The Probabilistic Method

Lecture Notes

Jiří Matoušek Jan Vondrák
Department of Applied Mathematics
Charles University
Malostranské nám. 25
118 00 Praha 1
Czech Republic

If you find errors, please let us know!
(e-mail: `matousek@kam.mff.cuni.cz`**)**

July 2002

Table of Contents

Preface

These are notes to a lecture taught by J. Matoušek at Charles University in Prague for several years. The audience were students of mathematics or computer science, usually with interest in combinatorics and/or theoretical computer science.

Generally speaking, an introductory text on the probabilistic method is rather superfluous, since at least two excellent sources are available: the beautiful thin book

> J. Spencer: *Ten lectures on the probabilistic method*, CBMS-NSF, SIAM, Philadelphia, PA, 1987

and the more modern and more extensive but no less readable

> N. Alon and J. Spencer: *The Probabilistic Method*, J. Wiley and Sons, New York, NY, 2nd edition, 2000.

The lecture was indeed based on these. However, these books were not generally available to students in Prague, and this was the main reason for starting with the present notes. For students, the notes may have another advantage too: they cover the material usually presented in the course relatively concisely. Chapters 8 and 9 go beyond the usual scope of the course and present, mostly without proofs, more recent and more advanced results on strong concentration.

Our presentation is slightly more formal in some cases and includes a brief review of the relevant probability theory notions. This keeps with the Prague mathematical tradition and should be closer to the presentation the students are used to from other math courses. Teaching experience also shows that the students' proficiency in application of the notions learned in probability theory is limited and that it is useful to demonstrate concrete applications of abstract probabilistic notions in some detail.

The techniques are usually illustrated with combinatorial examples. The notation and definitions not introduced here can be found in the book

> J. Matoušek and J. Nešetřil: *Invitation to Discrete Mathematics*, Oxford University Press, Oxford 1998

(Czech version: Kapitoly z diskrétní matematiky, Nakladatelství Karolinum 2000).

A large part of the material is taken directly from the Alon–Spencer book cited above, sometimes with a little different presentation. Readers wishing to pursue the subject in greater depth are certainly advised to consult that book. A more advanced source is

S. Janson, T. Łuczak, A. Ruciński: *Topics in random graphs*, J. Wiley and Sons, New York, NY, 2000.

A very nice book on probabilistic algorithms, also including a chapter on the probabilistic method per se, is

R. Motwani and P. Raghavan: *Randomized Algorithms*, Cambridge University Press, Cambridge, 1995.

Two journals in whose scope the probabilistic method occupies a central place are *Random Structures & Algorithms* and *Combinatorics, Probability & Computing*. Papers with applications of the probabilistic method are abundant and can be found in many other journals too.

A note for Czech students. Teorie pravděpodobnosti, podobně jako jiné matematické disciplíny, má ustálenou základní českou terminologii, která se v mnoha případech neshoduje s doslovným překladem terminologie anglické. Do textu jsme zahrnuli některé české termíny jako poznámky pod čarou, abychom nepodporovali bujení obratů typu "očekávaná hodnota", což je doslovný překlad anglického "expectation", místo správného *střední hodnota*.

1

Preliminaries

1.1 Probability Theory

This section summarizes the fundamental notions of probability theory and some results which we will need in the following chapters. In no way is it intended to serve as a substitute for a course in probability theory.

1.1.1 Definition. *A* **probability space**[1] *is a triple* $(\Omega, \Sigma, \mathrm{P})$, *where* Ω *is a set,* $\Sigma \subseteq 2^{\Omega}$ *is a* σ*-algebra on* Ω *(a collection of subsets containing* Ω *and closed on complements, countable unions and countable intersections), and* P *is a countably additive measure*[2] *on* Σ *with* $\mathrm{P}[\Omega] = 1$. *The elements of* Σ *are called* **events**[3] *and the elements of* Ω *are called* **elementary events**. *For an event* A, $\mathrm{P}[A]$ *is called the* **probability** *of* A.

In this text, we will consider mostly *finite probability spaces* where the set of elementary events Ω is finite and $\Sigma = 2^{\Omega}$. Then the probability measure is determined by its values on elementary events; in other words, by specifying a function $p : \Omega \to [0, 1]$ with $\sum_{\omega \in \Omega} p(\omega) = 1$. Then the probability measure is given by $\mathrm{P}[A] = \sum_{\omega \in A} p(\omega)$.

The basic example of a probability measure is the *uniform distribution*[4] on Ω, where

$$\mathrm{P}[A] = \frac{|A|}{|\Omega|} \quad \text{for all } A \subseteq \Omega.$$

Such a distribution represents the situation where any outcome of an experiment (such as rolling a die)[5] is equally likely.

1.1.2 Definition (Random graphs). [6] *The probability space of random graphs* $G(n,p)$ *is a finite probability space whose elementary events are all graphs on a fixed set of* n *vertices, and where the probability of a graph with* m *edges is*

$$p(G) = p^m (1-p)^{\binom{n}{2} - m}.$$

---

[1] probability space = pravděpodobnostní prostor
[2] measure = míra
[3] event = jev
[4] uniform distribution = rovnoměrné rozdělení
[5] rolling a die = hod kostkou
[6] random graph = náhodný graf

This corresponds to generating the random graph by including every potential edge independently with probability p. For $p = \frac{1}{2}$, we toss a fair coin[7] for each pair $\{u, v\}$ of vertices and connect them by an edge if the outcome is heads.[8] [9]

Here is an elementary fact which is used all the time:

1.1.3 Lemma. *For any collection of events $A_1, \dots, A_n$,*

$$\mathrm{P}\left[\bigcup_{i=1}^{n} A_i\right] \le \sum_{i=1}^{n} \mathrm{P}[A_i].$$

Proof. For $i = 1, \dots, n$, we define

$$B_i = A_i \setminus (A_1 \cup A_2 \cup \ldots \cup A_{i-1}).$$

Then $\bigcup B_i = \bigcup A_i$, $\mathrm{P}[B_i] \le \mathrm{P}[A_i]$, and the events $B_1, \dots, B_n$ are disjoint. By additivity of the probability measure,

$$\mathrm{P}\left[\bigcup_{i=1}^{n} A_i\right] = \mathrm{P}\left[\bigcup_{i=1}^{n} B_i\right] = \sum_{i=1}^{n} \mathrm{P}[B_i] \le \sum_{i=1}^{n} \mathrm{P}[A_i].$$

□

1.1.4 Definition. *Events A, B are* **independent**[10] *if*

$$\mathrm{P}[A \cap B] = \mathrm{P}[A]\,\mathrm{P}[B]\,.$$

More generally, events $A_1, A_2, \dots, A_n$ are **independent** *if for any subset of indices $I \subseteq [n]$*

$$\mathrm{P}\left[\bigcap_{i \in I} A_i\right] = \prod_{i \in I} \mathrm{P}[A_i].$$

We use the convenient notation $[n]$ for the set $\{1, 2, \dots, n\}$.

The independence of $A_1, A_2, \dots, A_n$ is not equivalent to all the pairs A_i, A_j being independent. Exercise: find three events A_1, A_2 and A_3 that are pairwise independent but not mutually independent.

Intuitively, the property of independence means that the knowledge of whether some of the events $A_1, \dots, A_n$ occurred does not provide any information regarding the remaining events.

1.1.5 Definition (Conditional probability). *For events A and B with $\mathrm{P}[B] > 0$, we define the conditional probability*[11] *of A, given that B occurs, as*

$$\mathrm{P}[A|B] = \frac{\mathrm{P}[A \cap B]}{\mathrm{P}[B]}.$$

[7] toss a fair coin = hodit spravedlivou mincí

[8] heads = líc (hlava)

[9] tails = rub (orel)

[10] independent events = nezávislé jevy

[11] conditional probability = podmíněná pravděpodobnost

Note that if A and B are independent, then $\mathrm{P}[A|B] = \mathrm{P}[A]$.

1.1.6 Definition (Random variable). *A real random variable*[12] *on a probability space* $(\Omega, \Sigma, \mathrm{P})$ *is a function* $X\colon \Omega \to \mathbf{R}$ *that is* P*-measurable. (That is, for any* $a \in \mathbf{R}$, $\{\omega \in \Omega\colon X(\omega) \le a\} \in \Sigma$.*)*

We can also consider random variables with other than real values; for example, a random variable can have complex numbers or n-component vectors of real numbers as values. In such cases, a random variable is a measurable function from the probability space into the appropriate space with measure (complex numbers or $\mathbf{R}^n$ in the examples mentioned above). In this text, we will mostly consider real random variables.

1.1.7 Definition. *The* **expectation**[13] *of a (real) random variable* X *is*

$$\mathbf{E}[X] = \int_\Omega X(\omega)\,\mathrm{dP}(\omega).$$

Any real function on a finite probability space is a random variable. Its expectation can be expressed as

$$\mathbf{E}[X] = \sum_{\omega\in\Omega} p(\omega)X(\omega).$$

1.1.8 Definition (Independence of variables). *Real random variables* X, Y *are independent if we have, for every two measurable sets* $A, B \subseteq \mathbf{R}$,

$$\mathrm{P}[X \in A \text{ and } Y \in B] = \mathrm{P}[X \in A] \cdot \mathrm{P}[Y \in B].$$

Note the shorthand notation for the events in the previous definition: For example, $\mathrm{P}[X \in A]$ stands for $\mathrm{P}[\{\omega \in \Omega\colon X(\omega) \in A\}]$.

Intuitively, the independence of X and Y means that the knowledge of the value attained by X gives us no information about Y, and vice versa. In order to check independence, one need not consider all measurable sets A and B; it is sufficient to look at $A = (-\infty, a]$ and $B = (-\infty, b]$. That is, if

$$\mathrm{P}[X \le a \text{ and } Y \le b] = \mathrm{P}[X \le a]\,\mathrm{P}[Y \le b]$$

for all $a, b \in \mathbf{R}$, then X and Y are independent.

As we will check in Chapter 3, $\mathbf{E}[X + Y] = \mathbf{E}[X] + \mathbf{E}[Y]$ holds for *any* two random variables (provided that the expectations exist). On the other hand, $\mathbf{E}[XY]$ is generally different from $\mathbf{E}[X]\,\mathbf{E}[Y]$. But we have

1.1.9 Lemma. *If* X *and* Y *are independent random variables, then*

$$\mathbf{E}[XY] = \mathbf{E}[X] \cdot \mathbf{E}[Y].$$

[12] random variable = náhodná proměnná

[13] expectation = **střední hodnota!!!**

Proof (for finite probability spaces). If X and Y are random variables on a finite probability space, the proof is especially simple. Let V_X, V_Y be the (finite) sets of values attained by X and by Y, respectively. By independence, we have $\mathrm{P}[X=a \text{ and } Y=b] = \mathrm{P}[X=a]\,\mathrm{P}[Y=b]$ for any $a \in V_X$ and $b \in V_Y$. We calculate

$$\begin{aligned}\mathbf{E}[XY] &= \sum_{a\in V_X, b\in V_Y} ab\cdot \mathrm{P}[X=a \text{ and } Y=b] \\ &= \sum_{a\in V_X, b\in V_Y} ab\cdot \mathrm{P}[X=a]\,\mathrm{P}[Y=b] \\ &= \Bigl(\sum_{a\in V_X} a\,\mathrm{P}[X=a]\Bigr)\Bigl(\sum_{b\in V_Y} b\,\mathrm{P}[Y=b]\Bigr) = \mathbf{E}[X]\,\mathbf{E}[Y]\,.\end{aligned}$$

For infinite probability spaces, the proof is formally a little more complicated but the idea is the same. □

1.2 Useful Estimates

In the probabilistic method, many problems are reduced to showing that certain probability is below 1, or even tends to 0. In the final stage of such proofs, we often need to estimate some complicated-looking expressions. The golden rule here is to start with the roughest estimates, and only if they don't work, one can try more refined ones. Here we describe the most often used estimates for basic combinatorial functions.

For the factorial function $n!$, we can often do with the obvious upper bound $n! \le n^n$. More refined bounds are

$$\left(\frac{n}{e}\right)^n \le n! \le en\left(\frac{n}{e}\right)^n$$

(where $e = 2.718281828\ldots$ is the basis of natural logarithms), which can be proved by induction. The well-known Stirling formula is very seldom needed in its full strength.

For the binomial coefficient $\binom{n}{k}$, the basic bound is $\binom{n}{k} \le n^k$, and sharper ones are

$$\left(\frac{n}{k}\right)^k \le \binom{n}{k} \le \left(\frac{en}{k}\right)^k.$$

For all k, we also have $\binom{n}{k} \le 2^n$. Sometimes we need better estimates of the middle binomial coefficient $\binom{2m}{m}$; we have

$$\frac{2^{2m}}{2\sqrt{m}} \le \binom{2m}{m} \le \frac{2^{2m}}{\sqrt{2m}}$$

(also see Section 5.2 for a derivation of a slightly weaker lower bound).

Very often we need the inequality $1+x \le e^x$, valid for all real x. In particular, for bounding expressions of the form $(1-p)^m$ from above, with $p > 0$ small, one uses

$$(1-p)^m \le e^{-mp}$$

almost automatically. For estimating such expressions from below, which is usually more delicate, we can often use

$$1 - p \geq e^{-2p},$$

which is valid for $0 \leq p \leq \frac{1}{2}$.

2

The Probabilistic Method

The probabilistic method is a remarkable technique for proving the existence of combinatorial objects with specified properties. It is based on probability theory but, surprisingly, it can be used for proving theorems that have nothing to do with probability. The usual approach can be described as follows.

We would like to prove the existence of a combinatorial object with specified properties. Unfortunately, an explicit construction of such a "good" object does not seem feasible, and maybe we do not even need a specific example; we just want to prove that something "good" exists. Then we can consider a random object from a suitable probability space and calculate the probability that it satisfies our conditions. If we prove that this probability is strictly positive, then we conclude that a "good" object must exist; if all objects were "bad", the probability would be zero.

