If a random variable X satisfies $\mathtt{E}X = 0$ and $a \leq X \leq b$, then

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

△

Problem 8.8 Prove Theorem 8.2 (Hint: Define the Doob martingale sequence $Z_i := \mathtt{E}[f|X_0, \ldots, X_i]$ and observe that $(D_i|\boldsymbol{X}_{i-1}) = D_i$. Apply Theorem 8.1). ▽

Problem 8.9 Prove Equation 8.8, i.e.

$$v_i := \sum_{a \in A_i} \Pr[X_i = a](D_i^2|X_i = a) \leq p(1-p)c_i^2$$

uner the hypothesis that A_i can be partitioned into two regions $A_i = G \cup B$, such that $\Pr[X_i \in G] = p$. ▽

Solution. Refer to [13]. △

Problem 8.10 Establish the bound in Equation 8.12 by using the MOBD in expected form. ▽

Problem 8.11 Let $G = (V, E)$ be a d-regular graph with n vertices. Consider the following algorithm for computing independent sets. Let $p : V \to [n]$ be a random permutation of the vertices. A vertex i enters the independent set if and only if $p_i < p_j$ for every j neighbour of i (the set so computed is clearly independent). Let X denote the size of the resulting independent set. Compute $\mathtt{E}X$ and show that X is concentrated around its expectation. ▽

Problem 8.12 Show a bound similar to Equation 8.12 for the edge coloring problem discussed in § 8.2.1. ▽

Problem 8.13 Repeat the analysis of Section 8.2.1 under the hypothesis $c = (1+\epsilon)d$. ▽

Problem 8.14 Show that

$$\max \sum_i x_i^2$$

subject to $\sum_i x_i = n$ and $0 \leq x_i \leq c$ is attained when $\lfloor \frac{n}{c} \rfloor$ terms are set equal to c and the remaining terms are set to 0. ▽

Problem 8.15 Let $X_1, \ldots, X_n$ be independent, with $a_k \leq X_k \leq b_k$ for each k where a_k and b_k are constants, and let $X := \sum_i X_i$. Prove that then, for any $t \geq 0$,

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

▽

Problem 8.16 Prove Equation 8.8. ▽

Problem 8.17 Consider the edge coloring algorithm described in § 8.2.1.

- Compute the expected number of colors that remain available for an edge.
- Show that this number is sharply concentrated around its expectation.
- Do the same for the intersection of the color lists of two edges incident upon the same vertex.

▽

Problem 8.18 Let G be a d-regular graph and consider the following randomized algorithm to compute a matching in the graph. Every edge enters a set S with probability $\frac{1}{d}$. If an edge in S does not have any neighbouring edges in S it enters the matching M. Edges in M and all their neighbours are removed from G.

- Compute the expected degree of a vertex that is not matched.
- Use the Method of Bounded Variances to prove that the degree of a vertex that is not matched is concentrated around its expectation. Can you use the MOBD in its simplest form?

- Show that the same is true if the above algorithm is repeated. How large a value of d (as a function of n) is needed for concentration to hold?

The point here is to show that the graph stays essentially regular during the execution of the algorithm, as long as the average degree is high enough. $\bigtriangledown$

Problem 8.19 [23]. We are given a d-regular graph G of girth at least 5 where each vertex has a list of $c := d/\log_2 d$ colours (the girth of a graph is the length of its smallest cycle). Consider the following algorithm. Each vertex *wakes up* with probability $p := 1/\log_2 d$. Each awaken vertex picks a tentative colour at random from its own list and checks for possible colour conflicts with the neighbours. If none of the neighbours pick the same tentative colour, the colour becomes final. If a colour c becomes the final colour of a neighbour of a vertex u, c is deleted from u's colour list.

- For a given vertex, let X be the number of its uncoloured neighbours. Prove that X is concentrated around its expectation.
- For a given vertex, let Y be the number of colours not chosen by its neighbours. Prove that Y is concentrated around its expectation.
- For given vertex u and colour c, let Z be the number of uncoloured neighbours of u that retain c in their list. Prove that Z is concentrated around its expectation.

$\bigtriangledown$

Chapter 9

The Infamous Upper Tail

9.1 Motivation: Non-Lipschitz Functions

Consider the random graph $G(n,p)$ with $p = p(n) = n^{-3/4}$. Let X be the number of triangles in this graph. We have $\mathtt{E}[X] = \binom{n}{3}p^3 = \Theta(n^{3/4})$. The randomvariable X is a function of the $\binom{n}{2}$ independent variables corresponding to whether a particular edge is present or not. Changing any of these variables could change the value of X by as much as $n-2$ in the worst case. Applying the MOBD with these Lipschitz coefficients is useless to obtain a non-trivial concentration result for deviations of the order of $\epsilon \mathtt{E}[X] = \Theta(n^{3/4})$ for a fixed $\epsilon > 0$.

Exercise 9.1 *Try to apply the MOABD or the MOBV and see if you get any meaningful results.*

The essential problem here is that the function under consideration is not Lipschitz with sufficiently small constants to apply the method of bounded differences. This initiated a renewed interest in methods to prove concentration for functions which are not "smooth" in the worst case Lipschitz sense of the MOBD but are nevertheless "smooth" in some "average" sense. We have already seen that the MOABD and MOBV are such methods. Here we briefly describe two new methods that apply well to problems such as counting triangles in the random graph.

9.2 Concentration of Multivariate Polynomials

Let $X_{i,j}$ be the indicator random variable for whether the edge (i,j) is included in the random graph $G(n,p)$. Then, the number of triangles X_{K_3} in $G(n,p)$ can be written as

$$X_{K_3} = \sum_{1 \le i < j < k \le n} X_{j,k} X_{k,i} X_{i,j}.$$

Formally, this can be seen as a multivariate polynomial in the $\binom{n}{2}$ variables $X_{i,j}$, and motivates the setting of the Kim-Vu inequality.

Let U be a base set and let $\mathcal{H}$ be a family of subsets of U of size at most k for some $0 < k \le n$. Let $X_u, u \in U$ be independent $0/1$ random variables with $\mathtt{E}[X_u] = p_u, u \in U$. Consider the function of $X_u, u \in U$ given by the following multi-variate polynomial:

$$Z := \sum_{I \in \mathcal{H}} w_I \prod_{u \in I} X_u.,$$

where $w_I, I \in \mathcal{H}$ are positive coefficients. In the example of the triangle above, the base set U is the set of all $\binom{n}{2}$ edges and the family $\mathcal{H}$ consists of the $\binom{n}{3}$ 3-element subsets of edges that form a triangle (so $k = 3$ and all coefiicients $w_I = 1$).

For each subset A of size at most k, define a polynomial Y_A as follows:

$$Z_A := \sum_{A \subseteq I \subseteq \mathcal{H}} w_I \prod_{u \in I \setminus A} X_u.$$

Formally, this is the partial derivative $\frac{\partial Z}{\partial X_A}$. Set

$$E_j(Z) := \max_{|A| \ge j} \mathtt{E}[Z_A], \quad 0 \le j \le k.$$

Heuristically, $E_j(Z)$ can be interpreted as the maximum *average* effect of a group of at least j underlying variables. – this will play the role of "average" Lipschitz coefficients in place of the worst case Lipschitz coefficients.

Theorem 9.2 (Kim-Vu Multivariate Polynomial Inequality) *For any $k \le n$, there are positive numbers a_k, b_k such that for any $\lambda \ge 1$*

$$\mathtt{Pr}\left[|Z - \mathtt{E}[Z]| \ge a_k \lambda^k \sqrt{E_0(Z) E_1(Z)}\right] \le b_k \exp\left\{-\lambda/4 + (k-1)\log n\right\}.$$

(For definiteness, we can take $a_k := 8^k \sqrt{k!}$ and $b_k := 2e^2$.)

To apply this to the number of triangles in the random graph $G(n,p)$, consider the base set $\binom{n}{2}$, take $\mathcal{H}$ to be the family of 3-element subsets forming a triangle and consider the multivariate polynomial:

$$Z := \sum_{1 \leq i < j < \ell \leq n} X_{j,\ell} X_{\ell,i} X_{i,j},$$

and $X_{i,j}$ is the indicator variable for whether the edge (i,j) is included. As we saw, with $p = n^{-3/4}$, we have $\mathtt{E}[Z] = \Theta(n^{3/4})$..

Now, if $A = \{i,j\}$, we have:

$$Z_A = \sum_{\ell \neq i,j} X_{j,\ell} X_{i,\ell},$$

and $\mathtt{E}[Z_A] = \Theta(np^2) = o(1)$. If A has two elements, then Z_A is either 0 or $t_{i,j}$ for some (i,j). Finally, if A has three elements, then Z_A is either 0 or 1. Thus, $E_1(Z) = \max_{|A| \geq 1} \mathtt{E}[Z_A] = 1$, and $E_0(Z) = \max_{|A| \geq 0} \mathtt{E}[Z_A] = \mathtt{E}[Z]$.

Setting $\lambda := cn^{1/8}$ for a constant c chosen to make $a_3 \lambda^3 \sqrt{\mathtt{E}[Z]} = \epsilon \mathtt{E}[Z]$, and applying Theorem 9.2 gives

$$\mathtt{Pr}\left[|Z - \mathtt{E}[Z]| \geq \epsilon \mathtt{E}[Z]\right] \leq b_3 \exp\left(-\lambda/4 + 2\log n\right) = e^{-\Theta(n^{1/8})}.$$

Stronger estimates can be obtained via refinements of this technique [68, 34], achieveing a factor of $\Theta(n^{-3/8})$ in the exponent.

9.3 The Deletion Method

The setting here is similar to that in the Kim-Vu inequality: Let $\mathcal{H}$ be a family of subsets of a base set U and suppose each set in $\mathcal{H}$ is of size at most k for some $k \leq n$. Let $(X_I, I \in \mathcal{H})$ be a family of non-negative random variables. These do not necessarily have the monomial structure as in Kim-Vu. Rather, only a local–dependence property is postulated: each X_I is independent of $(X_J \mid I \cap J = \emptyset)$. Note that this is true of the monomials in the Kim-Vu inequality. The object of study is the sum $Z := \sum_I X_I$.

Theorem 9.3 (Janson-Rucinski) *Let $\mathcal{H}$ be a family of subsets of $[n]$ of size at most k for smome $k \leq n$ and let $(X_I, I \in \mathcal{H})$ be a family of non-negative random variables such that each X_I is independent of $(X_J \mid I \cap J = \emptyset)$. Let $Z := \sum_I X_I$,* *and* $\mu := \mathtt{E}[Z] = \sum_I \mathtt{E}[X_I]$*. Further, for* $I \subseteq [n]$*, let* $Z_I := \sum_{I \subseteq J} X_J$ *and let* $Z_1^* := \max_u Z_{\{u\}}$*. If* $t > 0$*, then for every real* $r > 0$*,*

$$\begin{aligned} \mathtt{Pr}[Z \geq \mu + t] &\leq (1 + t/\mu)^{-r/2} + \mathtt{Pr}\left[Z_1^* > \frac{t}{2kr}\right] \\ &\leq (1 + t/\mu)^{-r/2} + \sum_u \mathtt{Pr}\left[Z_{\{u\}} > \frac{t}{2kr}\right]. \end{aligned}$$

The proof is surprisingly short and elementary, see [29, 30].

To apply this to the problem of counting the number of triangles in $G(n,p)$ with $p = n^{-3/4}$, consider again, base set $\binom{[n]}{2}$ of all possible edges and the family $\mathcal{H}$ to be the family of 3-element subsets of edges forming a triangle. For $I \in \mathcal{H}$,the variable X_I is the indicator for whether the triangle formed by the three edges in I exists in $G(n,p)$..To apply the Deletion method of Janson-Rucinski, note that the number of traingles containing a given edge (i,j), $Z_{\{i,j\}} = X_{i,j}B$ where $B \sim Bi(n-1, p^2)$ is the number of paths of length 2 between the endpoints i and j. Applying the CH bound to this yields

$$\mathtt{Pr}[Z_{\{i,j\}} > \mu/2r] \leq e^{-\mu/2r},$$

as long as $\mu/2r > 2n^{-1/2}$. Thus,

$$\mathtt{Pr}[Z > 2\mu] \leq e^{-r/9} + n^2 e^{-\mu/2r}.$$

Choosing $r = c\sqrt{\mu}$ gives

$$\mathtt{Pr}[Z > 2\mu] \leq n^2 e^{-cn^{3/8}},$$

which is stronger than the result obtained above using the multivariate polynomial inequality. To see a very revealing and exhaustive comparison of the use of various methods for the study of the "infamous upper tail" of problems such as counts of fixed subgraphs in the random graph, see [29]. We end by quoting [30],

> Neither of these methods seems yet to be fully developed and in a final version, and it is likely that further versions will appear and turn out to be important for applications. It would be most interesting to find formal relations and implications between Kim and Vu's method and our new method, possibly by finding a third approach that encompasses both methods.

9.4 Problems

Problem 9.4 Consider the number X_H of copies of a fixed graph H in the random graph $G(n,p)$ for different ranges of the parameter p. Let $\mu := \mathtt{E}[X_H]$ and apply the Kim-Vu multivariate polynomial bound.

(a) For $H := K_3$ (triangle), show that

$$\mathtt{Pr}[X_{K_3} \geq 2\mu] \leq n^4 \begin{cases} \exp\{-cn^{1/3}p^{1/6}\} & \text{if } p \geq n^{-1/2} \\ \exp\{-cn^{1/2}p^{1/2}\} & \text{otherwise.} \end{cases}$$

(b) For $H := K_4$, show that

$$\mathtt{Pr}[X_{K_4} \geq 2\mu] \leq n^{10} \begin{cases} \exp\{-cn^{1/6}p^{1/12}\} & \text{if } p \geq n^{-2/5} \\ \exp\{-cn^{1/3}p^{1/2}\} & \text{otherwise.} \end{cases}$$

(c) For $H := C_4$ (the cycle on 4 vertices), show that

$$\mathtt{Pr}[X_{C_4} \geq 2\mu] \leq n^6 \begin{cases} \exp\{-cn^{1/4}p^{1/8}\} & \text{if } p \geq n^{-2/3} \\ \exp\{-cn^{1/2}p^{1/2}\} & \text{otherwise.} \end{cases}$$

▽

Problem 9.5 Consider the number X_H of copies of a fixed graph H in the random graph $G(n,p)$ for different ranges of the parameter p. Let $\mu := \mathtt{E}[X_H]$ and apply the Janson-Rucinski Deletion method.

(a) For $H := K_3$ (triangle), show that

$$\mathtt{Pr}[X_{K_3} \geq 2\mu] \leq n^2 \exp\{-cn^{3/2}p^{3/2}\}.$$

(b) For $H := K_4$, show that

$$\mathtt{Pr}[X_{K_4} \geq 2\mu] \leq n^2 \begin{cases} \exp\{-cn^2p^3\} & \text{if } p \leq n^{-1/2} \\ \exp\{-cn^{4/3}p^{5/3}\} & \text{otherwise.} \end{cases}$$

(c) For $H := C_4$ (the cycle on 4 vertices), show that

$$\mathtt{Pr}[X_{C_4} \geq 2\mu] \leq n^2 \begin{cases} \exp\{-cn^{4/3}p\} & \text{if } p \geq n^{-2/3} \\ \exp\{-cn^2p^2\} & \text{otherwise.} \end{cases}$$

▽

Chapter 10

Isoperimetric Inequalities and Concentration

10.1 Isoperimetric inequalities

Everyone has heard about the mother of all isoperimetric inequalities:

$$\textit{Of all planar geometric figures with a given perimeter, the circle has the largest possible area.} \tag{10.1}$$

An abstract form of isoperimetric inequalities is usually formulated in the setting of a space (Ω, P, d) that is simultaneously equipped with a probability measure P and a metric d. We will call such a space a MM-space. Since our applications usually involve finite sets Ω and discrete distributions on them, we will not specify any more conditions (as would usually be done in a mathematics book).

Given $A \subseteq \Omega$, the *t-neighbourhood* of A is the subset $A_t \subseteq \Omega$ defined by

$$A_t := \{x \in \Omega \mid d(x, A) \leq t\}. \tag{10.2}$$

Here, by definition,

$$d(x, A) := \min_{y \in A} d(x, y).$$

An abstract isoperimetric inequality in such a MM-space (Ω, P, d) asserts that

$$\text{There is a “special” family of subsets } \mathcal{B} \text{ such that for any } A \subseteq \Omega, \text{ for all } B \in \mathcal{B} \text{ with } P(B) = P(A),\ P(A_t) \leq P(B_t). \tag{10.3}$$

To relate this to (10.1), take the underlying space to be the Euclidean plane with Lebesgue measure and Euclidean distance, and the family $\mathcal{B}$ to be balls in the plane. By letting $t \to 0$, an abstract isoperimetric inequality yields (10.1).

Often an abstract isoperimetric inequality is stated in the following form:

Assertion 10.1 *In a space (Ω, P, d), for any $A \subseteq \Omega$,*

$$P(A)P(\overline{A_t}) \leq g(t) \tag{10.4}$$

Such a result is often proved in two steps:

1. Prove an abstract isoperimetric inequality in the form (10.3) for s suitable family $\mathcal{B}$.
2. Explicitly compute $P(B)$ for $B \in \mathcal{B}$ to determine g.

(In § 10.4, there is an exception to this rule: the function g there is bounded from above directly.)

10.2 Isoperimetry and Concentration

An isoperimetric inequality such as (10.4) implies measure concentration if the function g decays sufficiently fast to zero as $t \to \infty$. Thus, if $A \subseteq \Omega$ satisfies $\Pr(A) \geq 1/2$, then (10.4) implies $\Pr(A_t) \geq 1 - 2g(t)$. If g goes sufficiently fast to 0, then $\Pr(A_t) \to 1$. Thus

> *"Almost all the meausre is concentrated around any subset of measure at least a half"*!

10.2.1 Concentration of Lipschitz functions

It also yields concentration of Lipschitz functions on a space (Ω, d, P). Let f be a Lipschitz function on Ω with constant 1, that is,

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

A *median Lévy Mean* of f is areal number $M[f]$ such that

$$P(f \geq M[f]) \geq 1/2, \quad \text{and} \quad P(f \leq M[f]) \geq 1/2.$$

Exercise 10.2 *Let (Ω, P) be a probability space and let f be a real-valued function on Ω. Define*

$$med(f) := \sup\{t \mid P[f \leq t] \leq 1/2\}.$$

Show that:

$$P[f < med(f)], \quad P[f > med(f)] \quad \leq \quad 1/2.$$

Set

$$A := \{x \in \Omega \mid f(x) \leq M[f]\}.$$

Then, by defintiion of a median, $\Pr(A) \geq 1/2$. Note that since f is Lipschitz,

$$\{x \mid f(x) > M[f] + t\} \subseteq \overline{A_t},$$

and hence,

$$\Pr[f(x) > M[f] + t] \leq \Pr(\overline{A_t}) \leq 2g(t) \rightarrow 0.$$

Exercise 10.3 *Show that (10.4) also implies a similar bound on*

$$\Pr[f(x) > M[f] - t].$$

.

Exercise 10.4 *Show that it suffices to impose a one-sided condition on f:*

$$f(x) \leq f(y) + d(x, y),$$

or

$$f(x) \geq f(y) - d(x, y).$$

to obtain two-sided concentration around a Lévy Mean.

Usually one has a concentration around the expectation. In Problem 10.15 you are asked to check that if the concentration is strong enough, concentration around the expectation or a median are essentially equivalent.

To get a quantitative bound on how good the concentration is, one needs to look at the behaviour of g in (10.4). Let (Ω, P, d) be a MM-space, and let

$$D := \max\{d(x, y) \mid x, y \in \Omega\}.$$

For $0 < \epsilon < 1$, let

$$\alpha(\Omega, \epsilon) := \max\{1 - P(A_{\epsilon D}) \mid P(A) \geq 1/2\}.$$

So a space with small $\alpha(\Omega, \epsilon)$ is one in which there is measure concentration around sets of measure at least 1/2.

A family of spaces $(\Omega_n, d_n, P_n), n \geq 1$ is called

- a *Lévy family* if
$$\lim_{n\to\infty} \alpha(\Omega_n, \epsilon) = 0.$$
- a *concentrated Lévy family* if there are constants $C_1, C_2 > 0$ such that
$$\alpha(\Omega_n, \epsilon) \leq C_1 \exp\left(-C_2 \epsilon \sqrt{n}\right).$$
- a *normal Lévy family* if there are constants $C_1, C_2 > 0$ such that
$$\alpha(\Omega_n, \epsilon) \leq C_1 \exp\left(-C_2 \epsilon^2 n\right).$$

10.3 Examples: Classical and Discrete

10.3.1 Euclidean Space with Lebesgue Measure

Consider Euclidean space R^n with the Eucledean metric and Lebesgue measure μ.

Theorem 10.5 (Isoperimetry for Euclidean Space) *For any compact subset $A \subseteq R^n$, and any $t \geq 0$,*
$$\mu(A_t) \geq \mu(B_t),$$
where B is a ball with $\mu(B) = \mu(A)$.

In Problem 10.16 you are asked to prove this using the famous Brunn-Minkowski inequality.

10.3.2 The Euclidean Sphere

For the sphere S^{n-1} with the usual Eucledean metric inherited from R^n, a r-ball is a sphereical cap i.e. an intersection of S^{n-1} with a half-space.

Theorem 10.6 (Isoperimetry for Euclidean Sphere) *For any measurable $A \subseteq S^{n-1}$, and any $t \geq 0$,*
$$\Pr(A_t) \geq \Pr(C_t),$$
where C is a spherical cap with $\Pr(C) = \Pr(A)$.

A calculation for spherical caps then yields:

Theorem 10.7 (Measure Concentration on the Sphere) *Let $A \subseteq S^{n-1}$ be a meqsurable set with* $\Pr(A) \geq 1/2$. *Then,*

$$P(A_t) \geq 1 - 2e^{-t^2 n/2}.$$

Note that the Sphere S^{n-1} has diameter 2 so this inequality shows that the faimily of spheres $\{S^{n-1} \mid n \geq 1\}$ is a normal Lévy family.

10.3.3 Euclidean Space with Gaussian Measure

Consider R^n with the Eucledean metric and the n-dimensional Gaussian measure γ:

$$\gamma(A) := (2\pi)^{-n/2} \int_A e^{-||x||^2/2} dx.$$

This is a probability distribution on R^n corresponding to the n-dimensional normal distribution. Let $Z_1, \ldots, Z_n$ be i.i.d. variables with the normal distribution $N(0,1)$ i.e. for any real z,

$$\Pr[Z_i \leq z] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2} dt.$$

Then the vector $(Z_1, \cdots, Z_n)$ is distributed according to the measure $.\gamma$. The distribution γ is spherically symmetric: the density function depends only on the distance from the origin.

The isoperimetric inequality for Gaussian measure asserts that among all subsets A with a given $\gamma(A)$, a half space has the smallest possible measure of the t-neighbourhood. By a simple calculation, this yields,

Theorem 10.8 (Gaussian Measure Concentration) *Let $A \subseteq R^n$ be measurable and satisfy $\gamma(A) \geq 1/2$. Then $\gamma(A_t) \geq 1 - e^{-t^2/2}$.*

10.3.4 The Hamming Cube

Consider the *Hamming cube* $Q_n := \{0,1\}^n$ with uniform measure and the *Hamming metric*:

$$d(x,y) := |\{i \in [n] \mid x_i \neq y_i\}.$$

A r-ball in this space is $B^r := \{x \mid d(x,0) \leq r\}$ i.e. the set of all 0/1 sequences that has at most r 1s. Clearly

$$\Pr(B^r) = \frac{1}{2^n} \sum_{0 \leq i \leq r} \binom{n}{i}.$$

Note that the t-neighbourhood of a r-ball is a $r+t$-ball: $B^r_t = B^{r+t}$.

Theorem 10.9 (Harper's Isoperimetric inequality) *If $A \subseteq Q_n$ satisfies $\Pr(A) \geq \Pr(B^r)$, then $\Pr(A_t) \geq \Pr(B^{r+t})$.*

Corollary 10.10 (Measure Concentration for the Hamming Cube) *Let $A \subseteq Q_n$ be such that $\Pr(A) \geq 1/2$. Then $\Pr(A_t) \geq 1 - e^{-2t^2/n}$.*

Since the diameter of Q_n is n, this shows that the family of cubes $\{Q^n \mid n \geq 1\}$ is a normal Lévy family.

Exercise 10.11 *Use the CH bound to deduce Corollary 10.10 from Harper's isoperimetric inequality.*

Exercise 10.12 *Deduce the Chernoff bound for iid variables corresponding to fair coin flips from Corollary 10.10.*

10.4 Martingales and Isoperimetric inequalities

In § 10.2 we saw that an isoperimetric inequality yields the method of bounded differences i.e. concentration for Lipschitz functions. In this section we see that conversely, isoperimetric inequalities can be derived via the method of bounded differences. So, isoperimetric inequalities and the concentration of Lipschitz functions are essentially equivalent.

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,

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

Consider now a weighted verison: the space is $\{0, 1\}^n$ with the uniform measure, but the metric is given by

$$d_\alpha(x, y) := \sum_{x_i \neq y_i} \alpha_i,$$

for fixed non=negative reals $\alpha_i, i \in [n]$.

Exercise 10.13 *Show that*

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

Exercise 10.14 *Check that the result of the previous exercise holds in arbitrary product spaces with arbitrary product distributions and a weighted Hamming metric.*

In the next chapter we will see a powerful extension of this inequality.

10.5 Bibliographic Notes

Ledoux [39][Chapter 1] has a thorough discussion of isoperimetric inequalities and concentration. The vexed issue of concentration around the mean or the median is addressed in Prop. 1.7 and the following discussion there. See also McDiarmid [50]. Examples of isoperimetric inequalities in different spaces are discussed in Ledoux [39][§2.1]. Matousek [46][Chapter 14] has a nice discussion and many examples.

10.6 Problems

Problem 10.15 [Expectation versus Median] In this problem, we check that concentration around the expectation or a median are essentially equivalent.

(a) Let $\Omega_n, P_n, d_n), n \geq 1$ be a normal Lévy family. let Ω_n have diameter D_n. Show that if f is a 1-Lipschitz function on Ω_n, then for some constant $c > 0$,

$$|M[f] - \mathtt{E}[f]| \leq c\frac{D_n}{\sqrt{n}}.$$

(b) Deduce that if $f : S^{n-1} \rightarrow R$ is 1-Lipschitz, then for some constant $c > 0$,

$$|M[f] - \mathtt{E}[f]| \leq c\frac{1}{\sqrt{n}}.$$

(c) Deduce that if $f : Q^n \rightarrow R$ is 1-Lipschitz, then for some constant $c > 0$,

$$|M[f] - \mathtt{E}[f]| \leq c\sqrt{n}.$$

▽

Problem 10.16 [Brunn-Minkowski] Recall the famous **Brunn-Minkowski** inequality: for any non-emty compact subsets $A, B \subseteq R^n$,

$$\text{vol}^{1/n}(A) + \text{vol}^{1/n}(B) \leq \text{vol}^{1/n}(A + B).$$

Deduce the isoperimetric inequality for R^n with Lebesgue measure and Euclidean distance form this. (Hint: Note that $A_t = A + tB$ where B is a ball of unit radius.) ▽

Problem 10.17 [Measure Concentration in Expander Graphs] The **edge expansion** or **conductance** $\Phi(G)$ of a graph $G = (V, E)$ is defined by:

$$\Phi(G) := \min\left\{\frac{e(A, V \setminus A)}{|A|} \mid \emptyset \neq A \subseteq V, |A| \leq |V|/2\right\}.$$

where $e(A, B)$ denotes the number of edges with one endpoint in A and the other in B. Regard G as a MM-space by G with the usual graph distance metric and equipped with the uniform measure P on V. Suppose $\Phi := \Phi(G) > 0$, and that the maximum degree of a vertex in G is Δ. Prove the following measure concentration inequality: if $A \subseteq V$ satisfies $P(A) \geq 1/2$, then $P(A_t) \geq 1 - \frac{1}{2}e^{-t\Phi/\Delta}$. (A constant degree **expander graph** G satisfies $\Phi(G) \geq c_1$ and $\Delta \leq c_2$ for constants $c_1, c_2 > 0$.) ▽

Problem 10.18 [Concentration for Permutations] Apply the average method of bounded differences to establish an isoperimetric inequality for the space of all permutations with the uniform measure and transposition distance. $\bigtriangledown$

Problem 10.19 [Measure Concentration and Length] Schectmann, generalizing Maurey, introduced the notion of **length** in a finite metric space (Ω, d). Say that (Ω, d) has length at most ℓ if there are constants $c_1, \cdots, c_n > 0$ with $\sqrt{\sum_i c_i^2} = \ell$ and a sequence of partitions $P_0 \preceq \cdots \preceq P_n$ of Ω with P_0 trivial, P_n discrete and such that whenever we have sets $A, B \in P_k$ with $A \cup B \subseteq C \in P_{k-1}$, then $|A| = |B|$ and there is a bijection $\phi : A \to B$ with $d(x, \phi(x)) \leq c_k$ for all $x \in A$.

(a) Show that the discrete Hamming Cube Q_n with the Hamming metric has length at most $\sqrt{n}$ by considering the partitions induced by the equivalence relations $x \equiv_k y$ iff $X_i = y_i, i \leq k$ for $0 \leq k \leq n$.

(b) Let $\alpha := (\alpha_1, \cdots, \alpha_n) \geq 0$. Show that the discrete Hamming Cube Q_n with the weighted Hamming metric $d_\alpha(x, y) := \sum_{x_i \neq y_i} \alpha_i$ has length at most $\|\alpha\|_2$.

(c) Show that the group of permutations S_n equipped with the usual transposition metric has small length.

(d) Show that Lipschitz functions on a finite metric space of small length are strongly concentrated around their mean. when the space is equppied with the uniform measure:

Theorem 10.20 *Let (Ω, d) be a finite metric space of length at most ℓ, and let f be a Lipschitz function i.e. $|f(x) - f(y)| \leq d(x, y)$ for all $x, y \in \Omega$. Then, if P is the uniform measure on Ω,*

$$P\left(f \geq E[f] + a\right), P\left(f \leq E[f] - a\right) \leq e^{-a^2/2\ell^2}.$$

(e) Generalize to the case when P is not the uniform distribution by requiring that the map $\phi : A \to B$ above is measure preserving. Show that a similar result holds for the concentration of Lipschitz functions with this condition.

$\bigtriangledown$

Problem 10.21 [Diameter, Laplace Functional and Concentration] Let (Ω, P, d) be a MM-space. The *Laplace functional*, $E = E_{\Omega,P,d}$ is defined by:

$$E(\lambda); = \sup\{\mathtt{E}[e^{\lambda f}] \mid f : \Omega \to R \text{ is 1-Lipschitz and } \mathtt{E}[f] = 0\}.$$

(a) Show that if $E(\lambda) \leq e^{a\lambda^2/2}$ for some $a > 0$, then $\mathtt{Pr}[|f - Ef| > t] \leq e^{-t^2/2a}$. (HINT: recall basic Chernoff bound argument!)