Let us start with an example illustrating how the probabilistic method works in its basic form.

2.1 Ramsey Numbers

The Ramsey theorem states that any sufficiently large graph contains either a clique or an independent set of a given size. (A *clique*[1] is a set of vertices inducing a complete subgraph and an *independent set*[2] is a set of vertices inducing an edgeless subgraph.)

2.1.1 Definition. *The* **Ramsey number** $R(k,\ell)$ *is*

$$\begin{aligned} R(k,\ell) \;=\; \min\{n\colon & \text{ any graph on } n \text{ vertices contains a clique} \\ & \text{of size } k \text{ or an independent set of size } \ell\}. \end{aligned}$$

The Ramsey theorem guarantees that $R(k,\ell)$ is always finite. Still, the precise values of $R(k,\ell)$ are unknown but for a small number of cases, and it is desirable at least to estimate $R(k,\ell)$ for large k and ℓ. Here we use the probabilistic method to prove a lower bound on $R(k,k)$.

[1]clique = klika (úplný podgraf)
[2]independent set = nezávislá množina

2.1.2 Theorem. *For any $k \geq 3$,*

$$R(k,k) > 2^{k/2-1}.$$

Proof. Let us consider a random graph $G(n,1/2)$ on n vertices where every pair of vertices forms an edge with probability $\frac{1}{2}$, independently of the other edges. (We can imagine flipping a coin for every potential edge to decide whether it should appear in the graph.) For any fixed set of k vertices, the probability that they form a clique is

$$p = 2^{-\binom{k}{2}}.$$

The same goes for the occurrence of an independent set, and there are $\binom{n}{k}$ k-tuples of vertices where a clique or an independent set might appear. Now we use the fact that the probability of a union of events is at most the sum of their respective probabilities (Lemma 1.1.3), and we get

$$\mathrm{P}\left[G(n,1/2) \text{ contains a clique or an indep. set of size } k\right] \leq 2\binom{n}{k} 2^{-\binom{k}{2}}.$$

It remains to choose n so that the last expression is below 1. Using the simplest estimate $\binom{n}{k} \leq n^k$, we find that it is sufficient to have $2n^k < 2^{k(k-1)/2}$. This certainly holds whenever $n \leq 2^{k/2-1}$. Therefore, there are graphs on $\lfloor 2^{k/2-1} \rfloor$ vertices that contain neither a clique of size k nor an independent set of size k. This implies $R(k,k) > 2^{k/2-1}$. □

Let us remark that, by using finer estimates in the proof, the lower bound for $R(k,k)$ can be improved a little, say to $2^{k/2}$. But a result even slightly better than this seems to require a more powerful technique. In particular, no lower bound is known of the form c^k with a constant $c > \sqrt{2}$, although the best upper bound is about 4^k.

One might object that the use of a probability space is artificial here and the same proof can be formulated in terms of counting objects. In effect, we are counting the number of bad objects and trying to prove that it is less than the number of all objects, so the set of good objects must be nonempty. In simple cases, it is indeed possible to phrase such proofs in terms of counting bad objects. However, in more sophisticated proofs, the probabilistic formalism becomes much simpler than counting arguments. Furthermore, the probabilistic framework allows us to use many results of probability theory—a mature mathematical discipline.

For many important problems, the probabilistic method has provided the only known solution, and for others, it has provided accessible proofs in cases where constructive proofs are extremely difficult.

2.2 Hypergraph Coloring

2.2.1 Definition. *A k-uniform hypergraph is a pair (X,S) where X is the set of vertices and $S \subseteq \binom{X}{k}$ is the set of edges (k-tuples of vertices).*

2.2.2 Definition. *A hypergraph is c-colorable if its vertices can be colored with c colors so that no edge is monochromatic (at least two different colors appear in every edge).*

This is a generalization of the notion of graph coloring. Note that graphs are 2-uniform hypergraphs and the condition of proper coloring requires that the vertices of every edge get two different colors.

Now we will be interested in the smallest possible number of edges in a k-uniform hypergraph that is not 2-colorable.

2.2.3 Definition. *Let* $\boldsymbol{m(k)}$ *denote the smallest number of edges in a k-uniform hypergraph that is not 2-colorable.*

For graphs, we have $m(2) = 3$, because the smallest non-bipartite graph is a triangle. However, the problem becomes much more difficult for larger k. As we will prove, $m(3) = 7$, but the exact value of $m(k)$ is unknown for $k > 3$.

Again, we can get a lower bound by probabilistic reasoning.

2.2.4 Theorem. *For any $k \geq 2$,*

$$m(k) \geq 2^{k-1}.$$

Proof. Let us consider a k-uniform hypergraph $\mathcal{H}$ with less than 2^{k-1} edges. We will prove that it is 2-colorable.

We color every vertex of $\mathcal{H}$ independently red or blue, with probability $\frac{1}{2}$. The probability that the vertices of a given edge are all red or all blue is $p = 2 \cdot (\frac{1}{2})^k$. Supposing $\mathcal{H}$ has $|S| < 2^{k-1}$ edges, the probability that there exists a monochromatic edge is at most $p|S| < p2^{k-1} = 1$. So there is a non-zero probability that no edge is monochromatic and a proper coloring must exist.

□

Note that for $k = 3$, we get $m(3) \geq 4$. On the other hand, the smallest known 3-uniform hypergraph that is not 2-colorable is the finite projective plane with 7 points, the *Fano plane*.

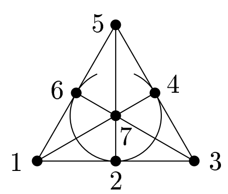
2.2.5 Definition. *The* **Fano plane** *is the hypergraph $\mathcal{H} = (X, S)$, where*

$$X = \{1, 2, 3, 4, 5, 6, 7\}$$

are the points and

$$S = \{\{1, 2, 3\}, \{3, 4, 5\}, \{5, 6, 1\}, \{1, 7, 4\}, \{2, 7, 5\}, \{3, 7, 6\}, \{2, 4, 6\}\}$$

are the edges.



2.2.6 Lemma. $m(3) \leq 7$.

Proof. We prove that the Fano plane is not 2-colorable. We give a quick argument using the fact that $\mathcal{H}$ is a projective plane, and thus for any two points, there is exactly one edge (line) containing both of them.

Suppose that we have a 2-coloring $A_1 \cup A_2 = X, A_1 \cap A_2 = \emptyset$, where A_1 is the larger color class.

If $|A_1| \geq 5$, then A_1 contains at least $\binom{5}{2} = 10$ pairs of points. Each pair defines a unique line, but as there are only 7 lines in total, there must be two pairs of points defining the same line. So we have three points of the same color on a line.

If $|A_1| = 4$ then A_1 contains $\binom{4}{2} = 6$ pairs of points. If two pairs among them define the same line, that line is monochromatic and we are done. So suppose that these 6 pairs define different lines $\ell_1, \ldots, \ell_6$. Then each point of A_1 is intersected by 3 of the ℓ_i. But since each point in the Fano plane lies on exactly 3 lines and there are 7 lines in total, there is a line not intersecting A_1 at all. That line is contained in A_2 and thus monochromatic. $\square$

Now we will improve the lower bound to establish that $m(3) = 7$.

2.2.7 Theorem. *Any system of 6 triples is 2-colorable; i.e.* $m(3) \geq 7$.

Proof: Let us consider a 3-uniform hypergraph $\mathcal{H} = (X, S), |S| \leq 6$. We want to prove that $\mathcal{H}$ is 2-colorable. We will distinguish two cases, depending on the size of X.

If $|X| \leq 6$, we apply the probabilistic method. We can assume that $|X| = 6$, because we can always add vertices that are not contained in any edge and therefore do not affect the coloring condition. Then we choose a random subset of 3 vertices which we color red and the remaining vertices become blue. The total number of such colorings is $\binom{6}{3} = 20$. For any edge (which is a triple of vertices), there are two colorings that make it either completely red or completely blue, so the probability that it is monochromatic is $\frac{1}{10}$. We have at most 6 edges, and so the probability that any of them is monochromatic is at most $\frac{6}{10} < 1$.

For $|X| > 6$, we proceed by induction. Suppose that $|X| > 6$ and $|S| \leq 6$. It follows that there exist two vertices $x, y \in X$ that are not "connected" (a pair of vertices is connected if they appear together in some edge). This is because every edge produces three connected pairs, so the number of connected pairs is at most 18. On the other hand, the total number of vertex pairs is at least $\binom{7}{2} = 21$, so they cannot be all connected.

Now if $x, y \in X$ are not connected, we define a new hypergraph by merging x and y into one vertex:

$$X' = X \setminus \{x, y\} \cup \{z\},$$

$$S' = \{M \in S\colon M \cap \{x, y\} = \emptyset\} \cup \{M \setminus \{x, y\} \cup \{z\}\colon M \in S, M \cap \{x, y\} \neq \emptyset\}.$$

This (X', S') is a 3-uniform hypergraph as well, $|S'| = |S| \leq 6$, and $|X'| = |X| - 1$, so by the induction hypothesis it is 2-colorable. If we extend the coloring of X' to X so that both x and y get the color of z, we obtain a proper 2-coloring for (X, S). $\square$

2.3 The Erdős–Ko–Rado Theorem

2.3.1 Definition. *A family $\mathcal{F}$ of sets is* **intersecting** *if for all $A, B \in \mathcal{F}$, $A \cap B \neq \emptyset$.*

2.3.2 Theorem (The Erdős–Ko–Rado Theorem). *If $|X| = n$, $n \geq 2k$, and $\mathcal{F}$ is an intersecting family of k-element subsets of X, then*

$$|\mathcal{F}| \leq \binom{n-1}{k-1}.$$

Clearly, this is tight, because a family of all the k-element subsets containing a particular point is intersecting and the number of such subsets is $\binom{n-1}{k-1}$. (This configuration is sometimes called a *sunflower* and the theorem is referred to as the Sunflower Theorem.)

2.3.3 Lemma. *Consider $X = \{0, 1, \ldots, n-1\}$ with addition modulo n and define $A_s = \{s, s+1, \ldots, s+k-1\} \subseteq X$ for $0 \leq s < n$. Then for $n \geq 2k$, any intersecting family $\mathcal{F} \subseteq \binom{X}{k}$ contains at most k of the sets A_s.*

Proof. If $A_i \in \mathcal{F}$, then any other $A_s \in \mathcal{F}$ must be one of the sets $A_{i-k+1}, \ldots, A_{i-1}$ or $A_{i+1}, \ldots, A_{i+k-1}$. These are $2k-2$ sets, which can be divided into $k-1$ pairs of the form (A_s, A_{s+k}). As $n \geq 2k$, $A_s \cap A_{s+k} = \emptyset$, and only one set from each pair can appear in $\mathcal{F}$. □

Proof of the theorem. We can assume that $X = \{0, 1, \ldots, n-1\}$ and $\mathcal{F} \subseteq \binom{X}{k}$ is an intersecting family. For a permutation $\sigma: X \to X$, we define

$$\sigma(A_s) = \{\sigma(s), \sigma(s+1), \ldots, \sigma(s+k-1)\},$$

addition again modulo n. The sets $\sigma(A_s)$ are just like those in the lemma, only with the elements relabeled by the permutation σ, so by the lemma at most k of these n sets are in $\mathcal{F}$. Therefore, if we choose random s and σ independently and uniformly,

$$\mathrm{P}[\sigma(A_s) \in \mathcal{F}] \leq \frac{k}{n}$$

(the underlying probability space here is the product $[n] \times S_n$ with the uniform measure, where S_n is the set of all permutations on $[n]$). But this choice of $\sigma(A_s)$ is equivalent to a random choice of a k-element subset of X, so

$$\mathrm{P}[\sigma(A_s) \in \mathcal{F}] = \frac{|\mathcal{F}|}{\binom{n}{k}}$$

and

$$|\mathcal{F}| = \binom{n}{k} \mathrm{P}[\sigma(A_s) \in \mathcal{F}] \leq \binom{n}{k} \frac{k}{n} = \binom{n-1}{k-1}.$$

□

2.4 Pairs of Sets

Let k and ℓ be fixed natural numbers. We are interested in the maximum $n = n(k,\ell)$ such that there exist sets $A_1, A_2, \ldots, A_n$ and $B_1, B_2, \ldots, B_n$ satisfying the following conditions

(C0) $|A_i| = k$, $|B_i| = \ell$ for all $i = 1, 2, \ldots, n$.

(C1) $A_i \cap B_i = \emptyset$ for all $i = 1, 2, \ldots, n$.

(C2) $A_i \cap B_j \neq \emptyset$ for all $i \neq j$, $i, j = 1, 2, \ldots, n$.

An example shows that $n(k,\ell) \geq \binom{k+\ell}{k}$: let $A_1, \ldots, A_n$ be all the k-element subsets of $\{1, 2, \ldots, k+\ell\}$ and let B_i be the complement of A_i. An ingenious probabilistic argument shows that this is the best possible (note that at first sight, it is not at all obvious that $n(k,\ell)$ is finite!).

2.4.1 Theorem. *For any $k, \ell \geq 1$, we have $n(k,\ell) = \binom{k+\ell}{k}$.*

Before we prove this theorem, we explain a motivation for this (perhaps strange-looking) problem. It is related to the *transversal number* of set systems, one of the central issues in combinatorics. Recall that a set $T \subseteq X$ is a *transversal* of a set system $\mathcal{F} \subseteq 2^X$ if $S \cap T \neq \emptyset$ for all $S \in \mathcal{F}$. The transversal number $\tau(\mathcal{F})$ is the size of the smallest transversal of $\mathcal{F}$.