(b) Show that the Laplace functional is sub-additive under products: let $(\Omega_i, P_i, d_i), i = 1, 2$ be two spaces, and let (Ω, P, d) be the product space with $\Omega := \Omega_1 \times \Omega_2$, $P := P_1 \times P_2$ and $d := d_1 + d_2$. Then

$$E_{\Omega,P,d} \leq E_{\Omega_1,P_1,d_1} \cdot E_{\Omega_2,P_2,d_2}.$$

(c) If (Ω, d) has diameter at most 1, show that $E(\lambda) \leq e^{-\lambda^2/2}$. (HINT: First note that by Jensen's inequality, $e^{\mathtt{E}[f]} \leq \mathtt{E}[e^f]$, hence if $\mathtt{E}[f] = 0$, then $\mathtt{E}[e^{-f}] \geq 1$. Now, let f be 1-Lipschitz, and let X and Y be two independent variables distributed according to P. Then,

$$\begin{aligned}
\mathtt{E}[e^{\lambda f(X)}] &\leq \mathtt{E}[e^{\lambda f(X)}]\mathtt{E}[e^{-\lambda f(Y)}] \\
&= \mathtt{E}[e^{\lambda (f(X) - f(Y))}] \\
&= \mathtt{E}\left[\sum_{i \geq 0} \frac{\lambda^i \left(f(X) - f(Y)\right)^i}{i!}\right] \\
&= \sum_{i \geq 0} \mathtt{E}\left[\frac{\lambda^i \left(f(X) - f(Y)\right)^i}{i!}\right]
\end{aligned}$$

Argue that the terms for odd i vanish and bound the terms for even i by using the Lipschitz condition on f.

(d) Deduce the Chernoff-Hoeffding bound from (b) and (c).

▽

Chapter 11

Talagrand's Isoperimetric Inequality

[Talagrand Inequality]

11.1 Statement of 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 $P(A_t) \geq 1 - \alpha(t)$ for some rapidly decreasing function α. (Recall that the neighbourhood set $A_t := \{x \in \Omega \mid d(x, A) \leq t\}$.

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.

Recall the normalized *weighted Hamming distance*, d_α specified by a given set of non-negative reals $\alpha_i, i \in [n]$:

$$d_\alpha(x, y) := \frac{\sum_{x_i \neq y_i} \alpha_i}{\sqrt{\sum_i \alpha_i^2}}. \tag{11.1}$$

Suppose now that each point $x \in \Omega$ is associated with a set of non-negative reals

$\alpha(x)_i, i \in [n]$. Consider the asymmetric "diatance" on Ω given by:

$$d_\alpha(x, y) := \frac{\sum_{x_i \neq y_i} \alpha(x)_i}{\sqrt{\sum_i \alpha(x)_i^2}} \tag{11.2}$$

This is the same as the normalized weighted Hamming distance (11.1), except that it involves a set of non-uniform sweights $\alpha(x)_i, i \in [n]$. As usual, for $A \subseteq \Omega$,

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

Theorem 11.1 (Talagrand's Inequality) *Let A be a subset in a product space with the "distance" (11.2). Then for any $t > 0$,*

$$\mathtt{Pr}[A]\mathtt{Pr}[\overline{A_t}] \leq e^{-t^2/4}. \tag{11.3}$$

Remarks':

1. The inequality will be stated in a seemingly stronger form in a later chapter where it will be proved. However, for all the applications we consider, the form given above suffices and is most conveninet.
2. The inequality should be compared to the statement of the isoperimetric inequality for product spaces and weighhted Hamming distance in a previous chapter. The main difference here is that the "distnace" here is non-uniform and asymmetric.

To gain some intuition about the Talagrand distance, let us set $\alpha(x)_i := 1$ for each $i \in [n]$ and ecah $x \in \Omega$; then we get

$$\begin{aligned} d^{d_\alpha}(x, A) &= \min_{y \in A} \frac{\sum_{x_i \neq y_i} 1}{\sqrt{n}} \\ &= d^H(x, A)/\sqrt{n}, \end{aligned} \tag{11.4}$$

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

$$A^{d_\alpha}_{t/\sqrt{n}} = A^H_t. \tag{11.5}$$

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 (11.5) and Talagrand's inequality (11.3), we get

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

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 11.2 *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.

11.2 Hereditary Functions of of Index Sets

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 [67] and explicit in in Steele [66] and Spencer [65]. We then illustrate the versatality of this framework by several examples.

11.2.1 A General Framework

For $x, y \in \Omega$ and $J \subseteq I$, we use the notation $x_J = y_J$ to mean $x_j = y_j, j \in J$, For $J \subseteq [n]$, let $J_{x=y} := \{j \in J \mid x_j = y_j\}$. Note that $x_{J_{x=y}} = y_{J_{x=y}}$.

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

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

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

Ket f_ϕ be the function determined by a hereditary property of index sets ϕ given by:

$$f_\phi(x) := \max_{\phi(x,J)} |J|. \tag{11.6}$$

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 11.3 *Let f be a hereditary function of index sets. Then for all $t > 0$,*

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

and

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

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

The Theorem follows from a more general result in the next section. Here we illustrate how to use it.

11.2.2 Increasing subsequences

Let $I(x_1, \ldots, x_n)$ denote the length of the largest increasing subsequence from $x_1, \ldots, x_n$. Let $X_1, \ldots, X_n$ be chosen independently at random from $[0, 1]$. Proposition 11.3 can be applied immediately to give a sharp concentration result on $I(X_1, \ldots, X_n)$.

Take the hereditary property $\phi(x, J)$ to be: for $jj' \in J$ such that $j < j'$, we have that $x_j \leq x_{j'}$ i.e. J corresponds to an increasing subsequence in x. Check that I is defined by ϕ, hence:

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

and

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

In this example, since $\mathtt{E}[I]$ is of the order of $\sqrt{n}$, this bound is a dramatic improvement over what could be achieved by the simple method of bounded differences.

11.2.3 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 $\texttt{Pr}[X_k = i]$ is the probability that ball k is thown into bin i. What herediatry function of index sets ϕ can we cook up so that f_ϕ is the the number of non–empty bins? Take $\phi(x, J)$ to hold iff $x(j) \neq x(j')$ for all $j, j' \in J$ with $j \neq j'$ i.e. the balls indexed by J go into distinct bins. A moment's thought shows tha ϕ is a hereditary function of index sets and that f_ϕ is the number of non-empty bins. Applying Theorem 11.3 we get:that if Z is the number of non=empty bins, then

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

and

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

11.2.4 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 (why?).Applying Theorem 11.3 yields bounds comparable with those obtained directly by isoperimetric inequalities in the theory of hereditary sets [8] (see also [66, p. 132].

11.3 Certifiable Functions

In this section, we consider a generalization of the previous section which is somewhat more flexible and powerful. A function f on a product space $\Omega := \prod_{i\in[n]} \Omega_i$ is said to be *lower bound certifiable* or just *certifiable* if:

Lower Bound Certificate (LBC): for each $x \in \Omega$, there is a subset $J(x) \subseteq [n]$ such that

(a) $f(x) \geq \rho |J(x)|$, for some constant $\rho > 0$.

(b) for any $y \in \Omega$ that agrees with x on $J(x)$ (i.e. $x_i = y_i, i \in J(x)$), we have $f(y) \geq f(x)$.

Intuitively, for each $x \in \Omega$, there is a an index set $J(x)$ that acts as a "certificate" for a lower bound of ρ times the cardinality of the certificate $J(x)$ on the value of f at any point that agrees with x on $J(x)$.

Exercise 11.4 *Show that a hereditary function of index sets is certifiable.*

Theorem 11.5 *Let $f : \Omega \to \mathbb{R}$ be certifiable and suppose it is Lipschitz with constant c (i.e. changing any co-ordinate changes the value of f by at most c). Then for all $t > 0$,*

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

and

$$\mathtt{Pr}[f < \mathtt{M}[f] - t] \leq 2 \exp\left(-\frac{\rho}{4c^2} \frac{u^2}{\mathtt{M}[f]}\right).$$

where $\mathtt{M}[f]$ is a median of f and $c^2 := \sum_i c_i^2$.

Proof. For each $x \in \Omega$. let $J(x)$ be the certifying interval for f. Set $\alpha(x)_i := c$ if $i \in J(x)$ and 0 otherwise. Note that

$$\alpha^2 := \sum_i \alpha_i^2(x) = c^2 |J(x)| \leq \frac{c^2}{\rho} f(x), \tag{11.7}$$

where in the final inequality, we use part (a) of the (LBC) condition.

Let $y := \operatorname{argmin}\{d_\alpha(x, z) \mid z \in A\}$. Define y' by setting

$$y'_i := \begin{cases} x_i & \text{if } i \in J(x), \\ y_i & \text{otherwise.} \end{cases} \tag{11.8}$$

Note that $f(y') \geq f(x)$ since $J(x)$ is a lower bound certificate for x.

Now,

$$\begin{aligned}
a &\geq f(y) \quad \text{by definition of } A \\
&\geq f(y') - \sum_{y'_i \neq y_i} c \quad \text{since } f \text{ is Lipschitz with constant } c \\
&= f(y') - \sum_{x_i \neq y_i} c \quad \text{by (11.8) } \textit{nonumber} \qquad (11.9) \\
&\geq f(x) - \sum_{x_i \neq y_i} c \quad \text{since } J(x) \text{ is a lower bound certificate for } x.(11.10)
\end{aligned}$$

Now consider the weighted distance with the normalized weights $\frac{c}{\alpha}$:

$$\begin{aligned}
d_{c/\alpha}(x, A) &= d_{c/\alpha}(x, y) \\
&= \sum_{x_i \neq y_i} \frac{c}{\alpha} \\
&\geq \frac{1}{\alpha}(f(x) - a) \quad \text{using (11.9)} \\
&\geq \frac{\sqrt{\rho}}{c} \frac{f(x) - a}{\sqrt{f(x)}} \text{using (11.7)} \qquad (11.11)
\end{aligned}$$

The function $u \mapsto (u - a)/\sqrt{u}$ is monotone increasing for $u \geq a$, so for any $a \geq 0$,

$$\begin{aligned}
\texttt{Pr}[f(X) \geq a + t] &= \texttt{Pr}\left[\frac{f(X) - a}{\sqrt{a + t}} \geq \frac{u}{\sqrt{a + t}}\right] \\
&\leq \texttt{Pr}\left[\frac{f(X) - a}{\sqrt{f(x)}} \geq \frac{u}{\sqrt{a + t}}\right] \\
&\leq \texttt{Pr}\left[d_{c/\alpha}(x, A) \geq \frac{\sqrt{\rho}}{c} \frac{u}{\sqrt{a + t}}\right] \quad \text{using (11.11)} \\
&\leq \frac{1}{P(A)} \exp\left(-\frac{\rho}{4c^2} \frac{u^2}{a + t}\right)
\end{aligned}$$

In the last step we applied Talagrand's inequality (11.3). That is, remembering the definition of A,

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

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

Exercise 11.6 *Deduce Theorem 11.3 from Theorem 11.5.*

Exercise 11.7 *Rework the examples of increasing subsequences and non-empty bins from the previous subsection using Theorem 11.5.*

11.3.1 Edge Colouring

In this example, we shall give an alternative analysis of a simple randomised algorithms for edge colouring a graph that we analysed in a previous chapter using Martingale methods. For convenience, we recall the problem and algorithm.

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: 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. The function f we consider is the number of edges around u that are *not* coloured succesfully. Clearly f is Lipschitz with all constants 1. Moreover,

Lemma 11.8 *The function f is a certifiable function with constants $\rho = 1/2$ and $c = 1$ function.*

Proof. For each edge e that is unsuccessful, there is at least another edge that gets the same tentative colour – fix one such edge $w(e)$ arbitrarily as a witness to this fact. For a given tentative colouring χ, the index set $J = J(\chi) \subseteq E(u)$ consists of all unsuccessful edges together with their witnesses. The condition (LBC) is now easily verified. First, the function is Lipschitz since changing the tentative colour of any edge changes f by at most 1. Second, the edge set J includes each unsucessful edge e and its witness, so it has at most twice as many edges as unsucessful ones (it is exactly twice if the witness for each unsuccessful edge is distinct from the others). Thus the (LBC) condition is satisfied with $\rho = 1/2$ and $c = 1$. ■

Applying Theorem 11.5, we get the result:

$$\mathtt{Pr}[f > \mathtt{M}[f] + t] \leq 2 \exp\left(-\frac{1}{8}\frac{u^2}{\mathtt{M}[f] + t}\right).$$

and

$$\mathtt{Pr}[f < \mathtt{M}[f] - t] \leq 2 \exp\left(-\frac{1}{8}\frac{u^2}{\mathtt{M}[f]}\right).$$

11.4 The Method of Non–uniformly Bounded Differences

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

Theorem 11.9 (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*

$$f(x) \leq f(y) + \sum_{x_i \neq y_i} \alpha_i(x), \quad \textit{for all } y \in \Omega. \tag{11.12}$$

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

$$\sum_i \alpha_i^2(x) \leq c \tag{11.13}$$

[1] In Steele [66][Lemma 6.2.1], this is stated with some additional superfluous conditions.

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

$$\texttt{Pr}[|f - \texttt{M}[f]| > t] \leq 2e^{-t^2/4c}. \tag{11.14}$$

Proof. Set $A = A(a) := \{y \mid f(y) \leq a\}$, where a is a parameter to be fixed later. By (11.12), 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 (11.13). Hence,

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

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

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

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

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

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

Note that the condition of (11.12) is just like the Lipschitz condition in the method of bounded differences except that the bounding parameters can be non–uniform i.e. a different set of parameters for each x. This is the crucial feature that makes this version substantially more powerful than the usual method of bounded differences as we illustrate with some examples below.

11.4.1 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 (11.12) is satisfied. Moreover $\sum_i \alpha_i^2 \leq n$. Hence,

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

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

11.4.2 Balls and Bins

Consider once again, the example of the number of non-empty bins when m balls are thrown independently and uniformly at random into n bins. For a agiven configuration x of balls in the bins, let $\alpha_i(x) := 1$ if ball i is the lowest numbered ball in its bin and 0 otherwise. Then if f is the number of non-empty bins,

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

Since $\sum_i \alpha_i^2(x) \leq n$, we get the bound:

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

11.4.3 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 11.9, we need to find suitable $\alpha(x)_i, i \in [n]$. That is, we need them to satisfy:

$$TSP(x) \leq TSP(y) + \sum_{x_i \neq y_i} \alpha(x)_i \tag{11.15}$$

Many proofs in the literature [66, 50] use the existence of *space filling curves* to do this. Actually, all one needs is the following simple but surprising fact:

Proposition 11.10 *There is a constant $c > 0$ such that or any set of n points $x_1, \ldots, x_n \in [0,1]^2$, there is a permutation $\sigma : [n] \to [n]$ satisfying $\sum_{i \in [n]} |x_{\sigma(i)} - x_{\sigma(i+1)}|^2 \leq c$, (where the final index $n+1$ is taken modulo n). That is, there is a tour through all points such that the sum of the squares of the lengths of all edges in the tour is at bounded by an absolute constant c.*

In Problem 11.14 we outline a completely elementary proof of this fact.

Let $C(x)$ be this tour corresponding to the points $x_1, \cdots, x_n$. We will use this tour to "stitch" in the points x into the optimal tour for the points $y_1, \cdots, y_n$

and satisfy (11.15). Take $\alpha(x)_i$ to be twice the lengths of the two edges incident to x_i in $C(x)$. Now, we verify that with this choice, (11.15) is satisfied, First, we note that the inequality is trivially true if $x \cap y = \emptyset$ i.e. x and y have no points in common. Otherwise, consider the cycle $C(x)$ and mark the points common to x and y on this tour. Double each edge in $C(x)$. Starting at a point in $x \cap y$ follow $C(x)$ until the last point before hitting another vertex of $x \cap y$. At this point, follow the cycle backwards (using the doubled edges) to the starting point. In this way, all the points in x have been attached by small cycles to a point in $x \cap y$. Let U' be the union of these cycles. Note that the sum of the lengths of the edges in U' is at most $\sum_{x_i \neq y_i} \alpha_i$. Finally consider the graph consisting of vertex set $x \cup y$ and edge set $TSP(y) \cup U'$. By "short circuiting", we can extract a tour of $x \cup y$ of length at most that of the edges in $TSP(y) \cup U'$. Since the length of a tour through x is at most that of a tour through $x \cup y$ and, $TSP(x)$ is an optimal tour through x, this verifies (11.15)

Since $\sum_i \alpha_i^2(x) \leq 4c$ where c is the constant given by Proposition 11.10, applying Theorem 11.9, we arrive at the truly Gaussian tail bound:

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

11.4.4 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 $\frac{B_k}{k} \rightarrow \gamma$ 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 1.17,

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

Thus, applying Theorem 11.9,

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

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

11.5 Bibliographic Notes

Other expositions of Talagrand's isoperimetric inequality are [66, 50, 2]. The original paper is the monumental *tour de force* [67]. Other applications in graph colouring problems can be found in [54]. McDiarmid [51] gives an extension of Talagrand's inequality to permutation distributions that is particularly useful in graph colouring applications. A further extension is given in [42].

11.6 Problems

Problem 11.11 [Independent Sets in Random Graphs] Let G be a graph on the vertex set $[n]$ and let $\alpha(G)$ denote the size of the largest independent set in G.

(a) Show that $\alpha(G)$ is a hereditary property of index sets.

(b) The Erdös-Renyi random graph $G(n, p)$ is the (undirected) graph on vertex set $[n]$ and edge set E defined by picking each possible edge $(i, j) \in [n] \times [n]$ independently with probability p. Deduce a sharp concentration result on $\alpha(G(n, p))$.

▽

Problem 11.12 [VC Dimension] One of the central notions in statistical learning theory is the *Vapnik-Chervonenkis* (VC) dimension. Let $\mathcal{A}$ be a collection of subsets of a base set X and let $x := (x_1, \cdots, x_n) \in X^n$. The *trace* of $\mathcal{A}$ on x is defined by:

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

That is, it is the collection of subsets that can be obtained by intersecting sets in $\mathcal{A}$ with $\{x_1, \cdots, x_n\}$. The number of such subsets, $T(x) := |\mathrm{tr}(x)|$ is called the *shatter coefficient* of $\mathcal{A}$ for x. A subset $\{x_{i_1}, \cdots, x_{i_k}\} \subseteq \{x_1, \cdots, x_n\}$ is said to be *shattered* if $T(x_{i_1}, \cdots, x_{i_k}) = 2^k$. Finally, the *VC dimension* $D(x) = D_{\mathcal{A}}(x)$ is defined to be the largest cardinality of a subset of $\{x_1, \cdots, x_n\}$ shattered by $\mathcal{A}$. Show that the VC dimension is a hereditary function of index sets and hence deduce a sharp concentration result for the VC dimension of a subset of points chosen independently at random. ▽

Problem 11.13 [Self-Bounding Functions] A non-negative function f on a product space $\Omega := \prod_{i \in [n]} \Omega_i$,has the *self-bounding* property if there exist functions $g_i, i \in [n]$ such that for all $x_1, \cdots x_n$ and all $i \in [n]$,

$$0 \leq g(x_1, \cdots, x_n) - g_i(x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_n) \leq 1,$$

and also

$$\sum_i (g(x_1, \cdots, x_n) - g_i(x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_n)) \leq g(x_1, \cdots, x_n).$$

(a) Show that a hereditary function of index sets has the self-bounding property.

(b) Show that a similar concentration result extends to hold for this wider class of functions.

(c) Show that the VC dimension (Problem 11.12) is a self-bounding function

▽

Problem 11.14 [An Amazing Fact] In this problem, we outline an elementary proof due to D.J. Newman of the following amazing fact: for any set of points in the unit square, there is a tour going through all the points such that the sum of the squares of the lengths of the edges in the tour is bounded by an absolute constant!

(a) Show that for any set of points in a right-angled triangle, there is a tour that starts at one endpoint of the hypotenuse, ends at the other endpoint and goes through all the points such that the sum of the lengths of the edges is bounded by the square of the hypotenuse. (Hint: Drop a perpendicular to the hypotenuse from the opposite vertex and use induction.)

(b) Use (a) to deduce the amazing fact with the constant 4.

▽

Problem 11.15 [Steiner Tree] Obtain a Gaussian concentration result for the length of a minimum *Steiner tree* containing a set of n points indepndently and uniformly distributed in the unit square. (A Steiner tree of a set of points is a tree containing the given subset among its vertices i.e. it could contain additional vertices.) (Hint: Use the fact that there is a universal constant bounding the sum of the squares of the lengths of the edges of a minimum spanning tree of any number of points in the unit square.) ▽

Chapter 12

Isoperimetric Inequalities and Concentration via Transportation Cost Inequalities

[Transportation Cost]

In this chapter, we give an introduction to the first of two recent approches to concentration via powerful information-theoretic inequalities: the so called transportation cost inequalities. These inequalities relate two different notions of ”distance” between probability distributions and lead easily to concentration results.

12.1 Distance Between Probability Distributions

Perhaps the best known notion of “distance” between probability distributions is the L_1 or *total variation* distance:

$$d_1(Q, R) := \frac{1}{2} \sum_x |Q(x) - R(x)| \, . \tag{12.1}$$

This is a special case of a more general way of defining a distance between two distributions Q and R on a metric space (Ω, d). the *coupling distance*:

$$d_1(Q, R) := \inf_{\pi(Y,Z)} \mathtt{E}_\pi \left[d(Y, Z) \right], \tag{12.2}$$

where the inf ranges over all couplings π with $\pi(Y) \sim Q$ and $\pi(Z) \sim R$ i.e. joint distributions $\pi(Y, Z)$ with the marginals $\pi(Y) \sim Q$ and $\pi(Z) \sim R$. The intuitive idea is: pick random variables Y and Z according to Q and R respectively and compute the expected distance between them. The added crucial qualification is that Y and Z are not picked independently, but via the best coupling.

Exercise 12.1 (Metric Properties) *Show that this definition defines a bonafide metric on the space of probability distributions on Ω^n.*

In Problem12.13, you are asked to show that when the distance on the space is the Dirac distance, $d(x, y) = 1[x \neq y]$, then this reduces to the total variation distance.

A *transportation cost* (TC) inequality in a MM-space (Ω, P, d) is an inequality of the form:

$$d_1(Q, P) \leq c\sqrt{D(Q||P)}, \quad \text{for any distribution } Q \text{ on } \Omega. \tag{12.3}$$

12.1.1 Distance in Product Spaces with Hamming Metric

Of special interest is a product space. Given MM-spaces $(\Omega_i, P_i, d_i), i \in [n]$, the product space (Ω, P, d) is defined by setting

- $\Omega := \Omega_1 \times \cdots \times \Omega_n$
- $P := P_1 \times \cdots \times P_n$,

and the distance $d = d_H$ is given by the Hamming metric,

$$d_H(x^n, y^n) := \sum_i d_i(x_i, y_i).$$

Recall the coupling distance (12.2) in this setting equals

$$d_1(Q^n, R^n) := \min_{\pi(Y^n, Z^n)} \sum_{i \in [n]} \mathtt{E}_\pi d_i(Y_i, Z_i),$$

where the minimum is over all couplings π of Q^n and R^n i.e. $\pi(Y^n, Z^n)$ is a joint distribution of random variables $Y^n := (Y_1, \cdots, Y_n)$ and $Z^n := (Z_1, \cdots, Z_n)$ with $\pi(Y^n) \sim Q^n$ and $\pi(Z^n) \sim R^n$.

Exercise 12.2 *Check this.*

Exercise 12.3 *Let $\Omega^n := [n]^n$ with the discrete Dirac metric in each component and consider the distributions*

- *The product distribution P^n,*

$$P^n(i_1, \cdots, i_n) := \prod_i \frac{1}{n} = \frac{1}{n^n}.$$

- *The* permutation distribution *Q^n which is concentrated and uniformly distributed on permutations σ of $[n]$:*

$$Q^n(\sigma(1), \cdots, \sigma(n)) = \frac{1}{n!}.$$

Compute $||P^n - Q^n||_1$, $d_1(P^n, Q^n)$, $D(Q^n||P^n)$. (Note that $D(P^n||Q^n)$ is undefined.)

12.2 TC Inequalities Imply Isoperimetric Inequalities and Concentration

A transportation cost inequality in a MM space (Ω, P, d) immediately yields an isoperimetric inequality. First, some notation: for a point $x \in \Omega$ and a subset $A \subseteq \Omega$, define

$$d_1(x, A) := \min_{y \in A} d_1(x, y),$$

and for subsets $A, B \subseteq \Omega$, define

$$\begin{aligned} d_1(A, B) &:= \min_{x \in A} d(x, B) \\ &= \min_{x \in A, y \in B} d(x, y). \end{aligned}$$

Proposition 12.4 (TC Implies Isometry) *Let (Ω, P, d) be a MM-space satisfying the TC inequality (12.3). Then, for $A, B \subseteq \Omega$,*

$$d_1(A, B) \leq c \left(\sqrt{\log \frac{1}{P(A)}} + \sqrt{\log \frac{1}{P(B)}} \right).$$

Proof. Take Q and R to be the measure P conditioned on A and B respectively:

$$Q(x) := \begin{cases} P(x)/P(A) & \text{if } x \in A, \\ 0 & \text{otherwise} \end{cases}$$

and

$$R(x) := \begin{cases} P(x)/P(B) & \text{if } x \in B, \\ 0 & \text{otherwise} \end{cases}$$

Note that

$$\begin{aligned} D(Q||P) &= \sum_{Q(x)>0} Q(x) \log \frac{Q(x)}{P(x)} \\ &= \sum_{x \in A} \frac{P(x)}{P(A)} \log \frac{1}{P(A)} \\ &= \log \frac{1}{P(A)}. \end{aligned} \tag{12.4}$$

Similarly,

$$D(R||P) = \log \frac{1}{P(B)}. \tag{12.5}$$

Then,

$$\begin{aligned} d_1(A,B) &\leq d_1(Q,R), \\ &\quad \text{since the min is at most an average} \\ &\leq d_1(Q,P) + d_1(R,P), \\ &\quad \text{by the triangle inequality} \\ &\leq c\sqrt{D(Q||P)} + c\sqrt{D(R||P)}, \\ &\quad \text{by the Transportation cost Inequality} \\ &= c\left(\sqrt{\log \frac{1}{P(A)}} + \sqrt{\log \frac{1}{P(B)}}\right), \\ &\quad \text{by (12.4) and (12.5)} \end{aligned}$$

■

To obtain the familiar product form of the isoperimetric inequality, take $B := \overline{A_t}$. then,

$$\begin{aligned} t &\leq d(A, \overline{A_t}) \\ &\leq c\left(\sqrt{\log \frac{1}{P(A)}} + \sqrt{\log \frac{1}{P(\overline{A_t})}}\right). \\ &\leq \sqrt{2}c\left(\sqrt{\log \frac{1}{P(A)} + \log \frac{1}{P(\overline{A_t})}}\right), \quad \text{concavity of } \sqrt{\cdot} \\ &= \sqrt{2}c\sqrt{\log \frac{1}{P(A)P(\overline{A_t})}} \end{aligned}$$

Hence,

$$P(A)P(\overline{A_t}) \leq e^{-t^2/2c^2}.$$

As we have seen before, such an insiperimetric inequality implies concentrations of Lipschitz functions. One can also deduce concentration results for Lipschitz functions directly from the transportation cost inequality as oulined in Problem12.14.

12.3 TC Inequality in Product Spaces with Hamming Distance

In this section, we state and prove a TC inequality for product measures with Hamming distance (with the discrete Dirac distance in each coordinate).

Theorem 12.5 (TC Inequality for Product Measures and Hamming Distance) *Let (Ω, P, d) be a product space i.e.for arbitrary MM-spaces $(\Omega_i, P_i, d_i), i \in [n]$,*

- $\Omega := \Omega_1 \times \cdots \times \Omega_n$,
- $P := P_1 \times \cdots \times P_n$, *and*
- $d(x^n, y^n) := \sum_i [x_i \neq y_i]$.

Then for any measure Q on Ω,

$$d_1(Q, P) \leq \sqrt{\frac{n}{2} D(Q||P)}.$$

Exercise 12.6 *Deduce a familiar isoperimetric inequality for product spaces from this TC inequality. (Hint: use Proposition12.2 above.)*

The proof is by induction on the dimension. *All the action takes place in the base case i.e. dimension one!*. The extension to higher dimensions is by abstract nonsense.

12.3.1 One dimension

In one dimension, the basic result is

Theorem 12.7 (Pinsker's inequality)

$$d_1(Q,R) \leq \sqrt{\frac{1}{2}D(Q||R)}.$$

Proof. First we prove the inequality in the special case when $\Omega = \{0,1\}$. Let $q := Q(1)$ and $r := R(1)$, and assume without loss of generaility that $q \geq r$. Then, we need to prove that:

$$q \log \frac{q}{r} + (1-q) \log \frac{1-q}{1-r} \geq 2(q-r)^2. \tag{12.6}$$

This is an exercise in elementary calculus.

For the general case, let $A^* := \{x \in \Omega \mid Q(x) \geq R(x)$, and define measures Q^* and R^* on $\{0,1\}$ by:

$$Q^*(1) := Q(A^*). \quad R^*(1) := R(A^*).$$

Then,

$$\begin{aligned} D(Q||R) &\geq D(Q^*||R^*), \quad \text{by Jensen's Inequality} \\ &\geq 2\left(Q^*(1) - R^*(1)\right)^2 \\ &= 2d_1^2(Q,R). \end{aligned}$$

∎

Exercise 12.8 *Establish (12.6) by calculus.*

12.3.2 Higher dimensions

The "tensorization" step to higher dimesions is by abstract nonsense. We will do it in an abstract general setting because, besides being natural, it is also useful in this form for other applications (other than the one above for simple product measures).

Recall that given MM-spaces $(\Omega_i, P_i, d_i), i \in [n]$, the product space (Ω, P, d) is defined by setting

- $\Omega := \Omega_1 \times \cdots \times \Omega_n$
- $P := P_1 \times \cdots \times P_n$,

- the distance $d = d_H$ is given by the Hamming metric,

$$d_H(x^n, y^n) := \sum_i d_i(x_i, y_i).$$

and The coupling distance 12.2 in this setting equals:

$$d_1(Q^n, R^n) := \inf_{\pi(Y^n, Z^n)} \sum_{i \in [n]} \mathtt{E}_\pi d_i(Y_i, Z_i), \tag{12.7}$$

where the inf is over all couplings π of Q^n and R^n i.e. $\pi(Y^n, Z^n)$ is a joint distribution of random variables $Y^n := (Y_1, \cdots, Y_n)$ and $Z^n := (Z_1, \cdots, Z_n)$ with $\pi(Y^n) \sim Q^n$ and $\pi(Z^n) \sim R^n$.