In order to understand a combinatorial parameter, one usually studies the *critical* objects. In our case, a set system $\mathcal{F}$ is called *τ-critical* if $\tau(\mathcal{F} \setminus \{S\}) < \tau(\mathcal{F})$ for each $S \in \mathcal{F}$. A question answered by the above theorem was the following: what is the maximum possible number of sets in a τ-critical system $\mathcal{F}$, consisting of k-element sets and with $\tau(\mathcal{F}) = \ell+1$? To see the connection, let $\mathcal{F} = \{A_1, A_2, \ldots, A_n\}$, and let B_i be an ℓ-element transversal of $\mathcal{F} \setminus \{A_i\}$. Note that by the τ-criticality of $\mathcal{F}$, the B_i exist and satisfy conditions (C0)–(C2). Thus $|\mathcal{F}| \leq n(k,\ell)$.

Proof of Theorem 2.4.1. Let $X = \bigcup_{i=1}^n (A_i \cup B_i)$ be the ground set. Arrange the elements of X in a random linear order (all the $|X|!$ orderings having the same probability). Let U_i be the event "each element of A_i precedes each element of B_i". We have $P[U_i] = \binom{k+\ell}{k}^{-1}$.

Crucially, we note that U_i and U_j cannot occur simultaneously for $i \neq j$. Indeed, since $A_i \cap B_j \neq \emptyset \neq A_j \cap B_i$, we have $\max A_i \geq \min B_j$ and $\max A_j \geq \min B_i$. If both U_i and U_j occurred, then $\max A_i < \min B_i$ and $\max A_j < \min B_j$, and we get a contradiction: $\max A_i \geq \min B_j > \max A_j \geq \min B_i > \max A_i$. Therefore

$$1 \geq \mathrm{P}\left[\bigcup_{i=1}^n U_i\right] = \sum_{i=1}^n \mathrm{P}[U_i] = \frac{n}{\binom{k+\ell}{k}}$$

and the theorem follows. $\square$

The same proof shows that if $A_1, A_2, \ldots, A_n$ and $B_1, B_2, \ldots, B_n$ are finite sets satisfying (C1) and (C2) then $\sum_{i=1}^n \binom{|A_i|+|B_i|}{|A_i|}^{-1} \leq 1$. This implies, among others, the famous *Sperner theorem*: If $\mathcal{F}$ is a family of subsets of $[m]$ with no two distinct sets $A, B \in \mathcal{F}$ satisfying $A \subset B$, then $|\mathcal{F}| \leq \binom{m}{\lfloor m/2 \rfloor}$. To see this, set $\mathcal{F} = \{A_1, A_2, \ldots, A_n\}$ and $B_i = [m] \setminus A_i$, and use the fact that $\binom{m}{k} \leq \binom{m}{\lfloor m/2 \rfloor}$ for all $k = 0, 1, \ldots, m$.

3

Linearity of Expectation

3.1 Computing Expectation Using Indicators

The proofs in this chapter are based on the following lemma:

3.1.1 Lemma. *The expectation is a linear operator; i.e., for any two random variables X, Y and constants $\alpha, \beta \in \mathbf{R}$:*

$$\mathbf{E}[\alpha X + \beta Y] = \alpha \mathbf{E}[X] + \beta \mathbf{E}[Y].$$

Proof. $\mathbf{E}[\alpha X + \beta Y] = \int_\Omega (\alpha X + \beta Y) \,\mathrm{dP} = \alpha \int_\Omega X \,\mathrm{dP} + \beta \int_\Omega Y \,\mathrm{dP} = \alpha \mathbf{E}[X] + \beta \mathbf{E}[Y].$ □

This implies that the expectation of a sum of random variables $X = X_1 + X_2 + \cdots + X_n$ is equal to

$$\mathbf{E}[X] = \mathbf{E}[X_1] + \mathbf{E}[X_2] + \cdots + \mathbf{E}[X_n].$$

This fact is elementary, yet powerful, since there is no restriction whatsoever on the properties of X_i, their dependence or independence.

3.1.2 Definition (Indicator variables). *For an event A, we define the indicator variable I_A:*

- $I_A(\omega) = 1$ *if* $\omega \in A$, *and*
- $I_A(\omega) = 0$ *if* $\omega \notin A$.

3.1.3 Lemma. *For any event A, we have* $\mathbf{E}[I_A] = \mathrm{P}[A]$.

Proof.

$$\mathbf{E}[I_A] = \int_\Omega I_A(\omega) \,\mathrm{dP} = \int_A \mathrm{dP} = \mathrm{P}[A].$$

□

In many cases, the expectation of a variable can be calculated by expressing it as a sum of indicator variables

$$X = I_{A_1} + I_{A_2} + \cdots + I_{A_n}$$

of certain events with known probabilities. Then

$$\mathbf{E}[X] = \mathrm{P}[A_1] + \mathrm{P}[A_2] + \cdots + \mathrm{P}[A_n].$$

Example. Let us calculate the expected number of fixed points of a random permutation σ on $\{1, \ldots, n\}$. If

$$X(\sigma) = |\{i\colon \sigma(i) = i\}|,$$

we can express this as a sum of indicator variables:

$$X(\sigma) = \sum_{i=1}^{n} X_i(\sigma)$$

where $X_i(\sigma) = 1$ if $\sigma(i) = i$ and 0 otherwise. Then

$$\mathbf{E}[X_i] = \mathrm{P}[\sigma(i) = i] = \frac{1}{n}$$

and

$$\mathbf{E}[X] = \frac{1}{n} + \frac{1}{n} + \cdots + \frac{1}{n} = 1.$$

So a random permutation has 1 fixed point (or "loop") on the average.

3.2 Hamiltonian Paths

We can use the expectation of X to estimate the minimum or maximum value of X, because there always exists an elementary event $\omega \in \Omega$ for which $X(\omega) \geq \mathbf{E}[X]$ and similarly, we have $X(\omega) \leq \mathbf{E}[X]$ for some $\omega \in \Omega$.

We recall that a *tournament* is an orientation of a complete graph (for any two vertices u, v, exactly one of the directed edges (u, v) and (v, u) is present). A *Hamiltonian path* in a tournament is a directed path passing through all vertices. The following result of Szele (1943) shows the existence of a tournament with very many Hamiltonian paths.

3.2.1 Theorem. *There is a tournament on n vertices that has at least $\frac{n!}{2^{n-1}}$ Hamiltonian paths.*

Proof. Let us calculate the expected number of Hamiltonian paths in a random tournament T (every edge has a random orientation, chosen independently with probability $\frac{1}{2}$). For a given permutation σ on $\{1, \ldots, n\}$, consider the sequence $\{\sigma(1), \sigma(2), \ldots, \sigma(n)\}$ and denote by X_σ the indicator of the event that all the edges $(\sigma(i), \sigma(i+1))$ appear in T with this orientation. Because the orientation of different edges is chosen independently,

$$\mathbf{E}[X_\sigma] = \mathrm{P}[(\sigma(i), \sigma(i+1)) \in T \text{ for } i = 1, 2, \ldots, n-1] = \frac{1}{2^{n-1}}.$$

The total number of Hamiltonian paths X equals the sum of these indicator variables over all potential Hamiltonian paths, i.e. permutations, and so

$$\mathbf{E}[X] = \sum_{\sigma} \mathbf{E}[X_\sigma] = \frac{n!}{2^{n-1}}.$$

So there is a tournament with at least $\frac{n!}{2^{n-1}}$ Hamiltonian paths. □

3.3 Splitting Graphs

The MAXCUT problem is the following important algorithmic problem: Given a graph $G = (V, E)$, divide the vertex set into two classes A and $B = V \setminus A$ so that the number of edges going between A and B is maximized. This problem is computationally hard (NP-complete). The following simple result tells us that it is always possible to achieve at least half of the edges going between A and B.

3.3.1 Theorem. *Any graph with m edges contains a bipartite subgraph with at least $\frac{m}{2}$ edges.*

Proof. Let $G = (V, E)$, and choose a random subset $T \subseteq V$ by inserting every vertex into T independently with probability $\frac{1}{2}$. For a given edge $e = \{u, v\}$, let X_e denote the indicator variable of the event that *exactly one* of the vertices of e is in T. Then we have

$$\mathbf{E}[X_e] = \mathrm{P}[(u \in T \ \&\ v \notin T) \text{ or } (u \notin T \ \&\ v \in T)] = \tfrac{1}{4} + \tfrac{1}{4} = \tfrac{1}{2}.$$

If X denotes the number of edges having exactly one vertex in T, then

$$\mathbf{E}[X] = \sum_{e \in E} \mathbf{E}[X_e] = \frac{m}{2}.$$

Thus for some $T \subseteq V$, there are at least $\frac{m}{2}$ edges crossing between T and $V \setminus T$, forming a bipartite graph. $\square$

4

Alterations

Sometimes the first attempt to find a "good" object by random construction fails, but we prove that there exists an object which *almost* satisfies our conditions. Often it is possible to modify it in a deterministic way so that we get what we need.

Before we begin with examples, let us mention one simple tool which is useful when we need to estimate the probability that a random variable exceeds its expectation significantly.

4.0.2 Lemma (Markov's inequality). *If X is a non-negative random variable and $a > 0$, then*

$$\mathrm{P}[X \geq a] \leq \frac{\mathbf{E}[X]}{a}.$$

Proof. If X is non-negative, then

$$\mathbf{E}[X] \geq a \cdot \mathrm{P}[X \geq a].$$

□

4.1 Independent Sets

4.1.1 Definition (Independence number). *For a graph G, $\alpha(G)$ denotes the size of the largest independent set in G (a set of vertices such that no two of them are joined by an edge).*

The independence number of a graph is one of its basic parameters. We would like to know how it depends on the number of edges in the graph; specifically, how small the independence number can be for a given average degree.

4.1.2 Theorem (A weak Turán theorem). *If n is the number of vertices of G, m is the number of edges, and $d = \frac{2m}{n} \geq 1$ is the average degree, then*

$$\alpha(G) \geq \frac{n}{2d}.$$

Note. By Turán's theorem, we actually have $\alpha(G) \geq \frac{n}{d+1}$, and this is the best possible in general. For d integral, the extremal graph is a union of disjoint cliques of size $d+1$.

Proof. First, let us select a random subset of vertices $S \subseteq V$ in such a way that we insert every vertex into S independently with probability p (we will choose a suitable value of p later). If X denotes the size of S and Y denotes the number of edges in $G[S]$ (the subgraph induced by S), then

$$\mathbf{E}[X] = np$$

(this follows immediately by the method of indicators; see Section 3.1) and

$$\mathbf{E}[Y] = mp^2 = \tfrac{1}{2}ndp^2$$

(because the probability that both vertices of a given edge are in S is p^2).

We get

$$\mathbf{E}[X - Y] = np(1 - \tfrac{1}{2}dp),$$

so there exists $S \subseteq V$ where the difference of the number of vertices and edges is at least $A(p) = np(1 - \frac{1}{2}dp)$.

Now observe that we can modify S by removing one vertex from each edge inside S. We obtain an independent set with at least $A(p)$ vertices. It remains to choose the value of p so as to maximize $A(p)$; the optimal value is $p = \frac{1}{d}$, which yields

$$A(p) = \frac{n}{2d}.$$

□

4.2 High Girth and High Chromatic Number

Now we turn to a famous problem (solved by Erdős). The question was whether the non-existence of short cycles in a graph implies that it can be colored with a small number of colors. The answer is negative: there are graphs that do not contain any short cycles and yet their chromatic number is arbitrarily large.

We recall that a *(proper) k-coloring* of a graph G is a mapping $c\colon V(G) \to [k]$ such that $c(u) \neq c(v)$ whenever $\{u, v\} \in E(G)$, and the *chromatic number*[1] of G, denoted by $\chi(G)$, is the smallest k such that G has a proper k-coloring. The *girth*[2] of a graph G, denoted by $g(G)$, is the length of its shortest cycle.

4.2.1 Theorem. *For any $k, \ell > 0$, there exists a graph G such that $\chi(G) > k$ and $g(G) > \ell$.*

Proof. Set $\varepsilon = \frac{1}{2\ell}$, $p = n^{\varepsilon - 1}$, and consider the random graph $G(n, p)$. First, we estimate the number of cycles of length at most ℓ, which we denote by X.

[1]chromatic number = barevnost
[2]girth = obvod

Since the number of potential cycles of length i is $\frac{1}{2}(i-1)!\binom{n}{i} \leq n^i$ and each of them is present with probability p^i, we get

$$\mathbf{E}[X] \leq \sum_{i=3}^{\ell} n^i p^i = \sum_{i=3}^{\ell} n^{\varepsilon i}.$$

Because $n^{\varepsilon i} = o(n)$ for all $i \leq \ell$, $\mathbf{E}[X] = o(n)$. If we choose n so large that $\mathbf{E}[X] < \frac{n}{4}$, we get by the Markov inequality

$$\mathrm{P}\left[X \geq \tfrac{n}{2}\right] < \tfrac{1}{2}.$$

Now we estimate the chromatic number of $G(n,p)$ by means of its independence number. If we set $a = \lceil \frac{3}{p} \ln n \rceil$, we have

$$\mathrm{P}[\alpha(G(n,p)) \geq a] \leq \binom{n}{a}(1-p)^{\binom{a}{2}} \leq n^a e^{-p\binom{a}{2}} = e^{(\ln n - p(a-1)/2)a},$$

which tends to zero as $n \to \infty$. Thus again, for n sufficiently large, we have

$$\mathrm{P}[\alpha(G(n,p)) \geq a] < \tfrac{1}{2}.$$

Consequently, there exists a graph G with $X < \frac{n}{2}$ and $\alpha(G) < a$. If we remove one vertex from each of the X short cycles, at least $\frac{n}{2}$ vertices remain and we get a graph G^* with $g(G^*) > \ell$ and $\alpha(G^*) < a$. Since in any proper coloring of G^*, the color classes are independent sets of size at most $a-1$,

$$\chi(G^*) \geq \frac{n/2}{a-1} \geq \frac{pn}{6\ln n} = \frac{n^{\varepsilon}}{6\ln n}.$$

It remains only to choose n sufficiently large so that $\chi(G^*) > k$. $\square$

5

The Second Moment

5.1 Variance and the Chebyshev Inequality

Besides the expectation, the other essential characteristic of a random variable is the variance.[1] It describes how much the variable fluctuates around its expectation. (For a constant random variable, the variance is zero.)