Proposition 12.9 (Tensorization of Transportation Cost) *Let $(\Omega_i, P_i, d_i), i \in [n]$ be MM-spaces that each satisfy the transportation cost inequality:*

$$d(Q_i, P_i) \leq c\sqrt{D(Q_i||P_i)}, \quad \text{for any distribution } Q_i \text{ on } \Omega_i.$$

for some constant $c > 0$. Let (Ω, P, d) be the product space as defined above. Then (Ω, P, d) satisfies the transportation cost inequality:

$$d(Q, P) \leq c\sqrt{nD(Q||P)}, \quad \text{for any distribution } Q \text{ on } \Omega.$$

Proof. It suffices to construct a coupling $\pi(Y^n, X^n)$ with $\pi(Y^n) \sim Q$ and $\pi(X^n) \sim P$ such that

$$\mathtt{E}_\pi \left[d(Y^n, X^n)\right] = \sum_i \mathtt{E}_\pi \left[d_i(Y_i, X_i)\right] \leq c\sqrt{nD(Q||P)}.$$

Introduce the notational abbreviations:

$$Q(y^i) := \pi(Y^i = y^i), \quad Q_i(y_i \mid y^{i-1}) := \pi(Y_i = y_i \mid Y^{i-1} = y^{i-1}).$$

Define:

$$\Delta_i(y^{i-1}) := D(Q_i(\cdot \mid y^{i-1})||P_i(\cdot \mid y^{i-1})) = D(Q_i(\cdot \mid y^{i-1})||P_i),$$

where the second equality is because P is a product measure. By the *chain rule for divergence*,

$$D(Q||P) = \sum_{i=1}^n \sum_{y^{i-1} \in \Omega^{i-1}} \Delta_i(y^{i-1}) Q(y^{i-1}).$$

We construct the coupling π inductively. Assume the joint distribution on (Y^{i-1}, X^{i-1}) has already been defined. To extend the distribution, we define the joint distribution of (Y_i, X_i) conditioned on $(Y^{i-1} = y^{i-1}, X^{i-1} = x^{i-1})$ for any y^{i-1}, x^{i-1}. First define the marginals by:

$$\pi(Y_i = z \mid Y^{i-1} = y^{i-1}, X^{i-1} = x^{i-1}) := Q_i(z \mid y^{i-1}),$$

and

$$\pi(X_i = z \mid Y^{i-1} = y^{i-1}, X^{i-1} = x^{i-1}) := P_i(z).$$

That is, noth Y_i and X_i are conditionally independent of X^{i-1} given $Y^{i-1} = y^{i-1}$.

Now, we use the transportation cost inequality satisfied by the component space Ω_i to construct a coupling of (Y_i, X_i) with these marginals so that for all y^{i-1},

$$\mathtt{E}_\pi\left[d_i(Y_i, X_i) \mid Y^{i-1} = y^{i-1}\right] \leq c\sqrt{\Delta_i(y^{i-1})}.$$

Finally we verify that this inductively constructed coupling satisfies the desired inequality:

$$\begin{aligned}
\sum_i \mathtt{E}_\pi\left[d_i(Y_i, X_i)\right] &= \sum_i \sum_{y^{i-1}} \mathtt{E}_\pi\left[d_i(Y_i, X_i) \mid Y^{i-1} = y^{i-1}\right] Q(y^{i-1}) \\
&\leq \sum_i \sum_{y^{i-1}} c\sqrt{\Delta_i(y^{i-1})} Q(y^{i-1}) \\
&= cn \sum_i \sum_{y^{i-1}} \sqrt{\Delta_i(y^{i-1})} \frac{Q(y^{i-1})}{n} \\
&\leq cn \sqrt{\sum_i \sum_{y^{i-1}} \Delta_i(y^{i-1}) \frac{Q(y^{i-1})}{n}}, \quad \text{by concavity of } \sqrt{\cdot} \\
&= c\sqrt{nD(Q||P)}, \quad \text{by the chain rule for divergence.}
\end{aligned}$$

∎

We can now complete the proof of the Transportation Cost Inequality in product spaces with the Hamming distance:

Proof. (of Theorem 12.5) Combine Pinsker's inequality with the abstract tensorization of Proposition 12.9. ∎

12.4 An Extension to Non-Product Measures

In this section, we state a theorem due to K. Marton which extends the TC inequality from independent distributions to certain dependent distributions where one has some handle to control the dependence. This extension is quite useful as shown by the application in Problem12.17.

Theorem 12.10 (TC Inequality with controlled dependence) *Let (Ω, Q, d) be MM-space with*

- $\Omega := \Omega_1 \times \cdots \times \Omega_n$.

- $d(x^n, y^n) := \sum_i d_i(x_i, y_i)$, *for arbitrary metrics d_i on Ω_i for each $i \in [n]$, and*

- *Q a measure on Ω such that for each $k \geq 0$ and each $x^k, \hat{x}^k$ differing only in the last co-ordinate (i.e. $x_i = \hat{x}_i, i < k$ and $x_i \neq \hat{x}_i$), there is a coupling $\pi(Y_k^n, Z_k^n)$ of the distributions $Q(\cdot \mid x^k)$ and $Q(\cdot \mid \hat{x}^k)$ such that*

$$\mathtt{E}_\pi \left[\sum_{i>k} d_i(Y_i, Z_i) \mid x^k, \hat{x}^k \right] \leq u.$$

Then for any other measure R,

$$d(R, Q) \leq (u+1)\sqrt{\frac{n}{2} D(R||Q)}.$$

Exercise 12.11 *Deduce the TC inequality for product measures from Theorem12.10*

12.5 Problems

Problem 12.12 Prove the following alternative characterizations of the total variation distance:

$$\begin{aligned} d_1(Q,R) &= \frac{1}{2}\mathtt{E}_Q\left[\left|1-\frac{R(Y)}{Q(Y)}\right|\right] && (12.8)\\ &= \mathtt{E}_Q\left[\left(1-\frac{R(Y)}{Q(Y)}\right)_+\right] && (12.9)\\ &= \sum_y\left(1-\frac{R(y)}{Q(y)}\right)_+ Q(y)\\ &= \sum_y\left(1-\frac{Q(y)}{R(y)}\right)_+ R(y)\\ &= \mathtt{E}_R\left[\left(1-\frac{Q(Y)}{R(Y)}\right)_+\right] && (12.10)\\ &= \max_{A\subseteq\Omega}|Q(A)-R(A)| && (12.11)\end{aligned}$$

▽

Problem 12.13 Show that the total variation distance is also given by:

$$d_1(Q,R) = \min_{\pi(Y,Z)} \mathtt{E}_\pi[Y\neq Z], \qquad (12.12)$$

where the minimum ranges over all couplings $\pi(Y,Z)$ of Q and R: $\pi(Y)\sim Q$ and $\pi(Z)\sim R$.

Proof. We start with the characterization (see Problem12.12)

$$d_1(Q,R) = \max_{A\subseteq\Omega}|Q(A)-R(A)|\,.$$

Let $A\subseteq\Omega$ achieve the maximum on the right hand side. Then,

$$\begin{aligned} d_1((Q,R) &= |Q(A)-R(A)|\\ &= |\pi(Y\in A)-\pi(Z\in A)|\\ &\leq \mathtt{E}_\pi\left[Y\neq Z\right].\end{aligned}$$

Equality is attained by the following coupling of Q and R. Let $\theta(x):=\min(Q(x),R(x))$. and let

$$\pi(Y=x,Z=x) := \theta(x),$$

and for $x \neq x'$, let

$$\pi(Y = x, Z = x') := \frac{(Q(x) - \theta(x))(R(x') - \theta(x'))}{1 - \sum_x \theta(x)}.$$

(Note that if the denominator vanishes then $Q = R$.) ■

▽

Problem 12.14 Use the Transportation Cost inequality to directly deduce a measure concentration result for Lipschitz functions. Let (Ω, P, d) be a MM-space satisffying a TC inequality:

$$d_1(Q, P) \leq c\sqrt{D(Q||P)},$$

and let f be a Lipschitz function on Ω. Let

$$A := \{x \in \Omega \mid f(x) > \mathtt{E}_P[f] + t\}.$$

Let Q be the measure P conditioned on A.

(a) Argue that
$$d_1(Q, P) \geq \mathtt{E}_Q[f] - \mathtt{E}_P[f] \geq t.$$

(b) Deduce that
$$P[f > \mathtt{E}_P[f] + t] \leq e^{-2t^2/c^2 n}.$$

(c) Similarly deduce the other tail inequality.

▽

Problem 12.15 [A Weighted Transportation Cost Inequality in Product Spaces] Let $\alpha := (\alpha_1, \cdots, \alpha_n) \geq 0$ and let (Ω, P_i, d_i) be arbitrary MM-spaces. Consider the product space (Ω, P, d_α) with Ω and P as usual, but with the *weighted Hamming metric*:

$$d_\alpha(x^n, y^n) := \sum_i \alpha_i d(x_i, y_i) \tag{12.13}$$

Prove:

Theorem 12.16 (TC Inequality in Product Spaces with Weighted Hamming Distance) *Let (Ω, P, d_α) be a product space with a weighted Hamming metric (12.13). Suppose the component spaces satisfy a transportation cost inequality:*

$$d(Q, P_i) \leq c\sqrt{D(Q||P_i)} \quad \textit{for } i \in [n].$$

Then, for any measure Q on Ω,

$$d(Q,P) \leq c||\alpha||_1\sqrt{D(Q||P)}.$$

In particular, if $||\alpha||_1 = 1$ i.e. α is a (non-negative) vector with unit L_1 norm, then,

$$d(Q,P) \leq c\sqrt{D(Q||P)}.$$

Verify that the unweighted case is a special case of this. ▽

Problem 12.17 [Transportation Cost and Concentration for Permutations] Consider the group of permutations S_n as a MM-space by endowing it with the uniform distribution P and the transposition distance d between permutations. Show that this space satisfies the transportation cost inequality

$$d(Q,P) \leq \sqrt{2nD(Q||P)}.$$

Deduce an isoperimetric inequality and a measure concentration result for Lipschitz functions on permutations. (Hint: Apply Marton's Theorem12.10.) ▽

Problem 12.18 Prove Theorem12.10 and give a weighted analogue. ▽

12.6 Bibliographic Notes

The approach to measure concentration via transportation cost was introduced by Marton [44]. The extension to dependent measures is from Marton [45]. Ledoux [39][Chapter 6] covers the Transportation cost approach in more detail.

Chapter 13

Quadratic Transportation Cost and Talagrand's Inequality

[Transportation Cost and Talagrand's Inequality]

13.1 Introduction

In this chapter, we will prove Talagrand's convex distance inequality via the transportation cost method, an approach pioneered by Kati Marton [45] and further developed by Amir Dembo [14]. This approach is particularly interesting because:

- It places both the theorem and its proof in its natural place within the context of isoperimetric inequalities.
- It places a standard structure on the proof as opposed to the somewhat *ad hoc* and mysterious nature of the original inductive proof of Talagrand..
- It isolates very clearly the essential content of the proof in one dimension, and shows that the extension to higher dimensions is routine.
- It also allows a stronger version of the method of bounded differences that leads to concrete improvements in applications.
- It allows generalization to dependent measures.

13.2 Review and Roadmap

Recall the setup for the isoperimetric inequality for product measures and a weighted Hamming distance: (Ω, P, d_α) where $\Omega := \prod_{i \in [n]} \Omega_i$, $P := \prod_i P_i$ for arbitrary spaces $(\Omega_i, P_i), i \in [n]$ and the weighted Hamming distance is defined by

$$d_\alpha(x, y) := \sum_{i \in [n]} \alpha_i [x_i \neq y_i], \tag{13.1}$$

for a fixed $\alpha := (\alpha_1, \ldots, \alpha_n) \geq 0$ with norm 1 i.e. $\sum_i \alpha_i^2 = 1$.

To prove this via the Transportation cost method, we introduced a distance between probability measures on Ω that reflected (13.1): namely, if Q and R are distributions on Ω, define

$$d_{1,\alpha}(Q, R) := \inf_{\pi(Y,Z)} \sum_{i \in [n]} \alpha_i [Y_i \neq Z_i] \tag{13.2}$$

We then proved the *Transportation cost inequality* for this distance in product spaces: for any other distribution Q on Ω,

$$d_{1,\alpha}(Q, P) \leq \sqrt{\frac{D(Q||P)}{2}}. \tag{13.3}$$

From this information-theoretic inequality, the isoperimetric inequality for product spaces and weighted Hamming distance followed readily: for any two subsets $A, B \subseteq \Omega$,

check constant in exponent

$$P(X \in A) \cdot P(d_{1,\alpha}(X, A) > t) \leq e^{-2t^2} \tag{13.4}$$

In the non-uniform setting, we have, for every point $x \in \Omega$, a non-negative unit norm vector $\alpha(x) := (\alpha_1(x), \ldots, \alpha_n(x))$ i.e. a function $\alpha : x \to \alpha(x)$ with $||\alpha(x)||_2 = 1$, and one defines an asymmetric notion of "distance" by:

$$d_{2,\alpha}(x, y) := \sum_{i \in [n]} \alpha_i(x) [x_i \neq y_i], \tag{13.5}$$

(The reason for the subscript "2" will emerge shortly.)

As usual, for $A \subseteq \Omega$,

$$d_{2,\alpha}(x, A) := \min_{y \in A} d_{2,\alpha}(x, y).$$

The goal is to prove the following isoperimetric inequality which is analogous to (13.4) which was used in the applications in the previous chapter:

Theorem 13.1 *For any $A \subseteq \Omega$,*

$$P(X \in A)P(d_{2,\alpha}(X,A) > t) \leq e^{-t^2/4}.$$

Some thought shows that proving such an inequality is tantamount to proving the inequality for all possible α simultaneously in the following sense. Define, for $x \in \Omega$ and $A \subseteq \Omega$,

$$d_2(x,A) := \sup_{||\alpha||=1} d_{2,\alpha}(x,A). \tag{13.6}$$

This is just the Talagrand convex distance between a point and a subset. Then we will prove,

Theorem 13.2 (Talagrand's Convex Distance Inequality) *For any $A \subseteq \Omega$,*

$$P(X \in A)P(d_2(X,A) > t) \leq e^{-t^2/4}.$$

To prove this via the transportation cost method, we need to introduce a distance between probability measures in Ω that reflects (13.5) and (13.6). For probability measures Q, R on Ω, define:

$$d_2(Q,R) = \inf_{\pi(Y,Z)} \sup_{\mathtt{E}_Q[||\alpha||_2]\leq 1} \mathtt{E}_\pi[\sum_{i\in[n]} \alpha(Y_i)[Y_i \neq Z_i] \tag{13.7}$$

(The sup is over all functions $\alpha : \Omega \to R^n$ such that $\mathtt{E}_Q[||\alpha(X)||] \leq 1$.) In Problem 13.17 you are asked to show that this notion of "distance" satisfies a traiangle inequality. We will show that this "distance" satisfies a transportation cost inequality and as a consequence yields Talagrand's convex distance inequality.

13.3 A L_2 (Pseudo)Metric on Distributions

13.3.1 One Dimension

A L_2 notion of "distance" between two distributions Q and R on a space is given by the following definition:

$$\begin{aligned} d_2(Q,R) &:= \left(\mathtt{E}_Q\left(1-\frac{R(Y)}{Q(Y)}\right)^2\right)^{1/2} \\ &= \left(\sum_y \left(1-\frac{R(y)}{Q(y)}\right)^2 Q(y)\right)^{1/2} \qquad (13.8) \\ &= \left(\sum_y \frac{R^2(y)}{Q(y)} - 1\right)^{1/2} \qquad (13.9) \end{aligned}$$

Note that this definition is *asymmetric*!

Compare this with the variational distance $d_1(Q,R)$:

$$\begin{aligned} d_1(Q,R) &:= \frac{1}{2}\mathtt{E}_Q\left[\left|1-\frac{R(Y)}{Q(Y)}\right|\right] \\ &= \mathtt{E}_Q\left[\left(1-\frac{R(Y)}{Q(Y)}\right)_+\right] \\ &= \sum_y \left(1-\frac{R(y)}{Q(y)}\right)_+ Q(y) \end{aligned}$$

An alternate characterization of d_2 is via couplings:

Proposition 13.3

$$\begin{aligned} d_2(Q,R) &= \inf_{\pi(Y,Z)} \sup_{\mathtt{E}_Q[\alpha]\le 1} \mathtt{E}_\pi\left[\alpha(Y)[Y\neq Z]\right]. \qquad (13.10) \\ &= \inf_{\pi(Y,Z)} \sum_y \left(\pi(Z\neq y \mid Y=y)\right)^2 Q(Y=y) \qquad (13.11) \end{aligned}$$

Here,

- *The* inf *is over all joint distributions π with marginals Q and R, and*
- *the* sup *is over all $\alpha : \Omega \to \mathbb{R}$.*

Proof. We will show that for *any* joint distribution π,

$$\sup_{\|\alpha\|\leq 1} \mathtt{E}_\pi[\alpha(Y)[Y \neq Z]] = \sum_y (\pi(Z \neq y \mid Y = y))^2 q(Y = y).$$

To show that the left hand side is at most the right hand side, we use the Cauchy-Schwartz inequality:

$$\begin{aligned} E_\pi[\alpha(Y)[Y \neq Z]] &= \sum_y \alpha(y)\pi(Z \neq y) \mid Y = y]q(Y = y) \\ &\leq \left(\sum_y (\alpha(y))^2 q(Y = y)\right)^{1/2} \left(\sum_y (\pi(Z \neq y \mid Y = y))^2 q(Y = y)\right)^{1/2} \\ &\leq \left(\sum_y (\pi(Z \neq y \mid Y = y))^2 q(Y = y)\right)^{1/2} \end{aligned}$$

∎

Exercise 13.4 *Choose α suitably to prove the other direction.*

13.3.2 Tensorization to Higher Dimensions

For probability measures Q, R on Ω^n, definition 13.7 reduces to:

$$d_2(Q, R) = \inf_{\pi(Y^n, Z^n)} \sup_{\mathtt{E}_q[\|\alpha\|_2]\leq 1} \mathtt{E}_\pi \left[\sum_{i\in[n]} \alpha(Y_i)[Y_i \neq Z_i]\right]$$

(The sup is over all functions $\alpha_i : \Omega_i \to \mathbb{R}$ such that $\mathtt{E}_Q[\|\alpha(X)\|_2] \leq 1$.) In Problem 13.17 you are asked to show that this notion of “distance” satisfies a triangle inequality.

An alternate characterization is:

$$d_2(Q, R) = \inf_{\pi(Y^n, Z^n)} \sum_i \sum_{y^n} (\pi(Z_i \neq y_i \mid Y^n = y^n))^2 Q(Y^n = y^n)$$

13.4 Quadratic Transportation Cost

Theorem 13.5 (Quadratic Transportation Cost Inequality in Product Spaces) *Let (Ω, P) be a product space with $\Omega := \prod_{i\in[n]} \Omega_i$ and $P := \prod_{i\in[n]} P_i$ where (Ω_i, P_i) are arbitrary spaces. Then, for any other measure Q on Ω,*

$$d_2(Q, P) \leq \sqrt{2D(Q\|P)}$$

The proof is by induction on dimension where all the action once again is in dimension one!

13.4.1 Base case: One Dimension

In one-dimension, for the L_1 distance d_1, the standard inequality is *Pinsker's inequality*:

$$d_1(Q, R) \leq \sqrt{\frac{1}{2} D(Q\|R)} \tag{13.12}$$

We need an analogous inequality for d_2. Notice that because the distance d_2 is not symmetric (unlike d_1), we actually need two inequalities. However there is an elegant symmetric version due to P-M Samson [63] from which the two asymmetric inequalities we need follow:

Theorem 13.6 *For any two distributions Q and R,*

$$d_2^2(Q, R) + d_2^2(R, Q) \leq 2D(R\|Q) \tag{13.13}$$

Hence,

$$d_2(Q, R), d_2(R, Q) \quad \leq \quad \sqrt{2D(R||Q)}. \tag{13.14}$$

Exercise 13.7 *Consider two distributions Q and R on the two point space $\Omega := \{0, 1\}$. Compute $d_1(Q, R)$ $d_2(Q, R)$ and $D(Q\|R)$. Verify that*

- $D_1(Q, R), d_2(Q, R) \leq D(Q\|R)$.
- $d_1(Q, R) \leq d_2(Q, R)$.

Exercise 13.8 *Write down the inequality in the case of a two point space and compare with Pinsker's inequality.*

Proof. (Of Theorem 13.6): Consider the function

$$\Psi(u) := u \log u - u + 1,$$

and

$$\Phi(u) := \Psi(u)/u.$$

By elementary calculus, it is easy to check that or $0 \leq u \leq 1$,

$$\Psi(u) \geq \frac{1}{2}(1-u)^2,$$

whereas for $u \geq 1$,

$$\Phi(u) \geq \frac{1}{2}(1-\frac{1}{u})^2.$$

Since

$$u \log u - u + 1 = \Psi(u)[u \leq 1] + u\Phi(u)[u > 1],$$

we have,

$$u \log u - u + 1 \geq \frac{1}{2}\,(1-u))_+^2 + \frac{u}{2}\left(1 - \frac{1}{u}\right)_+^2.$$

Putting $u := \frac{Q(X)}{R(X)}$ and taking expectations with respect to the measure $R(X)$ gives the lemma. ∎

Might add a few lines beacause this is a bit tricky ...

13.4.2 Tensorization to Higher Dimensions

Once we have the inequality in one dimension, it is routine (but tedious) to extend the inequality to higher dimensions. We phrase the tensorization lemma in a general abstract fashion to emphasise its generality (which is useful in other applications).

Proposition 13.9 (Tensorization of Quadratic Cost) *Let $(\Omega_i, P_i, d_i), i = 1, 2$ be spaces that separately satisfy a quadratic transportation cost inequality: for any measures Q_i on Ω_i,*

$$d_2(Q_i, P_i) \leq \sqrt{2D(Q_i \| P_i)}, \quad i = 1, 2.$$

Let $\Omega := \Omega_1 \times \Omega_2$ be the product space with product measure $P := P_1 \times P_2$ and distance $d(x, y) := d(x_1, y_1) + d(x_2, y_2)$. Then, the measure P also satisfies a quadratic transportation cost inequality: for any measure Q on Ω,

$$d_2(Q, P) \leq \sqrt{2D(Q \| P)}.$$

Proof. Co-ordinate by co-ordinate extension of the coupling, as in the previous chapter. See also Ledoux [39][Theorem 6.9]. pages 130-131. ∎

Now we can complete the proof of the Quadratic Transportation Cost inequality in product spaces:

Proof. (of Theorem 13.5) Induction using Proposition 13.9 with Theorem 13.6 as the base case. ∎

13.5 Talagrand's Inequality via Quadratic Transportation Cost

Exercise 13.10 *Verify that if $d_2(A, B) := \min_{x \in A} d_2(x, B)$ where $d_2(x, B)$ is the Talagrand convex distance and $d_2(Q, R)$ is the distance defined above for any probability distributions Q and R concentrated on A and B respectively, then $d_2(A, B) \leq d_2(Q, R)$,*

Corollary 13.11 (Talagrand's Convex Distance Inequality in Product Spaces)

$$d_2(A, B) \leq \sqrt{2 \log \frac{1}{P(A)}} + \sqrt{2 \log \frac{1}{P(B)}}.$$

Proof. Take $Q(C) := P(C \mid A), R(C) := P(C \mid B)$. Then,

$$\begin{aligned}
d_2(A, B) &\leq d_2(Q, R), \quad \text{since the min at at most an average} \\
&\leq d_2(Q, P) + d_2(R, P) \quad \text{triangle inequality} \\
&\leq \sqrt{2D(Q||P)} + \sqrt{2D(R||P)} \quad \text{TC inequality} \\
&= \sqrt{2 \log \frac{1}{P(A)}} + \sqrt{2 \log \frac{1}{P(B)}}.
\end{aligned}$$

■

To obtain the familiar product form of Talagrand's inequality, take $B := \overline{A_t}$. then,

$$\begin{aligned}
t &\leq d(A, \overline{A_t}) \\
&\leq \sqrt{2 \log \frac{1}{P(A)}} + \sqrt{2 \log \frac{1}{P(\overline{A_t})}}. \\
&\leq 2\sqrt{\log \frac{1}{P(A)} + \log \frac{1}{P(\overline{A_t})}}, \quad \text{concavity of } \sqrt{\cdot} \\
&= 2\sqrt{\log \frac{1}{P(A)P(\overline{A_t})}}
\end{aligned}$$

Hence,

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

13.6 Method of Bounded Differences Revisited

The Quadratic Transportation Cost inequality, Theorem13.5 can be used to give a direct proof of a somewhat stronger version of the method of bounded differences.

Theorem 13.12 (Method of Average Non-Uniform Bounded Differences) *Let Q be a measure in a product space $\Omega = \prod_{i\in[n]} \Omega_i$ satisfying a quadratic transportation cost inequality: there is a constant $c_1 > 0$ such that for any other measure R,*

$$d_2(Q,R) \le c_1 \sqrt{D(R\|Q)}.$$

Let f be a function such that there is a function $\beta : \Omega \to R^n$ with

$$\mathtt{E}_Q[\sum_i \beta_i^2(X)] \le c_2^2,$$

and such that

$$f(x^{(n)}) \le f(y^{(n)}) + \sum_{i\in[n]} \beta_i(x) d_i(x_i, y_i),$$

for any $x^{(n)}, y^{(n)} \in \Omega$. Then

$$\Pr[f < Ef - t] \le \exp\left(\frac{-t^2}{c_1^2 \cdot c_2^2}\right).$$

Proof. Set

$$A := \{x^{(n)} \in \Omega \mid f(x^{(n)}) < Ef - t\}.$$

Consider the measure R on Ω concentrated on A and defined by $R(x) := Q(x \mid A) = Q(x)/Q(A)$ for $x \in A$ and 0 otherwise. Consider

$$d_2(Q,R) = \inf_{\pi(X^n,Y^n)} \sup_{\mathtt{E}_Q[||\alpha||_2]\le 1} \mathtt{E}_\pi[\sum_{i\in[n]} \alpha(X_i) d(X_i, Y_i)]$$

where $\pi(X^n) \sim Q$ and $\pi(Y^n) \sim R$. Let π be the coupling attaining the infimum. Then

$$\begin{aligned}
d_2(Q,R) &= \sup_{\mathtt{E}_Q[||\alpha||_2]\le 1} \mathtt{E}_\pi[\sum_{i\in[n]} \alpha(X_i) d(X_i, Y_i)] \\
&\ge \mathtt{E}_\pi[\sum_{i\in[n]} \frac{\beta(X_i)}{c_2} d(X_i, Y_i)] \\
&\ge \frac{1}{c_2} \mathtt{E}_\pi(f(X^{(n)}) - f(Y^{(n)})) \\
&= \frac{1}{c_2} \mathtt{E}_Q[f(X^{(n)})] - \mathtt{E}_R[f(Y^{(n)})] \\
&\ge \frac{1}{c_2} t
\end{aligned}$$

But, by hypothesis,

$$d_2(Q,R) \leq c_1\sqrt{D(R||Q)}. = c_1\sqrt{\log\frac{1}{Q(A)}}.$$

Hence,

$$\Pr[f < Ef - t] = Q(A) \leq \exp\left(\frac{-t^2}{c_1^2 \cdot c_2^2}\right).$$

∎

Exercise 13.13 *Show that if we assume*

$$f(x^{(n)}) \geq f(y^{(n)}) - \sum_{i\in[n]} \beta_i(x)d_i(x_i,y_i),$$

then one obtains a similar concentration on $\Pr[f > Ef + t]$*.*

check!
Compare
Kim-Vu.

Example 13.14 [Subgraph Counts] Consider the random graph $G(n,p)$ with vertex set $[n]$ and where ecah possible edge $\{i,j\}$ is present with probbaility p independently. Let H be a fixed graph and let Y_H denote the number of copies of H in $G(n,p)$. The study of Y_H is a clasical topic in the theory of random graphs with a vast literature. We are interested concentration results obtained by estimating the probbaility $P[Y_H > (1+\epsilon)\mathtt{E}[Y_H]]$ for a fixed small constant $\epsilon > 0$.

Consider for illustration the case $H := K_3$. Clearly $\mathtt{E}[Y_{K_3}] = \binom{n}{3}p^3 = \Theta(p^3n^3)$. Vu obtained the first exponential bound:

$$P[Y_{K_3} > (1+\epsilon)\mathtt{E}[Y_{K_3}]] \leq \exp(-\Theta(p^{3/2}n^{3/2})).$$

Subsequently, Kim nad Vu by using a "Divide and Conquer" martingale argument improved this to the near optimal

$$P[Y_{K_3} > (1+\epsilon)\mathtt{E}[Y_{K_3}]] \leq \exp(-\Theta(p^2n^2)).$$

We show how to obtain this easily from the average version of the method of bounded differences above. The underlying product space is given by the indicator random variables $X := X_e, e \in E := \binom{[n]}{2}$ corresponding to the presence of edge e in $G(n,p)$ and the function $f(X_e, e \in E)$ is the number of triangles in the graph formed by the edges $X_e = 1$. Take $\beta_e(x)$ to be the number of triangles containing the edge e in the graph fromed by the edges $x_e = 1$.. Clearly,

$$f(x) \geq f(y) - \sum_{x_e \neq y_e} \beta_e(x).$$

The random variable $\beta_e(X)$ has distribution $Bin(n-2, p^2)$ and hence

$$\mathtt{E}[\beta_e^2(X)] = (n-2)p^2(1-p^2) + (n-2)^2p^4 = \Theta(n^2p^4),$$

and so

$$\sum_e \mathtt{E}[\beta_e^2(X)] = \Theta(n^4p^4).$$

Substituting inot he bound of Theorem 13.12 gives

$$P[Y_{K_3} > (1+\epsilon)\mathtt{E}[Y_{K_3}]] \leq \exp(-\Theta(p^2n^2)).$$

$\bigtriangledown$

13.7 Extension to Dependent Processes

In this section, we state an exetnsion of the Quadratic Transportation Cost inequality for certain classes of dependent measures. The result is due independently to Kati Marton and P-M. Samson . In the formulation below, we follow Samson [63].

Let Q be a a measure on Ω and let $X_1, \ldots X_n$ be distributed according to Q. To quantify the amount of dependence between these variables, introduce an upper triangular matrix $\Gamma = \Gamma(Q)$ with ones on the diagonal.