5.1.1 Definition. *The* **variance** *of a real random variable X is*

$$\mathrm{Var}\,[X] = \mathbf{E}\left[(X - \mathbf{E}\,[X])^2\right] = \mathbf{E}\left[X^2\right] - (\mathbf{E}\,[X])^2.$$

(The first equality is a definition, and the second one follows by an easy computation.) The standard deviation[2] *of X is $\sigma = \sqrt{\mathrm{Var}\,[X]}$.*

Unlike the expectation, the variance is *not* a linear operator. If we want to calculate the variance of a sum of random variables, we need to know something about their pairwise dependence.

5.1.2 Definition. *The covariance*[3] *of two random variables is*

$$\mathrm{Cov}\,[X,Y] = \mathbf{E}\,[(X - \mathbf{E}\,[X])(Y - \mathbf{E}\,[Y])] = \mathbf{E}\,[XY] - \mathbf{E}\,[X]\,\mathbf{E}\,[Y]\,.$$

5.1.3 Lemma. *The variance of a sum of random variables is equal to*

$$\mathrm{Var}\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n \mathrm{Var}\,[X_i] + \sum_{i\neq j} \mathrm{Cov}\,[X_i, X_j].$$

Proof.

$$\mathrm{Var}\left[\sum_{i=1}^n X_i\right] = \mathbf{E}\left[\sum_{i=1}^n X_i \sum_{j=1}^n X_j\right] - \mathbf{E}\left[\sum_{i=1}^n X_i\right]\mathbf{E}\left[\sum_{j=1}^n X_j\right] =$$

$$= \sum_{i=1}^n \mathbf{E}\left[X_i^2\right] + \sum_{i\neq j} \mathbf{E}\,[X_i X_j] - \sum_{i=1}^n (\mathbf{E}\,[X_i])^2 - \sum_{i\neq j} \mathbf{E}\,[X_i]\,\mathbf{E}\,[X_j] =$$

[1] variance = **rozptyl**
[2] standard deviation = **směrodatná odchylka**
[3] covariance = kovariance

$$= \sum_{i=1}^{n} \mathrm{Var}\,[X_i] + \sum_{i \neq j} \mathrm{Cov}\,[X_i, X_j].$$

□

Note. If $X_1, \ldots, X_n$ are independent, the covariance of each pair is 0. In this case, the variance of X can be calculated as the sum of variances of the X_i. On the other hand, $\mathrm{Cov}\,[X, Y] = 0$ does *not* imply independence of X and Y!

Once we know the variance, we can apply the *Chebyshev inequality*[4] to estimate the probability that a random variable deviates from its expectation at least by a given number.

5.1.4 Lemma (Chebyshev inequality). *Let X be a random variable with a finite variance. Then for any $t > 0$*

$$\mathrm{P}\left[|X - \mathbf{E}\,[X]\,| \geq t\right] \leq \frac{\mathrm{Var}\,[X]}{t^2}.$$

Proof.

$$\mathrm{Var}\,[X] = \mathbf{E}\left[(X - \mathbf{E}\,[X])^2\right] \geq t^2\, \mathrm{P}\left[|X - \mathbf{E}\,[X]\,| \geq t\right].$$

□

This simple tool gives the best possible result when X is equal to μ with probability p and equal to $\mu \pm t$ with probability $\frac{1-p}{2}$. In Chapter 7, we will examine stronger methods giving better bounds for certain classes of random variables. In this section, though, the Chebyshev inequality will be sufficient.

5.2 Estimating the Middle Binomial Coefficient

Among the binomial coefficients $\binom{2m}{k}$, $k = 0, 1, \ldots, 2m$, $\binom{2m}{m}$ is the largest and it often appears in various formulas (e.g. in the Catalan numbers, which count binary trees and many other things). The second moment method provides a simple way of bounding $\binom{2m}{m}$ from below. There are several other approaches, some of them yielding much more precise estimates, but the simple trick with the Chebyshev inequality gives the correct order of magnitude.

5.2.1 Proposition. *For all $m \geq 1$, we have $\binom{2m}{m} \geq 2^{2m}/(4\sqrt{m}+2)$.*

Proof. Consider the random variable $X = X_1 + X_2 + \cdots + X_{2m}$, where the X_i are independent and each of them attains values 0 and 1 with probability $\frac{1}{2}$. We have $\mathbf{E}\,[X] = m$ and $\mathrm{Var}\,[X] = \frac{m}{2}$. The Chebyshev inequality with $t = \sqrt{m}$ gives

$$\mathrm{P}\left[|X - m| < \sqrt{m}\,\right] \geq \frac{1}{2}.$$

The probability of X attaining a specific value $m + k$, where $|k| < \sqrt{m}$, is $\binom{2m}{m+k}2^{-2m} \leq \binom{2m}{m}2^{-2m}$ (because $\binom{2m}{m}$ is the largest binomial coefficient). So we have

$$\frac{1}{2} \leq \sum_{|k| < \sqrt{m}} \mathrm{P}\left[X = m + k\right] \leq (2\sqrt{m} + 1)\binom{2m}{m}2^{-2m}$$

[4] Chebyshev inequality = Čebyševova nerovnost

and the proposition follows. □

5.3 Threshold Functions

Now we return to random graphs and we consider the following question: What is the probability that $G(n,p)$ contains a triangle? Note that this is a *monotone property*; that means, if it holds for a graph G and $G \subset H$, it holds for H as well. It is natural to expect that for very small p, $G(n,p)$ is almost surely triangle-free, whereas for large p, the appearance of a triangle is very likely.

Let T denote the number of triangles in $G(n,p)$. For a given triple of vertices, the probability that they form a triangle is p^3. By linearity of expectation, the expected number of triangles is

$$\mathbf{E}[T] = \binom{n}{3} p^3$$

which approaches zero if $p(n) << \frac{1}{n}$ (the notation $f(n) << g(n)$ is equivalent to $f(n) = o(g(n))$ and $f(n) >> g(n)$ means $g(n) = o(f(n))$). Therefore, the probability that $G(n,p(n))$ contains a triangle tends to zero for $p(n) = o(\frac{1}{n})$.

On the other hand, let us suppose that $p(n) >> \frac{1}{n}$. Then the expected number of triangles goes to infinity with increasing n, yet this *does not* imply that $G(n,p)$ contains a triangle almost surely! It might be the case that there are a few graphs abounding with triangles (and boosting the expected value) while with a large probability the number of triangles is zero. This can also be illustrated with the following real-life scenario.

Example: fire insurance. The annual cost of insurance against fire, per household, is increasing. This reflects the growing damage inflicted by fire every year to an average household. But does this mean that the probability of a fire accident is rising, or even that in the limit, *almost every* household will be stricken by fire every year? Hardly. The rise in the expected damage costs is due to a few fire accidents every year which, however, are getting more and more expensive.

Fortunately, our triangles do not behave as erratically as fire accidents. Most random graphs have a "typical" number of triangles which is relatively close to the expectation. It is exactly the second moment method that allows us to capture this property and prove that if the expected number of triangles is sufficiently large, the random graph contains *some* triangle almost surely.

5.3.1 Lemma. *Consider a sequence $X_1, X_2, \ldots$ of non-negative random variables such that*

$$\lim_{n\to\infty} \frac{\mathrm{Var}[X_n]}{(\mathbf{E}[X_n])^2} = 0.$$

Then

$$\lim_{n\to\infty} \mathrm{P}[X_n > 0] = 1.$$

Proof. We choose $t = \mathbf{E}[X_n]$ in the Chebyshev inequality:

$$\mathrm{P}[|X_n - \mathbf{E}[X_n]| \geq \mathbf{E}[X_n]] \leq \frac{\mathrm{Var}[X_n]}{(\mathbf{E}[X_n])^2}$$

and we get

$$\lim_{n\to\infty} \mathrm{P}[X_n \leq 0] \leq \lim_{n\to\infty} \frac{\mathrm{Var}[X_n]}{(\mathbf{E}[X_n])^2} = 0.$$

□

Thus we need to estimate the variance of the number of triangles in $G(n,p)$. We have $T = \sum T_i$ where $T_1, T_2, \ldots$ are indicator variables for all the $\binom{n}{3}$ possible triangles in $G(n,p)$. The variance of a sum of random variables is

$$\mathrm{Var}[T] = \sum_i \mathrm{Var}[T_i] + \sum_{i\neq j} \mathrm{Cov}[T_i, T_j].$$

For every triangle

$$\mathrm{Var}[T_i] \leq \mathbf{E}\left[T_i^2\right] = p^3,$$

and for a pair of triangles sharing an edge

$$\mathrm{Cov}[T_i, T_j] \leq \mathbf{E}[T_i T_j] = p^5,$$

since $T_i T_j$ is the indicator variable of the appearance of 5 fixed edges.

The indicator variables corresponding to edge-disjoint triangles are independent and then the covariance is zero. So we only sum up over the pairs of triangles sharing an edge; the number of such (ordered) pairs is $12\binom{n}{4}$. In total, we get

$$\mathrm{Var}[T] \leq \binom{n}{3} p^3 + 12\binom{n}{4} p^5 \leq n^3 p^3 + n^4 p^5$$

$$\frac{\mathrm{Var}[T]}{(\mathbf{E}[T])^2} \leq \frac{n^3p^3 + n^4p^5}{(\binom{n}{3}p^3)^2} = O\left(\frac{1}{n^3p^3} + \frac{1}{n^2p}\right),$$

which tends to zero if $p(n) >> \frac{1}{n}$. Lemma 5.3.1 implies that the probability that $G(n,p)$ contains a triangle approaches 1 as $n \to \infty$.

As the reader can observe, the transition between random graphs that contain a triangle almost never or almost always is quite sharp. In order to describe this phenomenon more generally, Erdős and Rényi introduced the notion of a *threshold function.*

5.3.2 Definition. *A function* $r\colon \mathbf{N} \to \mathbf{R}$ *is a* **threshold function** *for a monotone graph property* A, *if for any* $p\colon \mathbf{N} \to [0,1]$

- $p(n) = o(r(n)) \Rightarrow \lim_{n\to\infty} \mathrm{P}[A \text{ holds for } G(n,p(n))] = 0$
- $r(n) = o(p(n)) \Rightarrow \lim_{n\to\infty} \mathrm{P}[A \text{ holds for } G(n,p(n))] = 1$

(a property A *is* **monotone** *if for any two graphs* G *and* H *with* $V(H) = V(G)$, $E(H) \subseteq E(G)$, *and* H *having property* A, G *has property* A *as well).*

Note that a threshold function may not exist, and if it exists, it is not unique. For our property "$G(n,p)$ contains a triangle", the threshold function is $r(n) = \frac{1}{n}$, but $r(n) = \frac{c}{n}$ (for any $c > 0$) could serve as well.

More generally, we can study the threshold functions for the appearance of other subgraphs (not necessarily induced; the question of induced subgraphs is much more difficult). It turns out that our approach can be extended to any subgraph H that is *balanced.*

5.3.3 Definition. *Let H be a graph with v vertices and e edges. We define the* **density** *of H as*

$$\rho(H) = \frac{e}{v}.$$

We call H **balanced** *if no subgraph of H has strictly greater density than H itself.*

5.3.4 Theorem. *Let H be a balanced graph with density ρ. Then*

$$r(n) = n^{-1/\rho}$$

is a threshold function for the event that H is a subgraph of $G(n,p)$.

Proof. Let H have v vertices and e edges, $\rho = \frac{e}{v}$. Denote the vertices of H by $\{a_1, a_2, \ldots, a_v\}$. For any ordered v-tuple $\beta = (b_1, b_2, \ldots, b_v)$ of distinct vertices $b_1, \ldots, b_v \in V(G(n,p))$, let A_β denote the event that $G(n,p)$ contains an appropriately ordered copy of H on $(b_1, \ldots, b_v)$. That is, A_β occurs if $\{b_i, b_j\} \in E(G(n,p))$ whenever $\{a_i, a_j\} \in E(H)$; in other words, whenever the mapping $a_i \mapsto b_i$ is a graph homomorphism.

Let X_β denote the indicator variable corresponding to A_β and let $X = \sum_\beta X_\beta$ be the sum over all the ordered v-tuples β. Note that due to the possible symmetries of H, some copies of H may be counted repeatedly, and so X is not exactly the number of copies of H in $G(n,p)$. However, the conditions $X = 0$ and $X > 0$ are equivalent to the absence and appearance of H in $G(n,p)$.

The probability of A_β is p^e. By linearity of expectation,

$$\mathbf{E}[X] = \sum_\beta \mathrm{P}[A_\beta] = \Theta(n^v p^e)$$

(note that v and e are constants, while p is a function of n).

If $p(n) << n^{-v/e}$, then

$$\lim_{n\to\infty} \mathbf{E}[X] = 0,$$

which completes the first part of the proof.