For $1 \leq i < j \leq n$, denote the vector $(X_i, \ldots, X_j)$ by X_i^j. For every $1 \leq i \leq n$, every $x_1, \ldots x_{i-1}$ with $x_k \in \Omega_k$ and $x_k, x_k' \in \Omega_k$, set:

$$a_j(x_1^{i-1}, x,, x_i') := d_1\left(Q(\cdot \mid X_1^{i-1} = x_i^{i-1}, X_i = x_i), Q(\cdot \mid X_1^{i-1} = x_i^{i-1}, X_i = x_i')\right).$$

That is, take the total variation distance between the two conditional distributions of Q where the two conditionings differ only at one point. Set

$$\Gamma_{i,j}^2 := \sup_{x_i, x_i'} \sup_{x_1, \ldots, x_{i-1}} a_j(x_1^{i-1}, x,, x_i').$$

Theorem 13.15 (Transportation Cost Inequality for Dependent Measures) *For any probability measure R on Ω,*

$$d_2(R,Q),\ d_2(Q,R) \ \leq\ \|\Gamma(Q)\|\ \sqrt{2D(R\|Q)}. \tag{13.15}$$

Exercise 13.16 *Recover the inequality for independent measures from this one.*

13.8 Bibliographic Notes

The transportation cost approach to proving Talagrand's inequality was pioneered by Kati Marton. Dembo [14] contains systematic generalizations to several other geometric inequalities. The proof of the inequality in one dimension and the extension to dependent measures are from Samson [63]. Ledoux [39][§ 6.3] contains a complete exposition.

13.9 Problems

Problem 13.17 Show that the asymmetric and non-uniform notion of distance in (13.5) satsifies a triangle inequality. $\triangledown$

Chapter 14

Log-Sobolev Inequalities and Concentration

[Log-Sobolev Inequalities]

14.1 Introduction

In this chapter, we give an introduction to Log-Sobolev inequalities and their use in deriving concentration of measure results. This is a third importnat methodology for concentration of measure (the other two being martingales and transportation cost) and it appears to be the most powerful of the three.

Given a probability space (Ω, μ), and a function $f : \Omega \to R$, define the *entropy* of f by

$$\mathtt{Ent}_\mu(f) := \mathtt{E}_\mu[f \log f] - \mathtt{E}_\mu[f] \log \mathtt{E}_\mu[f]. \tag{14.1}$$

By Jensen's inequality applied to the convex function $\psi(x) := x \log x$, $\mathtt{Ent}_\mu(f) \geq 0$ for any f.

A *logarithmic Sobolev inequality* or just log-Sobolev inequality bounds $\mathtt{Ent}_\mu[f]$, for a "smooth" function f, by an expression involving its gradient. In R^n which is the original context in which log-Sobolev inequalities were introduced, a measure μ satisfies a log-Sobolev inequality if, for some $C > 0$ and all smooth enough functions f,

$$\mathtt{Ent}_{mu}(f) \leq 2C\mathtt{E}_\mu[|\nabla f|^2]. \tag{14.2}$$

14.2 A Discrete Log-Sobolev Inequality on the Hamming Cube

We are interested here in discrete settings: what is the analogue of ∇f is a discrete setting in order to formulate a version of (14.2)?

Consider the familiar Hamming cube $\{0,1\}^n$. Here a natural analogue of ∇f would be:

$$\nabla f := (D_1 f, \ldots, D_n f),$$

where, for each $i \in [n]$,

$$D_i f(x) := f(x) - f(\sigma_i x),$$

and $\sigma_i(x)$ is the result of flipping the bit in the ith position in x.

Theorem 14.1 (Log-Sobolev Inequality in the Hamming Cube) *For any function $f : \{0,1\}^n \rightarrow R$,*

$$\mathtt{Ent}_\mu(f^2) \leq \frac{1}{2} \sum_{1 \leq i \leq n} \mathtt{E}_\mu[|D_i f|^2]. \tag{14.3}$$

14.3 Concentration: The Herbst Argument

The log-Sobolev inequality (14.3) in Theorem 14.1 yields the familiar measure concentration results for Lipschitz functions on the Hamming cube. Ledoux [39] attributes the basic argument to Herbst.

Let F be 1-Lipschitz (with respect to the Hamming metric in the cube) and apply (14.3) to the function $f^2 := e^{sF}$ for some $s \in R$ to be chosen later.

To bound the right hand side in (14.3), we use the Lipschitz property of F and elementary calculus to get:

$$\begin{aligned} |D_i(e^{sF/2})| &:= |e^{sF(x)/2} - e^{sF(\sigma_k(x))/2}| \\ &\leq |s| e^{sF(x)/2}. \end{aligned}$$

Putting this into (14.3),

$$\mathtt{Ent}_\mu(e^{sF}) \leq \frac{ns^2}{2} \mathtt{E}_\mu[e^{sF}]. \tag{14.4}$$

Now, we introduce some *generatingfunctionology*: let

$$G(s) := \mathtt{E}_{\mu}[e^{sF}].$$

be the (exponential moment) generating function of F. Then, the left hand side is (with $\mathtt{E} = \mathtt{E}_{\mu}$),

$$s\mathtt{E}[Fe^{sF}] - \mathtt{E}_{[}e^{sF}] \log \mathtt{E}_{[}e^{sF}] = sG'(s) - G(s) \log G(s),$$

and the right hand side is

$$\frac{ns^2}{2}G(s).$$

Hence we arrive at the following differential inequality for $G(s)$:

$$sG'(s) - G(s) \log G(s) \leq \frac{ns^2}{2}G(s). \tag{14.5}$$

Let $\Psi(s) := \frac{\log G(s)}{s}$; then from (14.5), we get:

$$\begin{aligned} \Psi'(s) &\leq \frac{ns}{2} \\ &\leq \frac{n}{2}, \quad \text{since } s \leq 1. \end{aligned}$$

Thus

$$\Psi(s) \leq \frac{ns}{2} + a,$$

for some constant a. The constant is determined by noting that

$$\lim_{s \to 0} \Psi(s) = \lim_{s \to 0} \frac{G'(0)}{G(0)} = \mathtt{E}[f].$$

Hence,

$$\Psi(s) \leq \mathtt{E}[f] + \frac{ns}{2},$$

i.e.

$$\mathtt{E}[e^{sF}] =: G(s) \leq \exp\left(sE[F] + \frac{ns^2}{2}\right). \tag{14.6}$$

Thus we have arrived at a bound on the moment generating function of F and this yields as usual, via Markov's inequality applied to e^{sF} , the concentration bound:

$$\mu\left(F > \mathtt{E}[F] + t\right) \leq \exp\left(\frac{-t^2}{2n}\right).$$

14.4 Tensorization

The following theorem enables one to reduce the proof of a log-Sobolev inequality in product spaces to a single dimension:

Theorem 14.2 (Tensorization of Entropy) *Let $X_1, \ldots, X_n$ be independent random variables with (X_i taking values in $(\Omega_i, \mu_i), i \in [n]$). Let f be a non-negative function on $\prod_i \Omega_i$. Then, with $\mu := \prod_i \mu_i$ and $\mu_{-i} := \prod_{j \neq i} \mu_j$,*

$$\mathtt{Ent}_\mu(f) \leq \sum_{i \in [n]} \mathtt{E}_{\mu_{-i}}[\mathtt{Ent}_{\mu_i}[f \mid X_j, j \neq i]]. \tag{14.7}$$

As a first application of Theorem 14.2, we prove Theorem 14.1:

Proof. (of Theorem 14.1): By Theorem 14.2, it suffices to prove the inequality in one dimension, where it amounts to:

$$u^2 \log u^2 + v^2 \log v^2 - (u^2 + v^2) \log \frac{u^2 + v^2}{2} \leq (u - v)^2, \tag{14.8}$$

for any real u, v. This is easily checked by elementary calculus. Thus,

$$\begin{aligned}
\mathtt{Ent}_\mu(f^2) &\leq \sum_{i \in [n]} \mathtt{E}_{\mu_{-i}}[\mathtt{E}_{\mu_i}[f^2 \mid X_j, j \neq i]] \\
&\leq \sum_{i \in [n]} \mathtt{E}_{\mu_{-i}}[\frac{1}{2}\mathtt{E}_{\mu_i}[D_i f^2 \mid X_j, j \neq i]] \\
&= \frac{1}{2} \sum_{1 \leq i \leq n} \mathtt{E}_\mu[|D_i f|^2].
\end{aligned}$$

∎

Exercise 14.3 *Verify (14.8).*

Theorem 14.2 itself follows faily easily from an basic inequality in information theory.

Theorem 14.4 (Han's Inequality for Entropy) *Let $X_1, \ldots, X_n$ be any set of (discrete) random variables, with X_i taking values in Ω_i for $i \in [n]$ and let Q be their distribution on the product space $\Omega := \prod_{i \in [n]} \Omega$. Then,*

$$H_Q(X_1, \ldots, X_n) \leq \frac{1}{n-1} \sum_{i \in [n]} H_Q(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n).$$

Given a distributions Q on a product space $\Omega := \prod_i \Omega_i$, let Q_{-i} denote the distribution on the product space $\Omega_{-i} := \prod_{j \neq i} \Omega_j$ and given by:

$$Q_{=i}(x_{-i}) := \sum_{x_i \in \Omega_i} Q(x),$$

where $x := (x_1, \ldots, x_n)$ and $x_{-i} := (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$.

Theorem 14.5 (Han's Inequality for Relative Entropy) *Let P be the product measure on Ω and let Q be any other measure on Ω. Then,*

$$D(Q||P) \geq \frac{1}{n-1} \sum_{i \in [n]} D(Q_{-i}||P_{-i}),$$

or,

$$D(Q||P) \leq \sum_{i \in [n]} \left(D(Q||P) - D(Q_{-i}||P_{-i}) \right).$$

Proof. (Of Theorem 14.2): First note that if the inequality is true for a random variable f, it is also true for cf for any constant $c > 0$, so we may rescale to assume $E[f] = 1$. Define

$$Q(x) := f(x)\mu(x),$$

so that

$$D(Q\|\mu) = \mathtt{Ent}_\mu[f]$$

Thus,

$$\begin{aligned} \mathtt{Ent}_\mu[f] &= D(Q\|\mu) \\ &\leq \sum_{i \in [n]} \left(D(Q\|\mu) - D(Q_{-i}\|\mu_{-i}) \right) \\ &= \sum_{i \in [n]} \mathtt{E}_{\mu_{-i}}[\mathtt{Ent}_{\mu_i}[f \mid X_j, j \neq i]] \end{aligned}$$

∎

14.5 Modified Log-Sobolev Inequalities in Product Spaces

Let $X_1, \ldots, X_n$ be independent random variables and let $X'_i, \ldots, X'_n$ be an independent identical copy of the variables $X_1, \ldots, X_n$. Let $Z := f(X_1, \ldots, X_n$ be a positive valued random variable and for each $i \in [n]$, set

$$Z'_i := f(X_1, \ldots, X_{i-1}.X'_i, X_{i+1}, \ldots, X_n).$$

Theorem 14.6 (Symmetric Log-Sobolev Inequality in Product Spaces) *Let $X_i, X'_i, i \in [n]$ and $Z, Z'_i, i \in [n]$ be as above. Then,*

$$\mathtt{Ent}[e^{sZ}] \leq \sum_{i \in [n]} \mathtt{E}[e^{sZ} \psi(-s(Z - Z'_i))], \tag{14.9}$$

where $\psi(x) := e^x - x - 1$. Moreover,

$$\mathtt{Ent}[e^{sZ}] \leq \sum_{i \in [n]} \mathtt{E}\left[e^{sZ} \tau(-s(Z - Z'_i))[Z > Z'_i]\right]. \tag{14.10}$$

and

$$\mathtt{Ent}[e^{sZ}] \leq \sum_{i \in [n]} \mathtt{E}\left[e^{sZ} \tau(-s(Z'_i - Z))[Z < Z'_i]\right].. \tag{14.11}$$

where $\tau(x) := x(e^x - 1)$.

Proof. We use Theorem (14.2) applied to the function e^{sZ} and bound each term in the sum on the right hand side. Lemma 14.7 below implies that if Y' is any positive function of $X_1, \ldots, X_{i-1}, X'_i, X_{i+1}, \ldots, X_N$, then,

$$\mathtt{E}_i[Y \log Y] - E_i[Y] \log E_i[Y] \leq \mathtt{E}_i[Y(\log Y - \log Y') - (Y - Y')].$$

Applying this to $Y := e^{sZ}$ and $Y' := e^{Z'_i}$, we get:

$$\mathtt{E}_i[Y \log Y] - E_i[Y] \log E_i[Y] \leq \mathtt{E}_i\left[e^{sZ} \psi\left(-s(Z - Z'_i)\right)\right]$$

This yields (14.9),

To prove the other two inequalities, write

$$e^{sZ} \psi\left(-s(Z - Z'_i)\right) = e^{sZ} \psi\left(-s(Z - Z'_i)\right)[Z > Z'_i] + e^{sZ} \psi\left(s(Z'_i - Z)\right)[Z < Z'_i].$$

By symmetry, the conditional expectation of the second term may be written

$$\begin{aligned} \mathtt{E}_i\left[e^{sZ} \psi\left(s(Z'_i - Z)\right)[Z < Z'_i]\right] &= \mathtt{E}_i\left[e^{sZ'_i} \psi\left(s(Z - Z'_i)\right)[Z > Z'_i]\right] \\ &= \mathtt{E}_i\left[e^{sZ} e^{-s(Z - Z'_i)} \psi\left(s(Z - Z'_i)\right)[Z > Z'_i]\right]. \end{aligned}$$

Thus,

$$\mathtt{E}_i\left[e^{sZ} \psi\left(-s(Z - Z'_i)\right)\right] = \mathtt{E}_i\left[e^{sZ} \psi\left(-s(Z - Z'_i)\right) + e^{-s(Z - Z'_i)} \psi\left(s(Z - Z'_i)\right)[Z > Z'_i]\right].$$

Now (14.10) follows by noting that $\psi(x) + e^x \psi(-x) = x(e^x - 1) =; \tau(x)$.

The proof of (14.11) is symmetric to that of (14.10). ■

Lemma 14.7 *Let Y be a positive random variable. Then, for any $u > 0$,*

$$\mathtt{E}[Y \log Y] - (\mathtt{E}[Y]) \log(\mathtt{E}[Y]) \leq \mathtt{E}[Y \log Y - Y \log u - (Y - u)].$$

Proof. For any $x > 0$, $\log x \leq x - 1$, hence

$$\log \frac{u}{\mathtt{E}[Y]} \leq \frac{u}{\mathtt{E}[Y]} - 1,$$

and so,

$$\mathtt{E}[Y] \log \frac{u}{\mathtt{E}[Y]} \leq u - \mathtt{E}[Y],$$

which is equivalent to the statement in the lemma. ∎

14.6 The Method of Bounded Differences Revisited

Theorem 14.8 (Method of Bounded Differences) *If*

$$\sum_{i \in [n]} (Z - Z_i')^2 \leq C,$$

for some constant $C > 0$, then

$$\Pr[Z > \mathtt{E}[Z] + t], \Pr[Z < \mathtt{E}[Z] - t] \leq \exp(-t^2/4C).$$

Proof. Observe that for $x < 0$, $\tau(-x) \leq x^2$ and hence for any $s > 0$, we have by (14.10),

$$\begin{aligned} \mathtt{Ent}[e^{sZ}] &\leq \mathtt{E}\left[e^{sZ} \sum_{i \in [n]} s^2 (Z - Z_i')^2 [Z > Z_i']\right]. \\ &\leq \mathtt{E}\left[e^{sZ} \sum_{i \in [n]} s^2 (Z - Z_i')^2\right] \\ &\leq s^2 C \mathtt{E}[e^{sZ}], \end{aligned}$$

where in the last step, we used the hypothesis.