Now assume $p(n) >> n^{-v/e}$ and apply the second moment method:

$$\mathrm{Var}[X] = \sum_\beta \mathrm{Var}[X_\beta] + \sum_{\beta \neq \gamma} \mathrm{Cov}[X_\beta, X_\gamma].$$

Note that $\mathrm{Var}[X_\beta] = \mathrm{Cov}[X_\beta, X_\beta]$, so we can also write

$$\mathrm{Var}[X] = \sum_{\beta,\gamma} \mathrm{Cov}[X_\beta, X_\gamma].$$

The covariances are non-zero only for the pairs of copies that share some edges. Let β and γ share $t \geq 2$ vertices; then the two copies of H have at most $t\rho$ edges in common (because H is balanced), and their union contains at least $2e - t\rho$ edges. Thus

$$\mathrm{Cov}\,[X_\beta, X_\gamma] \leq \mathbf{E}\,[X_\beta X_\gamma] \leq p^{2e-t\rho}.$$

The number of pairs β, γ sharing t vertices is $O(n^{2v-t})$, because we can choose a set of $2v-t$ vertices in $\binom{n}{2v-t}$ ways and there are only constantly many ways to choose β and γ from this set (since H is fixed and so its size of H is a constant). For a fixed t, we get

$$\sum_{|\beta\cap\gamma|=t} \mathrm{Cov}\,[X_\beta, X_\gamma] = O(n^{2v-t}\, p^{2e-t\rho}) = O((n^v p^e)^{2-t/v}).$$

For the variance of X, we get

$$\mathrm{Var}\,[X] = O\left(\sum_{t=2}^{v} (n^v p^e)^{2-t/v}\right)$$

and

$$\lim_{n\to\infty} \frac{\mathrm{Var}\,[X]}{(\mathbf{E}\,[X])^2} = \lim_{n\to\infty} O\left(\sum_{t=2}^{v} (n^v p^e)^{-t/v}\right) = 0$$

since $\lim_{n\to\infty} n^v p^e = \infty$. This completes the second part of the proof because by Lemma 5.3.1,

$$\lim_{n\to\infty} \mathrm{P}\,[X > 0] = 1$$

and there is almost always a copy of H in $G(n, p)$. □

The question of a general subgraph H was solved by Erdős and Rényi: The threshold function for H is determined by the subgraph $H' \subset H$ with the maximum density. We give here only the result without a proof.

5.3.5 Theorem. *Let H be a graph and $H' \subset H$ a subgraph of H with the maximum density. Then*

$$r(n) = n^{-1/\rho(H')}$$

is a threshold function for the event that H is a subgraph of $G(n, p)$.

5.4 The Clique Number

Now we consider the clique number of a random graph. For simplicity, suppose that the probability of each edge is $p = \frac{1}{2}$. Let us choose a number k and count the number of cliques of size k. For each set S of k vertices, let X_S denote the indicator variable of the event "S is a clique". Then $X = \sum_{|S|=k} X_S$ is the number of k-cliques in the graph. The expected number of k-cliques is

$$\mathbf{E}\,[X] = \sum_{|S|=k} \mathbf{E}\,[X_S] = \binom{n}{k} 2^{-\binom{k}{2}}.$$

This function drops below 1 approximately at $k = 2\log_2 n$ and, indeed, this is the typical size of the largest clique in $G(n, 1/2)$.

5.4.1 Lemma.
$$\lim_{n\to\infty} \mathrm{P}\left[\omega(G(n,1/2)) > 2\log_2 n\right] = 0.$$

Proof. We set $k(n) = \lceil 2\log_2 n \rceil$ and calculate the average number of cliques of this size:
$$\mathbf{E}[X] = \binom{n}{k} 2^{-\binom{k}{2}} \leq \frac{(2^{k/2})^k}{k!} 2^{-k(k-1)/2} = \frac{2^{k/2}}{k!}$$
which tends to 0 as $n \to \infty$. Therefore
$$\lim_{n\to\infty} \mathrm{P}\left[\omega(G(n,1/2)) > 2\log_2 n\right] = 0.$$
□

However, it is more challenging to argue that there will almost always be a clique of size near the threshold of $2\log_2 n$. We prove the following result.

5.4.2 Theorem. *Let $k(n)$ be a function such that*
$$\lim_{n\to\infty} \binom{n}{k(n)} 2^{-\binom{k(n)}{2}} = \infty.$$
Then
$$\lim_{n\to\infty} \mathrm{P}\left[\omega(G(n,1/2)) \geq k(n)\right] = 1.$$

Proof. Here the calculations are somewhat more demanding than usual. For brevity, let us write $E(n,k) = \binom{n}{k}2^{-\binom{k}{2}}$. First we note that we may assume n to be sufficiently large and
$$\tfrac{3}{2}\log_2 n \leq k < 2\log_2 n$$
(where $\frac{3}{2}$ can be replaced by any constant smaller than 2). As for the second inequality, we already know that $E(n, 2\log_2 n) \to 0$. For the first inequality, we have $\log_2 E(n,k) \geq \log_2\left[(\frac{n}{k})^k 2^{-k^2/2}\right] = k\log_2 n - k\log_2 k - \frac{k^2}{2}$, and so $\log_2 E(n, \frac{3}{2}\log_2 n) \geq \frac{3}{2}\log_2^2 n - o(\log^2 n) + \frac{9}{8}\log_2^2 n \to \infty$ as $n \to \infty$.

For convenience, we also suppose that $k = k(n)$ is even.

Let $X = \sum_{|S|=k(n)} X_S$ denote the number of cliques of size $k(n)$ in $G(n,1/2)$. The condition on $k(n)$ guarantees that $\lim_{n\to\infty} \mathbf{E}[X] = \infty$. It remains to estimate the variance of X:
$$\mathrm{Var}[X] = \sum_{|S|=|T|=k} \mathrm{Cov}[X_S, X_T]$$
(note that this includes the terms where $S = T$, which are equal to $\mathrm{Var}[X_T]$).

The variables X_S, X_T are independent whenever S and T share at most one vertex (and therefore the corresponding cliques have no edges in common). So we are interested only in those pairs S, T with $|S \cap T| \geq 2$, and we can write
$$\mathrm{Var}[X] = \sum_{t=2}^{k} C(t),$$

where

$$C(t) = \sum_{|S \cap T| = t} \mathrm{Cov}\,[X_S, X_T].$$

For a fixed $t = |S \cap T|$, the cliques on S and T have $2\binom{k}{2} - \binom{t}{2}$ edges in total, so we have

$$\mathrm{Cov}\,[X_S, X_T] \le \mathbf{E}\,[X_S X_T] = 2^{\binom{t}{2} - 2\binom{k}{2}}$$

and since a pair of subsets (S, T) with $|S| = |T| = k$ and $|S \cap T| = t$ can be chosen in $\binom{n}{k}\binom{k}{t}\binom{n-k}{k-t}$ ways,

$$C(t) \le \binom{n}{k}\binom{k}{t}\binom{n-k}{k-t} 2^{\binom{t}{2} - 2\binom{k}{2}}.$$

We need to prove that

$$\frac{\mathrm{Var}\,[X]}{(\mathbf{E}\,[X])^2} = \sum_{t=2}^{k} \frac{C(t)}{(\mathbf{E}\,[X])^2} \to 0$$

(see Lemma 5.3.1). We split the sum over t into two ranges.

In the *first range*, $2 \le t \le \frac{k}{2}$, we show that the sum goes to 0 for $k < 2\log_2 n$. When dealing with a product of several binomial coefficients, it is often a good idea to expand them, as many terms usually cancel out or can be matched conveniently. We have

$$\begin{aligned}
\frac{C(t)}{(\mathbf{E}\,[X])^2} &\le \frac{\binom{k}{t}\binom{n-k}{k-t}}{\binom{n}{k}} 2^{\binom{t}{2}} \\
&\le \frac{k^t}{t!} \cdot \frac{(n-k)(n-k-1)\cdots(n-2k+t+1)}{(k-t)!} \cdot \frac{k!}{n(n-1)\cdots(n-k+1)} \cdot 2^{\binom{t}{2}} \\
&\le k^{2t} \cdot \frac{1}{n(n-1)\cdots(n-t+1)\cdot t!} \cdot 2^{t^2/2} \le k^{2t} n^{-t} 2^{t^2/2} \\
&\le k^{2t}(2^{-k/2})^t 2^{t^2/2} \le (k^2 2^{-k/2} 2^{t/2})^t.
\end{aligned}$$

Since $t \le \frac{k}{2}$, the expression in parentheses is at most $k^2 2^{-k/4} = o(1)$. We can thus bound $\sum_{t=2}^{k/2} C(t)/(\mathbf{E}\,[X])^2$ by the sum of the geometric series, $\sum_{t=2}^{\infty} q^t$, with $q = k^2 2^{-k/4} = o(1)$ and so the sum tends to 0.

For the *second range*, $\frac{k}{2} < t \le k$, we show that $\sum_{t=k/2}^{k} C(t)/\mathbf{E}\,[X] = o(1)$ for $k \ge \frac{3}{2}\log_2 n$. Consequently, since $\mathbf{E}\,[X] \to \infty$ by the condition in the theorem, we have $\sum_{t=k/2}^{k} C(t)/(\mathbf{E}\,[X])^2 \to 0$ as well. This time we can afford to bound the binomial coefficients quite roughly:

$$\begin{aligned}
\frac{C(t)}{\mathbf{E}\,[X]} &\le \binom{k}{t}\binom{n-k}{k-t} 2^{\binom{t}{2} - \binom{k}{2}} \le \binom{k}{k-t}\binom{n}{k-t} 2^{\binom{t}{2} - \binom{k}{2}} \\
&\le k^{k-t} n^{k-t}\, 2^{(t^2 - k^2 - t + k)/2} \\
&= (kn)^{k-t}\, 2^{-(k-t)(k+t-1)/2} = (kn 2^{-(k+t-1)/2})^{k-t} \\
&\le (2^{\log_2 k + (2/3)k - (k+t-1)/2})^{k-t} \\
&\le (2^{\log_2 k + (2/3)k - (3/4)k})^{k-t}
\end{aligned}$$

as $t > \frac{k}{2}$. The expression in parentheses is $o(1)$. Bounding by a geometric series again, it follows that $\sum_{t=k/2}^{k} C(t)/\mathbf{E}[X] \to 0$ as claimed. Altogether we have proved $\lim_{n\to\infty} \mathrm{Var}[X]/(\mathbf{E}[X])^2 = 0$. □

Remark. If we choose $k(n) = (2-\varepsilon)\log_2 n$, the condition of the theorem holds for any $\varepsilon > 0$. This means that the clique number $\omega(G(n,1/2))$ almost always lies between $(2-\varepsilon)\log_2 n$ and $2\log_2 n$. However, the concentration of the clique number is even stronger. In 1976, Bollobás, Erdős and Matula proved that there exists a function $k(n)$ such that

$$\lim_{n\to\infty} \mathrm{P}[k(n) \leq \omega(G(n,1/2)) \leq k(n)+1] = 1.$$

6

The Lovász Local Lemma

6.1 Statement and Proof

The typical goal of the probabilistic method is to prove that the probability that nothing "bad" happens is greater than zero. Usually, we have a collection of bad events $A_1, A_2, \ldots, A_n$ that we are trying to avoid. (These may be, for example, the occurrences of a monochromatic edge in a hypergraph, as in Theorem 2.2.4.) If the sum of their probabilities $\sum \mathrm{P}[A_i]$ is strictly less than 1, then clearly there is a positive probability that none of them occurs. However, in many cases this approach is not powerful enough, because the sum of probabilities of the bad events $\sum \mathrm{P}[A_i]$ may be substantially larger than the probability of their union $\mathrm{P}[\bigcup A_i]$.

One case where we can do better is when the events $A_1, \ldots, A_n$ are *independent* (and non-trivial). Then their complements are independent as well, and we have

$$\mathrm{P}\left[\overline{A_1} \cap \overline{A_2} \cap \ldots \cap \overline{A_n}\right] = \mathrm{P}\left[\overline{A_1}\right] \mathrm{P}\left[\overline{A_2}\right] \cdots \mathrm{P}\left[\overline{A_n}\right] > 0$$

even though the probabilities $\mathrm{P}[A_i]$ can be very close to 1 and their sum can be arbitrarily large.

It is natural to expect that something similar holds even if the events are not entirely independent. The following definitions conveniently express "limited dependence" of events using a directed graph.

6.1.1 Definition. *An event A is* **independent of events** *$B_1, \ldots, B_k$ if for any nonempty $J \subseteq [k]$,*

$$\mathrm{P}\left[A \cap \bigcap_{j \in J} B_j\right] = \mathrm{P}[A]\, \mathrm{P}\left[\bigcap_{j \in J} B_j\right].$$

6.1.2 Definition. *Let $A_1, A_2, \ldots, A_n$ be events in a probability space. A directed graph $D = (V, E)$ with $V = [n]$ is a* **dependency digraph** *for $A_1, \ldots, A_n$ if each event A_i is independent of all the events A_j with $(i, j) \notin E$.*

Note that a dependency digraph need not be determined uniquely.

The local lemma, discovered by Lovász, is a powerful tool which allows us to exclude all bad events, provided that their probabilities are relatively small and their dependency digraph does not have too many edges. We begin with a simple symmetric form of the local lemma, the one used most often.

6.1.3 Lemma (Symmetric Lovász Local Lemma). *Let $A_1, \dots, A_n$ be events such that $\mathrm{P}[A_i] \le p$ for all i and all outdegrees in a dependency digraph of the A_i are at most d; that is, each A_i is independent of all but at most d of the other A_j. If $ep(d+1) \le 1$ (where $e = 2.71828\dots$ is the basis of natural logarithms), then*

$$\mathrm{P}\Big[\bigcap_{i=1}^{n} \overline{A_i}\Big] > 0.$$

If some of the events A_i have probability considerably larger than the others, then the following general version can be useful:

6.1.4 Lemma (Lovász Local Lemma). *Let $A_1, A_2, \dots, A_n$ be events, $D = (V, E)$ their dependency digraph, and $x_i \in [0,1)$ real numbers assigned to the events, in such a way that*

$$\mathrm{P}[A_i] \le x_i \prod_{(i,j)\in E} (1 - x_j).$$

Then

$$\mathrm{P}\Big[\bigcap_{i=1}^{n} \overline{A_i}\Big] \ge \prod_{i=1}^{n} (1 - x_i) > 0.$$

If all the $\mathrm{P}[A_i]$ are below $\frac{1}{6}$, say, then a good choice in applications is usually $x_i = 3\mathrm{P}[A_i]$ (the exact value 3 is not important). Then it is easy to show that if $\sum_{j:(i,j)\in E} \mathrm{P}[A_j] \le \frac{1}{6}$ for all i, and the assumptions of the Lovász Local Lemma hold.

In the rest of the section, we prove both versions of the local lemma. It seems that at first reading, the proof does not give much insight why the lemma holds. The reader not particularly interested in the proof may safely continue with the examples in the next sections and perhaps return to the proof later.

Proof of Lemma 6.1.4. The complementary events $\overline{A_i}$ have positive probabilities but we want them all to occur simultaneously. This would be impossible if the occurrence of a combination of $\overline{A_j}$ forced some other A_i to hold. Therefore, we need to bound the probability of A_i on the condition of the other events *not occurring*, and this is where the parameters x_i come into play. First we prove that for any subset $S \subset \{1, \dots, n\}$ and $i \notin S$

$$\mathrm{P}\Big[A_i \,\Big|\, \bigcap_{j\in S} \overline{A_j}\Big] \le x_i.$$

We proceed by induction on the size of S. For $S = \emptyset$, the statement follows directly from the assumption of the lemma:

$$\mathrm{P}[A_i] \le x_i \prod_{(i,j)\in E} (1 - x_j) \le x_i.$$

Now suppose it holds for any $S', |S'| < |S|$ and set $S_1 = \{j \in S\colon (i,j) \in E\}, S_2 = S \setminus S_1$. We can assume $S_1 \neq \emptyset$, for otherwise, A_i is independent of $\bigcap_{j\in S} \overline{A_j}$ and the statement follows trivially. We have

$$\mathrm{P}\left[A_i \,\middle|\, \bigcap_{j\in S} \overline{A_j} \right] = \frac{\mathrm{P}\left[A_i \cap \bigcap_{j\in S_1} \overline{A_j} \,\middle|\, \bigcap_{l\in S_2} \overline{A_l}\right]}{\mathrm{P}\left[\bigcap_{j\in S_1} \overline{A_j} \,\middle|\, \bigcap_{l\in S_2} \overline{A_l}\right]}$$

Since A_i is independent of the events $\{A_l\colon l \in S_2\}$, we can bound the numerator as follows:

$$\mathrm{P}\left[A_i \cap \bigcap_{j\in S_1} \overline{A_j} \,\middle|\, \bigcap_{l\in S_2} \overline{A_l} \right] \leq \mathrm{P}\left[A_i \,\middle|\, \bigcap_{l\in S_2} \overline{A_l} \right] = \mathrm{P}[A_i] \leq x_i \prod_{(i,j)\in E} (1 - x_j).$$

To bound the denominator, suppose $S_1 = \{j_1, \ldots, j_r\}$ and use the induction hypothesis:

$$\begin{aligned}
\mathrm{P}\left[\overline{A_{j_1}} \cap \cdots \cap \overline{A_{j_r}} \,\middle|\, \bigcap_{l\in S_2} \overline{A_l}\right] &= \mathrm{P}\left[\overline{A_{j_1}} \,\middle|\, \bigcap_{l\in S_2} \overline{A_l}\right] \mathrm{P}\left[\overline{A_{j_2}} \,\middle|\, \overline{A_{j_1}} \cap \bigcap_{l\in S_2} \overline{A_l}\right] \\
&\quad \cdots \times \mathrm{P}\left[\overline{A_{j_r}} \,\middle|\, \overline{A_{j_1}} \cap \cdots \cap \overline{A_{j_{r-1}}} \cap \bigcap_{l\in S_2} \overline{A_l}\right] \\
&\geq (1 - x_{j_1})(1 - x_{j_2}) \cdots (1 - x_{j_r}) \\
&\geq \prod_{(i,j)\in E} (1 - x_j).
\end{aligned}$$

We conclude that $\mathrm{P}\left[A_i | \bigcap_{j\in S} \overline{A_j}\right] \leq x_i$ and now the lemma follows easily, because

$$\mathrm{P}\left[\bigcap_{i=1}^{n} \overline{A_i}\right] = \mathrm{P}\left[\overline{A_1}\right] \mathrm{P}\left[\overline{A_2} \,\middle|\, \overline{A_1}\right] \cdots \mathrm{P}\left[\overline{A_n} \,\middle|\, \overline{A_1} \cap \cdots \cap \overline{A_{n-1}}\right] \geq \prod_{i=1}^{n} (1 - x_i).$$

□

Proof of the symmetric version (Lemma 6.1.3). For $d = 0$ the events are mutually independent and the result follows easily. Otherwise set $x_i = \frac{1}{d+1} < 1$. In the dependency digraph, the outdegree of any vertex is at most d, so

$$x_i \prod_{(i,j)\in E} (1 - x_j) \geq \frac{1}{d+1} \left(1 - \frac{1}{d+1}\right)^d \geq \frac{1}{e(d+1)} \geq p$$

and we can apply the general local lemma. □

Algorithmic remark. In the basic probabilistic method, we usually prove that almost all of the considered objects are good. So if we want to find a good object, we can select an object at random, and we have a very good chance of selecting a good one (of course, *verifying* that an object is good can still be difficult, but this is another matter). In contrast, the Lovász Local Lemma guarantees that the probability of avoiding all bad events is positive, but this probability is typically very small! For example, if $A_1, \ldots, A_n$ are independent events, with probability

$\frac{1}{3}$ each, say, in which case the Local Lemma applies, then the probability of none A_i occurring is only $(\frac{2}{3})^n$. So good objects guaranteed by the Local Lemma can be extremely rare. Nevertheless, algorithmic versions of the Local Lemma, where a good object can be found efficiently, are known; the first one, for a particular application, was discovered by Beck, and for quite general recent results the reader may consult

> M. Molloy, B. Reed: Further algorithmic aspects of the Local Lemma, *Proc. of the 30th ACM Symposium of Theory of Computing*, 1998, pages 524–530.

Now we present several combinatorial results which can be obtained with the help of the Local Lemma.

6.2 Hypergraph Coloring Again

In section 2.2, we proved that any k-uniform hypergraph with less than 2^{k-1} edges is 2-colorable. By applying the Local Lemma, we prove a similar result which holds for a hypergraph with arbitrarily many edges provided that they do not intersect too much.

6.2.1 Theorem. *Let $\mathcal{H}$ be a hypergraph in which every edge has at least k vertices and intersects at most d other edges. If $e(d+1) \leq 2^{k-1}$, then $\mathcal{H}$ is 2-colorable.*

Proof. Let us color the vertices of $\mathcal{H}$ independently red or blue, with probability $\frac{1}{2}$. For every edge f, let A_f denote the event that f is monochromatic. As any edge has at least k elements, the probability of A_f is at most $p = 2^{1-k}$. Clearly, the event A_f is independent of all A_g but those (at most d) events where f intersects g. Since $ep(d+1) \leq 1$, we can use the Local Lemma, which implies that there is a non-zero probability that no edge is monochromatic. □

6.3 Directed Cycles

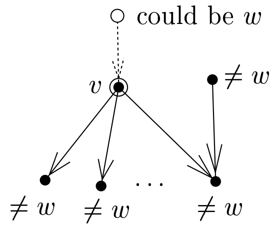
6.3.1 Theorem. *Let $D = (V, E)$ be a directed graph with minimum outdegree δ and maximum indegree Δ. Then for any $k \in \mathbf{N}$ such that*

$$k \leq \frac{\delta}{1 + \ln(1 + \delta\Delta)},$$

D contains a directed cycle of length divisible by k.

Proof. Consider a subgraph of D where every outdegree is exactly δ. Let $f\colon V \to \{0, 1, \ldots, k-1\}$ be a random coloring obtained by choosing $f(v)$ for each $v \in V$ independently and uniformly. Let $N^+(v)$ denote the set of vertices $\{w\colon (v, w) \in E\}$ and A_v the event that no vertex in $N^+(v)$ is colored by $f(v) + 1 \pmod{k}$.

The probability of A_v is $p = (1 - \frac{1}{k})^\delta$. We claim that each A_v is independent of all the events A_w with $N^+(v) \cap (N^+(w) \cup \{w\}) = \emptyset$. That is, w is not a successor of v and w and v have no common successor:



Note that v may be a successor of w (as indicated by the dashed arrow). In this case, the independence is not so obvious, but it still holds: Even if the color is fixed for all vertices except for $N^+(v)$ and it is chosen randomly on $N^+(v)$, the probability of A_v is still $(1-\frac{1}{k})^\delta$.

The number d of vertices w not satisfying the above conditions is at most $\delta + \delta(\Delta-1) = \delta\Delta$. Hence

$$ep(d+1) \leq e(1-\frac{1}{k})^\delta(\delta\Delta+1) \leq e^{1-\delta/k}(\delta\Delta+1) \leq 1,$$

and by the Local Lemma, there is a coloring such that for every $v \in V$, there is a $w \in N^+(v)$ such that $f(w) = f(v)+1 \pmod k$. Now starting at any vertex v_0, we can generate a sequence of vertices $v_0, v_1, v_2, \ldots$ such that $(v_i, v_{i+1}) \in E$ and $f(v_{i+1}) = f(v_i) + 1 \pmod k$, until we find a directed cycle in D. The coloring scheme guarantees that the length of the cycle is divisible by k. $\square$

6.4 Ridiculous Injections

This is a silly example which, nonetheless, shows how strong the Local Lemma is, compared to an elementary probabilistic argument. Let us consider two finite sets M and N; $|M| = m, |N| = n$. We will attempt to prove by the probabilistic method that under favorable circumstances, there exists an injective mapping from M to N. The first result is based only on elementary probabilistic reasoning, and it is also relatively weak. :–)

6.4.1 Theorem. *If $n > \binom{m}{2}$, then an injective mapping $f\colon M \to N$ exists.*

Proof. Consider a random mapping $f\colon M \to N$, where the image of each element of M is chosen from N at random, uniformly and independently. Let A_{xy} denote the event that, for $x, y \in M$, $f(x) = f(y)$. The probability of A_{xy} is $p = \frac{1}{n}$. Since there are $\binom{m}{2}$ such events A_{xy} that must be avoided in order for f to be injective, we have

$$\mathrm{P}\Big[\bigcap_{x,y\in M} \overline{A_{xy}}\Big] \geq 1 - \binom{m}{2}\frac{1}{n} > 0$$

and therefore an injective mapping exists. $\square$

Now, with the Local Lemma at hand, we are ready for a substantial improvement. Instead of $n > \binom{m}{2}$, we will need only a linear number of elements!

6.4.2 Theorem. *If $n > 6m$, then an injective mapping $f\colon M \to N$ exists.*

Proof. Again, we define the events A_{xy} for $x \neq y$ as $f(x) = f(y)$ and we observe that $p = \mathrm{P}[A_{xy}] < \frac{1}{6m}$ and A_{xy} is independent of all but the $d < 2m$ events $A_{x'y'}$ with $\{x, y\} \cap \{x', y'\} \neq \emptyset$. So we have $ep(d+1) < 1$ and the Local Lemma says that

$$\mathrm{P}\Big[\bigcap_{x,y \in M} \overline{A_{xy}}\Big] > 0.$$

□

6.5 Coloring of Real Numbers

This is a problem which appeared in the original paper containing the Local Lemma by Erdős and Lovász. They asked whether it is possible, for a given finite set $S \subset \mathbf{R}$, to color the real numbers with k colors in such a way that every translation (shifted copy) of S contains all the k colors.

6.5.1 Definition. *Let $c\colon \mathbf{R} \to [k]$ be a coloring of the real numbers. A set $T \subset \mathbf{R}$ is called* **colorful** *if $c(T) = [k]$.*

6.5.2 Theorem. *For any k there is m such that for any m-point set $S \subset \mathbf{R}$, the real numbers can be colored with k colors so that any translation of S is colorful.*

Proof. First, we prove a result about finite sets of translates.

> *Statement F: For any k, there exists $m = m(k)$ such that for any m-point $S \subset \mathbf{R}$ and finite $X \subset \mathbf{R}$, there is a coloring c of the set $T = \bigcup_{x \in X}(S + x)$ with k colors under which each translation $S + x$ with $x \in X$ is colorful.*

Let $c\colon T \to [k]$ be a random coloring obtained by choosing $c(y)$ for each $y \in T$ independently and uniformly at random. For each $x \in X$, let A_x denote the event that $c(S + x)$ does not contain all the k colors. The probability of A_x is at most $p = k(1 - \frac{1}{k})^m$. Moreover, each A_x is independent of all the other events but those $A_{x'}$ with $(S + x) \cap (S + x') \neq \emptyset$. The number of such events is at most $d = m(m - 1)$. If we choose m sufficiently large so that

$$ep(d+1) = ek\left(1 - \tfrac{1}{k}\right)^m (m(m-1) + 1) \leq 1,$$

then the Local Lemma implies that there is a coloring such that all the sets $S + x$, $x \in X$, are colorful. Statement F is proved.

Here it should be noted that the Local Lemma itself cannot take us any further, because it requires that the number of events in question is finite. The proper coloring of all real numbers can be obtained by a compactness argument (which requires the axiom of choice).

First, we will show a weaker result by an elementary argument. (This weaker result is included just for illustration and it is not needed in the proof of Theorem 6.5.2 that will be presented later.) Let $Q = \{q_1, q_2, q_3, \ldots\} \subset \mathbf{R}$

be a countable set, for example the rationals. We are going to color the set $T = \bigcup_{q\in Q}(S+q)$. Let $T_i = \bigcup_{j=1}^{i}(S+q_j)$. For every T_i, using Statement F above, we fix a coloring $c_i: T_i \to [k]$ such that all the sets $S+q_j$, $j \le i$, are colorful. We are going to define a coloring $c: T \to [k]$ by a diagonal argument.