Now we complete the Herbst argument via generatingfunctionology. Introduce the generating function $G(s) : -\mathtt{E}[e^{sZ}]$ and observe that the left hand side is

$$\mathtt{Ent}[e^{sZ}] = sG'(s) - G(s) \log G(s).$$

so,

$$\mathtt{Ent}[e^{sZ}] = sG'(s) - G(s)\log G(s) \leq s^2 CG(s).$$

Divide both sides by $s^2F(s)$ and observe that the LHS is then the derivative of

$$\Psi(s) := \frac{\log G(s)}{s}.$$

Hence, we have

$$\Psi'(s) \leq C,$$

which integrates to

$$\Psi(s) \leq sC + a,$$

for some constant a. The constant is determined by noting that

$$\lim_{s \to 0} \Psi(s) = \lim_{s \to 0} \frac{G'(s)}{G(s)} = \frac{G'(0)}{G(0)} = \mathtt{E}[Z],$$

so

$$\Psi(s) \leq \mathtt{E}[Z] + Cs,$$

which gives a bound on the moment generating function

$$G(s) \leq \exp\left(\mathtt{E}[Z]s + s^2 C\right).$$

This bound yields the desired concentration via the usual argument of applying Markov's inequality to e^{sZ}. ■

Exercise 14.9 *Check that it is sufficient to assume*

$$\sum_{i \in [n]} (Z - Z_i')^2 [Z > Z_i'] \leq C,$$

for the proof above.

14.7 Talagrand's Inequality Revisited

In this section we show how Talagrand's inequality follows easily via log-Sobolev inequalities.

Recall the setting of Talagrand's inequality: we have a product distribution in a product space, and the Talagrand convex distance:between a point x and a subset A in the space:

$$d_T(x, A) := \sup_{\|\alpha|\|=1} d_\alpha(x, A),$$

where

$$\begin{aligned} d_\alpha(x, A) : &= \min_{y \in A} d_\alpha(x, y) \\ &= \min_{y \in A} \sum_{i \in [n]} \alpha_i [x_i \neq y_i] \end{aligned}$$

Eqvivalently, we may write:

$$d_T(x, A) = \inf_{\nu \in D(A)} \sup_{\|\alpha\|=1} \sum_i \alpha_i \mathtt{E}_\nu [x_i \neq Y_i], \tag{14.12}$$

where $D(A)$ is the set of probability distributions concentrated on A.

Exercise 14.10 *Check that (14.12) is equivalent to the usual definition.*

Now we apply Sion's MiniMax Theorem: if $f : X \times Y$ is convex, lower-semicontinuous with respect to the first argument, concave and upper semi-continuous with respect to the second argument, and X is convex and compact, then

$$\inf_x \sup_y f(x, y) = \sup_y \inf_x f(x, y) = \min_x \sup_y f(x, y).$$

Applying this to the characterization (14.12), we have,

$$\begin{aligned} d_T(x, A) &= \inf_{\nu \in D(A)} \sup_{\|\alpha\|=1} \sum_i \alpha_i \mathtt{E}_\nu [x_i \neq Y_i] \\ &= \sup_{\|\alpha\|=1} \inf_{\nu \in D(A)} \sum_i \alpha_i \mathtt{E}_\nu [x_i \neq Y_i] \end{aligned}$$

and the saddle point is achieved by some pair (ν, α).

Let Z denote the random variable $d_T(X, A)$. Given $X = (X_1, \ldots, X_n)$, let $(\hat{\nu}, \hat{\alpha})$ denote the saddle point corresponding to X. Then,

$$\begin{aligned} Z_i' &:= \inf_\nu \sup_\alpha \sum_j \alpha_j \mathtt{E}_\nu \left[X_j^{(i)} \neq Y_j \right] \\ &\geq \inf_\nu \sum_j \hat{\alpha}_j \mathtt{E}_\nu \left[X_j^{(i)} \neq Y_j \right] \end{aligned}$$

where $X_j^{(i)} = X_j$ if $j \neq i$ and $X_i^{(i)} = X_i'$. Let $\tilde{\nu}$ denote the distribution achieving the infimum in the last line. Then

$$\begin{aligned} Z &= \inf_\nu \sum_j \hat{\alpha}_j \mathtt{E}_\nu \left[X_j \neq Y_j \right] \\ &\leq \sum_j \hat{\alpha}_j \mathtt{E}_{\tilde{\nu}} \left[X_j \neq Y_j \right] \end{aligned}$$

Hence,

$$\begin{aligned} Z - Z_i' &\leq \sum_j \hat{\alpha}_j \mathtt{E}_{\tilde{\nu}} \left([X_j \neq Y_j] - [X_j^{(i)} \neq Y_i] \right) \\ &= \hat{\alpha}_i \mathtt{E}_{\tilde{\nu}} \left([X_i \neq Y_j] - [X_i' \neq Y_i] \right) \\ &\leq \hat{\alpha}_i \end{aligned}$$

Hence,

$$\sum_i (Z - Z_i')^2 [Z > Z_i'] \leq \sum_i \hat{\alpha}_i^2 = 1.$$

Now form the observation of the proof in Theorem 14.8 needed in Exercise 14.9, we get the result.

14.8 Problems

Problem 14.11 Consider the Hamming Cube with non-homogeneous product measure.

(a) Derive a log-Sobolev inequality analogous to (14.3).

(b) Use the log-Sobolev inequality to derive a concentration result for Lipschitz functions on the cube.

▽

Problem 14.12 Consider the convex cube $[0,1]^n$ non-homogeneous product measure where the expected value on co-ordinate $i \in [n]$ is p_i.

(a) Derive a log-Sobolev inequality analogous to (14.3). (Hint: use a convexity argument to reduce this to the previous problem.)

(b) Use the log-Sobolev inequality to derive a concentration result for Lipschitz functions on the convex cube.

▽

Problem 14.13 Relax the codition of Theorem (14.8) as follows to get a average version of the method of bounded differences. Show that if

$$\mathtt{E}\left[\sum_i (Z - Z_i')^2 [Z > Z_i'] \mid X_1, \ldots, X_n \right] \leq C,$$

then for all $t > 0$,

$$\Pr[Z > \mathtt{E}[Z] + t] \leq e^{-t^2/4C},$$

while if

$$\mathtt{E}\left[\sum_i (Z - Z_i')^2 [Z < Z_i'] \mid X_1, \ldots, X_n\right] \leq C,$$

then for all $t > 0$,

$$\Pr[Z < \mathtt{E}[Z] - t] \leq e^{-t^2/4C},$$

▽

14.9 Bibliographic Notes

Our exposition is based on a combination of Ledoux [39][§ 5.1, 5.4] and the notes of Gabor Lugosi [43]. A nice survey of the Entropy method in the context of other techniques is in [62]. The original article developing the modified Log-sobolev inequalities with many other variations is [61]. Bobkov and Götze [6] compare the relative strengths of the transportaion cost and log-Sobolev inequalities.

Bibliography

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

[2] N. Alon and J. Spencer. *The Probabilistic Method.* John Wiley, 2 edition, 2000.

[3] Noga Alon, Jeong-Han Kim, and Joel Spencer. Nearly perfect matchings in regular simple hypergraphs. *Israel J. Math.*, 100:171–187, 1997.

[4] M. Bellare and J. Rompel. Randomness-efficient oblivious sampling. Foundations of Computer Science, FOCS. IEEE, 1994.

[5] David Blackwell. An analog of the minimax theorem for vector payoffs. *Pacific J. Math.*, 6:1–8, 1956.

[6] S.G. Bobkov and F. Götze. Exponential integrability and transportation cost related to logarithmic sobolev inequalities. *Journal of Functional Analysis*, 163, 1999.

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

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

[9] Béla Bollobás. Chromatic number, girth and maximal degree. *Discrete Math.*, 24(3):311–314, 1978.

[10] Shiva Chaudhuri and Devdatt Dubhashi. Probabilistic recurrence relations revisited. *Theoret. Comput. Sci.*, 181(1):45–56, 1997. Latin American Theoretical INformatics (Valparaíso, 1995).

[11] 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.

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

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

[14] A. Dembo. Information inequalities and concentration of measure. *Ann. of Prob.*, 25(2):927–939, 1997.

[15] Frank den Hollander. *Large deviations*, volume 14 of *Fields Institute Monographs.* American Mathematical Society, Providence, RI, 2000.

[16] D. P. Dubhashi and D. Ranjan. Balls and Bins: a Study in Negative Dependence. *Random Structures and Algorithms*, 13(2):99–124, 1998.

[17] Dean P. Foster and Rakesh Vohra. Regret in the on-line decision problem. *Games Econom. Behav.*, 29(1-2):7–35, 1999. Learning in games: a symposium in honor of David Blackwell.

[18] Yoav Freund and Robert E. Schapire. Adaptive game playing using multiplicative weights. *Games Econom. Behav.*, 29(1-2):79–103, 1999. Learning in games: a symposium in honor of David Blackwell.

[19] Naveen Garg, Goran Konjevod, and R. Ravi. A polylogarithmic approximation algorithm for the group Steiner tree problem. In *Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms (San Francisco, CA, 1998)*, pages 253–259, New York, 1998. ACM.

[20] David Gillman. A Chernoff Bound for Random Walks on Expander Graphs. *Siam J. Computing*, 27(4):1203–1220, August 1998.

[21] D. Grable. A large deviation inequality for functions of independent, multi-way choices. *Comb. Prob. & Computing*, 7(1):57–63, 1998.

[22] D. Grable and A. Panconesi. Nearly-optimal, distributed edge-colouring in $O(\log\log n)$ rounds. *Random Structures & Algorithms*, 10(3):385–405, 1997.

[23] David A. Grable and Alessandro Panconesi. Fast distributed algorithms for Brooks-Vizing colorings. *J. Algorithms*, 37(1):85–120, 2000. Special issue for the best papers of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms (San Francisco, CA, 1998).

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

[25] O. Häggström. *Finite Markov Chains and Algorithmic Applications*. Cambridge University Press, 2002.

[26] Thomas P. Hayes. Randomly colouring graphs of girth at least five. Symposium on the Theory of Computing (STOC). ACM, 2003.

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

[28] S. Janson. Large deviations for sums of partially dependent random variable. *Random Structures and Algorithms*, pages 234–248, 2004.

[29] Svante Janson and Andrzej Ruciński. The infamous upper tail. *Random Structures Algorithms*, 20(3):317–342, 2002. Probabilistic methods in combinatorial optimization.

[30] Svante Janson and Andrzej Ruciński. The deletion method for upper tail estimates. *Combinatorica*, 24(4):615–640, 2004.

[31] J. Considine J.W. Byers and M. Mitzenmacher. Geometric generalizations of the power of two choices. Symp. on Parallel Algorithms and Architectures (SPAA). ACM, 2004.

[32] Nabil Kahale. Large deviation bounds for markov chains. *Combinatorics, Probability and Computing*, 6:465–474, 1997.

[33] Richard M. Karp. Probabilistic recurrence relations. *J. Assoc. Comput. Mach.*, 41(6):1136–1150, 1994.

[34] J. H. Kim and V. H. Vu. Divide and conquer martingales and the number of triangles in a random graph. *Random Structures Algorithms*, 24(2):166–174, 2004.

[35] Jeong Han Kim. On Brooks' theorem for sparse graphs. *Combin. Probab. Comput.*, 4(2):97–132, 1995.

[36] J. Kleinberg and E. Tardos. *Algorithm Design*. Pearson - Addison Wesley, 2005.

[37] Goran Konjevod, R. Ravi, and Aravind Srinivasan. Approximation algorithms for the covering Steiner problem. *Random Structures Algorithms*, 20(3):465–482, 2002. Probabilistic methods in combinatorial optimization.

[38] Dexter C. Kozen. *The design and analysis of algorithms*. Texts and Monographs in Computer Science. Springer-Verlag, New York, 1992.

[39] Michel Ledoux. *The Concentration of Measure Phenomenon*. American Mathematical Society, 2001.

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

[41] Michael Luby. A simple parallel algorithm for the maximal independent set problem. *SIAM J. Comput.*, 15(4):1036–1053, 1986.

[42] Malwina J. Luczak and Colin McDiarmid. Concentration for locally acting permutations. *Discrete Math.*, 265(1-3):159–171, 2003.

[43] G. Lugosi. Concentration of Measure Inequalities. lecture Notes, 2005.

[44] K. Marton. Bounding $\overline{d}$-distance by Information Divergence: A method to prove Measure Concentration. *Ann. of Prob.*, 24(2):857–866, 1996.

[45] K. Marton. Measure Concentration for a class of Random Processes. *Prob. Theory Relat. Fields*, 110(2):427–439, 1998.

[46] Jiří Matoušěk. *Lectures on Discrete Geometry.* Springer, 2002.

[47] C. J. H. McDiarmid and R. B. Hayward. Large deviations for Quicksort. *J. Algorithms*, 21(3):476–507, 1996.

[48] 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.

[49] Colin McDiarmid. Centering sequences with bounded differences. *Combin. Probab. Comput.*, 6(1):79–86, 1997.

[50] Colin McDiarmid. Concentration. In *Probabilistic methods for algorithmic discrete mathematics*, volume 16 of *Algorithms Combin.*, pages 195–248. Springer, Berlin, 1998.

[51] Colin McDiarmid. Concentration for independent permutations. *Combin. Probab. Comput.*, 11(2):163–178, 2002.

[52] Gary L. Miller and John H. Reif. Parallel tree contraction.and its applications. *Randomness and Computation*, pages 47–72, 1989.

[53] M.Molloy and B.Reed. *Graph Colouring and the Probabilistic Method.* Springer, 2002.

[54] M. Molloy and B Reed. *Graph Colouring and the Probabilistic Method.* Number 23 in Algorithms and Combinatorics. Springer, 2002.

[55] Rajeev Motwani and Prabhakar Raghavan. *Randomized algorithms.* Cambridge University Press, Cambridge, 1995.

[56] Noam Nisan and Avi Wigderson. Hardness vs. randomness. *J. Comput. System Sci.*, 49(2):149–167, 1994.

[57] A. Panconesi and J. Radhakrishnan. Expansion properties of (secure) wireless networks. Symp. on Parallel Algorithms and Architectures (SPAA). ACM, 2004.

[58] 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.

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

[60] Alessandro Mei Alessandro Panconesi Roberto Di Pietro, Luigi Mancini and Jaikumar Radhakrishnan. Connectivity properties of secure sensor networks. Workshop on Security of Ad Hoc and Sensor Networks (SASN '04). ACM, 2004.

[61] G. Lugosi S. Boucheron and P. Massart. Concentration Inequalities using the Entropy Method. *Annals of Probability*, 31, 2003.

[62] O. Bosquet S. Boucheron and G. Lugosi. Concentration inequalities. In O. Bosquet eta al, editor, *Advanced Lectures in Machine Learning*, pages 208–240. Springer, 2004.

[63] P-M. Samson. Concentration of measure inequalities for Markov chains and Φ-mixing processes'. *Ann. of Prob.*, 28(1):416–461, 2000.

[64] Siegel Alan Schmidt, Jeanette P and Aravind Srinivasan. Chernoff-hoeffding bounds for applications with limited independence. *SIAM J. Discrete Math.*, 8(2):255–280, 1995.

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

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

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

[68] V. H. Vu. Concentration of non-Lipschitz functions and applications. *Random Structures Algorithms*, 20(3):262–316, 2002. Probabilistic methods in combinatorial optimization.