There are finitely many ways of coloring the set $S+q_1$, and we have the infinite sequence $(c_1, c_2, \ldots)$ of colorings, so there is an infinite subsequence $(c_{i_1}, c_{i_2}, \ldots)$ all of whose colorings coincide on $S+q_1$ (and $S+q_1$ is colorful under them). For simpler notation, let us write $c_j^{(1)} = c_{i_j}$, so we have the infinite sequence $(c_1^{(1)}, c_2^{(1)}, c_3^{(1)}, \ldots)$. All of these colorings, except possibly for $c_1^{(1)}$, are defined on $S+q_2$, and can have only finitely many patterns there, so we can select an infinite subsequence $(c_1^{(2)}, c_2^{(2)}, c_3^{(2)}, \ldots)$, all of whose colorings coincide on $S+q_2$. Continuing in this manner, after ℓ steps, we get an infinite sequence $(c_1^{(\ell)}, c_2^{(\ell)}, \ldots)$ whose colorings coincide on $T_\ell = \bigcup_{i=1}^{\ell}(S+q_i)$ and such that each $S+q_i$, $i = 1, 2, \ldots, \ell$ is colorful. Note that the coloring of T_ℓ remains fixed after the ℓth step, and each $c_j^{(r)}$, $r \ge \ell$, coincides with $c_1^{(\ell)}$ on T_ℓ.

Now we define a "diagonal" coloring $c: T \to [k]$ by letting $c(x) = c_1^{(\ell)}(x)$, where ℓ is the smallest index such that $x \in T_\ell$. Note that we also have $c(x) = c_1^{(r)}$ for all r such that $x \in T_r$. Since each $S+q_r$ is colorful under $c_1^{(r)}$ by the construction, it follows that it is colorful under c as well.

Finally, we prove the existence of the desired coloring of the real numbers. We need to recall two facts about compact topological spaces. First, if $\mathcal{C}$ is a system of closed subsets in a compact space such that $\bigcap_{C\in\mathcal{F}} C \ne \emptyset$ for any finite subsystem $\mathcal{F} \subseteq \mathcal{C}$, then $\bigcap_{C\in\mathcal{C}} C \ne \emptyset$. And second, an arbitrary Cartesian product of compact topological spaces is compact (Tychonoff's theorem),[1] and in particular, the space M of all mappings $f: \mathbf{R} \to [k]$ is compact. The topology on this space is that of the Cartesian power $[k]^{\mathbf{R}}$; explicitly, any set of mappings of the form

$$\{f \in M\colon f(i) = g(i) \text{ for all } i \in I\}, \tag{6.1}$$

where $I \subset \mathbf{R}$ is finite and $g: I \to [k]$ is arbitrary, is closed in M.

Coming back to our coloring problem, let $C_x \subset M$ denote the set of all colorings for which $S+x$ is colorful. Each C_x is a finite union of sets of the form (6.1) and so it is closed in M. Statement F implies that for any finite set $X \subset \mathbf{R}$, $\bigcap_{x\in X} C_x \ne \emptyset$. From the compactness of M, we obtain the existence of a $c \in \bigcap_{x\in\mathbf{R}} C_x$, and such a coloring c makes all the sets $S+x$ $(x \in \mathbf{R})$ colorful. $\square$

[1] Tychonoff's theorem = Tichonovova věta (čte se s Ť)

7

Strong Concentration Around the Expectation

What is typically the maximum degree of the random graph $G(n, \frac{1}{2})$? This maximum degree is a quite complicated random variable, and it is not even clear how to compute its expectation. For each vertex, the expected degree is $d = \frac{1}{2}(n-1)$, but this alone does not tell us much about the maximum over all vertices. But suppose that we can show, for some suitable number t much smaller than n, that the degree of any given vertex exceeds $d+t$ with probability smaller than n^{-2}, say (as we will see later, the appropriate value of t is about $const \cdot \sqrt{n \log n}$). Then we can conclude that the maximum degree is below $d+t$ with probability at least $1 - \frac{1}{n}$, i.e. almost always.

In this case, and in many other applications of the probabilistic method, we need to bound probabilities of the form $\mathrm{P}[X \geq \mathbf{E}[X] + t]$ for some random variable X (and usually also probabilities of negative deviations from the expectation, i.e. $\mathrm{P}[X \leq \mathbf{E}[X] - t]$). Bounds for these probabilities are called *tail estimates*.[1] In other words, we want to show that X almost always lives in the interval $(\mathbf{E}[X] - t, \mathbf{E}[X] + t)$; we say that X *is concentrated* around its expectation.

The Chebyshev inequality is a very general result of this type, but usually it is too weak, especially if we need to deal with many random variables simultaneously. It tells us that

$$\mathrm{P}[|X - \mathbf{E}[X]| \geq \lambda\sigma] \leq \lambda^{-2},$$

where $\sigma = \sqrt{\mathrm{Var}[X]}$ and $\lambda \geq 0$ is a real parameter. If X is the degree of a fixed vertex in $G(n, \frac{1}{2})$, we have $\sigma = \frac{1}{2}\sqrt{n-1}$. Since the largest deviations we may ever want to consider in this case are smaller than $\frac{1}{2}(n-1)$, λ^{-2} is never below $\frac{1}{n}$, and the Chebyshev inequality is useless for the above consideration of the maximum degree. But as we will see below, for our particular X, a much better inequality holds, with λ^{-2} replaced by the exponentially small bound $2e^{-\lambda^2/2}$. This is already sufficient to conclude that, for example, the maximum degree of $G(n, \frac{1}{2})$ almost never exceeds $\frac{n}{2} + O(\sqrt{n \log n})$.

---

[1] tail estimate = odhad pravděpodobnosti velkých odchylek

7.1 Sum of Independent Uniform ± 1 Variables

We will start with the simplest result about strong concentration, which was mentioned in the above discussion of the maximum degree of $G(n, \frac{1}{2})$. We note that the degree of a given vertex v in $G(n, \frac{1}{2})$ is the sum of the indicators of the $n-1$ potential edges incident to v. Each of these indicators attains values 0 and 1, both with probability $\frac{1}{2}$, and they are all mutually independent.

For a more convenient notation in the proof, we will deal with sums of variables attaining values -1 and $+1$ instead of 0 and 1. One advantage is that the expectation is now 0. Results for the original setting can be recovered by a simple re-scaling.

7.1.1 Theorem. *Let $X_1, X_2, \ldots, X_n$ be independent random variables, each attaining the values $+1$ and -1, both with probability $\frac{1}{2}$. Let $X = X_1 + X_2 + \cdots + X_n$. Then we have, for any real $t \geq 0$,*

$$\mathrm{P}[X \geq t] < e^{-t^2/2\sigma^2} \quad \textit{and} \quad \mathrm{P}[X \leq -t] < e^{-t^2/2\sigma^2},$$

where $\sigma = \sqrt{\mathrm{Var}[X]} = \sqrt{n}$.

This estimate is often called Chernoff's[2] inequality in the literature (although Chernoff proved a more general and less handy inequality in 1958, and the above theorem goes back to Bernstein's paper from 1924).

Note that in this case, we can write down a formula for $\mathrm{P}[X \geq t]$, which will involve a sum of binomial coefficients. We could try to prove the inequality by estimating the binomial coefficients suitably. But we will use an ingenious trick from probability theory (due to Bernstein) which also works for sums of more general random variables, where explicit formulas are not available.

Proof. We prove only the first inequality; the second one follows by symmetry. The key steps are to consider the auxiliary random variable $Y = e^{uX}$, where $u > 0$ is a (yet undetermined) real parameter, and to apply Markov's inequality to Y.

We have $\mathrm{P}[X \geq t] = \mathrm{P}[Y \geq e^{ut}]$. By Markov's inequality, we obtain $\mathrm{P}[Y \geq q] \leq \mathbf{E}[Y]/q$. We calculate

$$\mathbf{E}[Y] = \mathbf{E}\left[e^{u(\sum_{i=1}^n X_i)}\right] = \mathbf{E}\left[\prod_{i=1}^n e^{uX_i}\right] = \prod_{i=1}^n \mathbf{E}\left[e^{uX_i}\right]$$

(by independence of the X_i)

$$= \left(\frac{e^u + e^{-u}}{2}\right)^n \leq e^{nu^2/2}.$$

The last estimate follows from the inequality $(e^x + e^{-x})/2 = \cosh x \leq e^{x^2/2}$ valid for all real x (this can be established by comparing the Taylor series of both sides). We obtain

$$\mathrm{P}\left[Y \geq e^{ut}\right] \leq \frac{\mathbf{E}[Y]}{e^{ut}} \leq e^{nu^2/2 - ut}.$$

[2] Chernoff = Černov

The last expression is minimized by setting $u = t/n$, which yields the value $e^{-t^2/2n} = e^{-t^2/2\sigma^2}$. Theorem 7.1.1 is proved. $\square$

Combinatorial discrepancy. We show a nice application. Let X be an n-point set, and let $\mathcal{F}$ be a system of subsets of X. We would like to color the points of X red and blue, in such a way that each set of $\mathcal{F}$ contains approximately the same number of red and blue points (we want a "balanced" coloring). The *discrepancy* of the set system $\mathcal{F}$ measures how well this can be done. We assign the value $+1$ to the red color and value -1 to the blue color, so that a coloring can be regarded as a mapping $\chi: X \to \{-1, +1\}$. Then the imbalance of a set $S \in \mathcal{F}$ is just $\chi(S) = \sum_{x \in S} \chi(x)$. The discrepancy $\mathrm{disc}(\mathcal{F}, \chi)$ of $\mathcal{F}$ under the coloring χ is $\max_{S \in \mathcal{F}} |\chi(S)|$, and the discrepancy of $\mathcal{F}$ is the minimum of $\mathrm{disc}(\mathcal{F}, \chi)$ over all χ.

If we take $\mathcal{F} = 2^X$ (all sets), then $\mathrm{disc}(\mathcal{F}) = \frac{n}{2}$. Using the Chernoff inequality, we show that the discrepancy is much smaller; namely, if the number of sets in $\mathcal{F}$ is not too large, then the discrepancy is not much larger than $\sqrt{n}$,

7.1.2 Proposition. *Let $|X| = n$ and $|\mathcal{F}| = m$. Then $\mathrm{disc}(\mathcal{F}) \le \sqrt{2n \ln(2m)}$. If the maximum size of a set in $\mathcal{F}$ is at most s, then $\mathrm{disc}(\mathcal{F}) \le \sqrt{2s \ln(2m)}$.*

Proof. Let $\chi: X \to \{-1, +1\}$ be a random coloring, the colors of points being chosen uniformly and independently. For any fixed set $S \subseteq X$, the quantity $\chi(S) = \sum_{x \in S} \chi(x)$ is a sum of $|S|$ independent random ± 1 variables. Theorem 7.1.1 tells us that

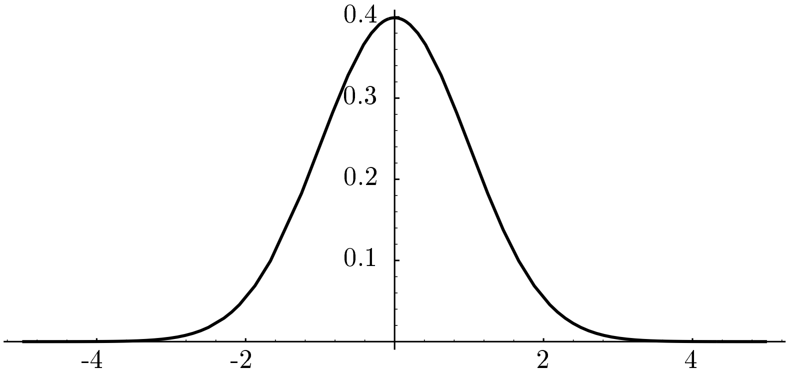
$$\mathrm{P}[|\chi(S)| > t] < 2e^{-t^2/2|S|} \le 2e^{-t^2/2s}.$$

For $t = \sqrt{2s \ln(2m)}$, $2e^{-t^2/2s}$ becomes $\frac{1}{m}$. Thus, with a positive probability, a random coloring satisfies $|\chi(S)| \le t$ for all $S \in \mathcal{F}$ simultaneously. $\square$

7.2 Sums of Bounded Independent Random Variables

Estimates like that in Theorem 7.1.1 hold in much greater generality. For understanding such results, it is useful to keep in mind a marvelous result of probability theory: the Central Limit Theorem. We remark that the following discussion, up until Theorem 7.2.1, is not necessary for understanding the subsequent results, and so a reader who does not feel at ease with continuous distributions, say, can skip this part.

First we recall that a real random variable Z has the *standard normal distribution* $N(0, 1)$ if its density is given by the function $\frac{1}{\sqrt{2\pi}} e^{-x^2/2}$:



(so $\mathrm{P}[Z \leq t] = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}\, \mathrm{d}x$). We have $\mathbf{E}[Z] = 0$ and $\mathrm{Var}[Z] = 1$, and Z is concentrated around its expectation: the probability of deviating from 0 by more than λ is roughly proportional to $e^{-\lambda^2/2}$ for large λ.

The Central Limit Theorem asserts that if S is the sum of many *independent* random variables, none of them with unreasonably large variance compared to the others, then the normalized random variable

$$\frac{S - \mathbf{E}[S]}{\sqrt{\mathrm{Var}[S]}}$$

has approximately the standard normal distribution $N(0,1)$. This looks like magic, since the distributions of the summands can be rather arbitrary and have nothing to do with the normal distribution. One simple formulation of the Central Limit Theorem is as follows. Let $X_1, X_2, \ldots$ be a sequence of independent random variables with $\mathbf{E}[X_i] = 0$, let $S_n = \sum_{i=1}^{n} X_i$, and suppose that for all i, $\mathrm{Var}[X_i]/\mathrm{Var}[S_n] \to 0$ as $n \to \infty$. Then the distribution function of the normalized random variable $Z_n = S_n/\sqrt{\mathrm{Var}[S_n]}$ converges to the distribution function of $N(0,1)$, i.e. for any real t, $\mathrm{P}[Z_n \leq t] \to \mathrm{P}[Z \leq t]$ as $n \to \infty$. (The condition on the $\mathrm{Var}[X_i]$, called Feller's condition, can be considerably weakened—see a probability theory textbook.)

This theorem as stated doesn't tell us anything about the speed of the convergence to the normal distribution, and so it cannot be used for obtaining concrete tail estimates for sums of finitely many random variables. But it is a useful heuristic guide, suggesting what behavior of a sum of independent random variables we should expect. Here we state a useful and quite general concentration result.

7.2.1 Theorem. *Let $X_1, X_2, \ldots, X_n$ be independent random variables, each of them attaining values in $[0,1]$, let $X = X_1 + X_2 + \cdots + X_n$, and let $\sigma^2 = \mathrm{Var}[X] = \sum_{i=1}^{n} \mathrm{Var}[X_i]$. (In particular, if $X_i = 1$ with probability p and $X_i = 0$ with probability $1-p$, then $\mathrm{Var}[X] = np(1-p)$, and so we can use $\sigma \leq \sqrt{np}$.) Then, for any $t \geq 0$,*

$$\mathrm{P}[X \geq \mathbf{E}[X] + t] < e^{-t^2/2(\sigma^2 + t/3)} \quad \textit{and} \quad \mathrm{P}[X \leq \mathbf{E}[X] - t] < e^{-t^2/2(\sigma^2 + t/3)}.$$

This theorem can be proved along the same lines as Theorem 7.1.1, only the estimates become more complicated. Note that in a wide range of t, say up to $t = \sigma^2$, the estimate is close to $e^{-t^2/2\sigma^2}$, and this is approximately the value predicted by the approximation of the distribution of X by the appropriately scaled normal distribution. For larger t, though, the correction factor $t/3$ gradually makes the estimate weaker than $e^{-t^2/2\sigma^2}$. Some correction like this is actually necessary in general for these very large deviations.

Let us remark that many other estimates of this kind can be found in the literature (associated with the names of Bernstein, Hoeffding, and some others), and sometimes they are slightly sharper.

Randomized rounding. This is a general technique in combinatorial optimization, which in many cases allows us to compute approximate solutions for NP-hard problems. The analysis is based on Theorem 7.2.1. Here we present one specific example: randomized rounding applied to the *k-matching problem*. Let $V = \{v_1, v_2, \ldots, v_n\}$ be a set and let $\mathcal{F} = \{S_1, S_2, \ldots, S_m\}$ be a system of subsets of V. A subsystem $\mathcal{M} \subseteq \mathcal{F}$ is called a *k-matching*[3] (or sometimes a k-packing[4]) if no point of V is contained in more than k sets of $\mathcal{M}$. Given V, $\mathcal{F}$, and k, we would like to find a k-matching $\mathcal{M}$ with as many sets as possible.

Let A denote the $n \times m$ incidence matrix of the system $\mathcal{F}$, with rows corresponding to points and columns to sets; that is, $a_{ij} = 1$ if $v_i \in S_j$ and $a_{ij} = 0$ otherwise. Let $\mathbf{1}$ denote the (column) vector of 1's (of appropriate length). Then the k-matching problem for $\mathcal{F}$ can be expressed as the following integer program:

$$\max\{\mathbf{1}^T x \colon x \in \{0,1\}^m,\ Ax \le k\mathbf{1}\}.$$

The correspondence to the original problem is simple: the set S_j is put into the k-matching $\mathcal{M}$ exactly when $x_j = 1$.

With the restriction $x \in \{0,1\}^m$, this is an NP-hard problem (since the k-matching problems is known to be NP-hard). But efficient algorithms for linear programming allow us to solve the *linear relaxation* in polynomial time: compute an optimal solution x^* of the linear program

$$\max\{\mathbf{1}^T x \colon x \in [0,1]^m,\ Ax \le k\mathbf{1}\}.$$

Let $OPT^* = \mathbf{1}^T x^*$ denote the optimal value. We note that $OPT^* \ge OPT$, where OPT is the optimal value of the integer program, i.e. the number of sets in a largest k-matching.

In order to get an approximate solution to the k-matching problem, we want to round each component of x^* to 0 or 1. The idea of randomized rounding is to use the real number x_j^* as the probability of rounding the jth component to 1. We begin with a preliminary consideration, which does not yet quite work.

Let us define a random vector $y \in \{0,1\}^m$ by choosing $y_j = 1$ with probability x_j^* and $y_j = 0$ with probability $1 - x_j^*$, the choices for various j being mutually independent. By linearity of expectation, we have $\mathbf{E}\left[\mathbf{1}^T y\right] = \mathbf{1}^T x^* = OPT^*$ and $\mathbf{E}\left[(Ay)_i\right] = (Ax^*)_i \le k$ for all i. Moreover, the quantity $\mathbf{1}^T y = \sum_{j=1}^m y_j$ is the sum of 0/1 independent random variables, and the tail estimates in Theo-

[3] matching = párování
[4] packing = pakování

rem 7.2.1 show that with high probability, its value is close to OPT^*. Similarly, for each i, $(Ay)_i$ is likely to be near $(Ax^*)_i$ and thus not much larger than k.

In this way, we would get a solution which is "nearly" a k-matching but some points are typically contained in somewhat more than k sets. In order to get an actual k-matching by the rounding procedure, we slightly lower the probabilities of 1's. Namely, now we set y_j to 1 with probability only $(1-\frac{\varepsilon}{2})x_j^*$. This works if k is sufficiently large:

7.2.2 Proposition. *Let $\varepsilon \in (0,1]$ be a parameter, and let us suppose that $k \geq \frac{10}{\varepsilon^2}\ln(2n+2)$. Then with probability at least $\frac{1}{2}$, the vector y obtained by the just described randomized rounding procedure defines a k-matching with at least $(1-\varepsilon)OPT$ sets.*

Proof. Let us write $X = \sum_{j=1}^m y_j = \mathbf{1}^T y$. First we estimate the probability $\mathrm{P}[X < (1-\varepsilon)OPT^*]$. We note that $OPT^* \geq k$, since any 0/1 vector x with k ones satisfies $Ax \leq k\mathbf{1}$. We have $\mathbf{E}[X] = (1-\frac{\varepsilon}{2})OPT^*$ and $\mathrm{Var}[X] \leq \mathbf{E}[X]$ (this is always true for a sum of independent random 0/1 variables). So we use the second inequality in Theorem 7.2.1 with $t = \frac{\varepsilon}{2}\,OPT^*$ and $\sigma^2 \leq OPT^*$. This yields $\mathrm{P}[X < (1-\varepsilon)OPT^*] \leq e^{-(\varepsilon^2/10)OPT^*} \leq e^{-(\varepsilon^2/10)k} \leq \frac{1}{2n+2}$.

Next, we write $Y_i = (Ay)_i$ and we estimate $\mathrm{P}[Y_i > k]$ in a very similar way. This time $\mathbf{E}[Y_i] = (1-\frac{\varepsilon}{2})(Ax^*)_i \leq (1-\frac{\varepsilon}{2})k$, and we can set $t = \frac{\varepsilon}{2}\,k$ and $\sigma^2 = k$ in the first inequality in Theorem 7.2.1. We obtain $\mathrm{P}[Y_i > k] \leq \frac{1}{2n+2}$. Therefore, with probability at least $\frac{1}{2}$, we have $Ay \leq k\mathbf{1}$ as well as $\mathbf{1}^T y \geq (1-\varepsilon)OPT^* \geq (1-\varepsilon)OPT$. □

The same approach can be used for many other problems expressible as integer programs with 0/1 variables. These include problems in VLSI design (routing), multicommodity flows, and independent sets in hypergraphs, to name just a few. Some recent results in this direction can be found, for example, in

A. Srinivasan: Improved approximation guarantees for packing and covering integer programs, *SIAM J. Computing* 29(1999) 648–670.

7.3 A Lower Bound For the Binomial Distribution

Sometimes we need a lower bound for probabilities like $\mathrm{P}[X \geq \mathbf{E}[X] + t]$; we need to know that the probability of deviation t is not *too* small. The Central Limit Theorem suggests that the distribution of the sum of many independent random variables is approximately normal, and so the bounds as in Theorems 7.1.1 and 7.2.1 should not be far from the truth. It turns out that this is actually the case, under quite general circumstances. Such general and precise bounds can be found in

W. Feller: Generalization of a probability limit theorem of Cramér, *Trans. Am. Math. Soc*, 54:361–372, 1943.

For example, the following is an easy consequence of Feller's results:

7.3.1 Theorem. *Let X be a sum of independent random variables, each attaining values in $[0,1]$, and let $\sigma = \sqrt{\mathrm{Var}\,[X]} \geq 200$. Then for all $t \in [0, \frac{\sigma^2}{100}]$, we have*

$$\Pr[X \geq \mathbf{E}\,[X] + t] \geq ce^{-t^2/3\sigma^2}$$

for a suitable constant $c > 0$.

Here we will prove just a counterpart of Theorem 7.1.1:

7.3.2 Proposition. *For n even, let $X_1, X_2, \ldots, X_n$ be independent random variables, each attaining the values 0 and 1, both with probability $\frac{1}{2}$. Let $X = X_1 + X_2 + \cdots + X_n$. Then we have, for any integer $t \in [0, \frac{n}{4}]$,*

$$\mathrm{P}\left[X \geq \tfrac{n}{2} + t\right] \geq \tfrac{1}{15}\, e^{-16t^2/n}.$$

Proof. A good exercise in elementary estimates. Write $n = 2m$. For $t \leq \frac{m}{2}$, we have

$$\mathrm{P}[X \geq m + t] = 2^{-2m} \sum_{j=t}^{m} \binom{2m}{m+j} \geq 2^{-2m} \sum_{j=t}^{2t} \binom{2m}{m+j}.$$

Using $\binom{2m}{m} \geq 2^{2m}/2\sqrt{m}$ and $1 - x \geq e^{-2x}$ for $0 \leq x \leq \frac{1}{2}$ we get

$$\begin{aligned} \binom{2m}{m+j} &= \binom{2m}{m} \frac{m}{m+j} \cdot \frac{m-1}{m+j-1} \cdots \frac{m-j+1}{m+1} \\ &= \binom{2m}{m} \prod_{i=1}^{j} \left(1 - \frac{j}{m+i}\right) \\ &\geq \frac{2^{2m}}{2\sqrt{m}} \left(1 - \frac{j}{m}\right)^j \geq \frac{2^{2m}}{2\sqrt{m}} \cdot e^{-2j^2/m}. \end{aligned}$$

So

$$\mathrm{P}[X \geq m + t] \geq \frac{1}{4\sqrt{m}}\, t\, e^{-8t^2/m}.$$

For $t \geq \frac{1}{4}\sqrt{m}$, the right-hand side is at least $\frac{1}{8}e^{-16t^2/n}$. For $0 \leq t < \frac{1}{4}\sqrt{m}$, we have $\mathrm{P}[X \geq m + t] \geq \mathrm{P}\left[X \geq m + \frac{1}{4}\sqrt{m}\right] \geq \frac{1}{8}e^{-1/2} \geq \frac{1}{15}$. Thus, the claimed bound holds for all $t \leq \frac{m}{2}$. The constants in the estimate could be improved, of course. □

A lower bound for discrepancy. We show that the upper bound of $O(\sqrt{n \log(2m)}\,)$ for the discrepancy of m sets on n points (Proposition 7.1.2) is nearly the best possible in a wide range of values of m.

7.3.3 Proposition. *For all m, $15n \leq m \leq 2^n$, there are systems of m sets on n points with discrepancy at least $\Omega(\sqrt{n \ln(m/15n)}\,)$.*

For $m \geq n^2$, say, the lower and upper bounds in Propositions 7.1.2 and 7.3.3 are the same up to a constant. For m close to n, there is a gap. It turns out that it is the upper bound which can be improved (by a very sophisticated probabilistic argument). The correct bound for the maximum discrepancy of m sets on n points, $m \geq n$, is of order $\sqrt{n \ln(2m/n)}$.

Proof. Consider a random set system $\mathcal{F} = \{S_1, S_2, \ldots, S_m\}$ on the ground set $[n]$, n even, where the S_i are independent random subsets of $[n]$; that is, each $x \in [n]$ is included in S_i independently with probability $\frac{1}{2}$.

Let $\chi\colon [n] \to \{-1,+1\}$ be an arbitrary fixed coloring, and suppose that the number of -1's is a and the number of $+1$'s is $n-a$. A point $x \in [n]$ with $\chi(x) = 1$ contributes 1 to $\chi(S_i)$ if $x \in S_i$ and 0 if $x \notin S_i$. Since $x \in S_i$ has probability $\frac{1}{2}$, the contribution of x to $\chi(S_i)$ is a random variable attaining values 0 and 1 with probability $\frac{1}{2}$. Similarly, the contribution of an x with $\chi(x) = -1$ attains values 0 and -1 with probability $\frac{1}{2}$. Therefore, $\chi(S_i)$ is a sum of n independent random variables, a of them attaining values -1 and 0 with probability $\frac{1}{2}$ and $n-a$ of them attaining values 0 and 1 with probability $\frac{1}{2}$. Then $\chi(S_i) + a$ is the sum of n independent random variables, each with values 0 and 1. For $a \leq \frac{n}{2}$, we have

$$\mathrm{P}\left[|\chi(S_i)| \geq t\right] \geq \mathrm{P}\left[\chi(S_i) + a \geq t + a\right] \geq \mathrm{P}\left[\chi(S_i) + a \geq \tfrac{n}{2} + t\right].$$

By Proposition 7.3.2, the last probability is at least $\frac{1}{15} e^{-16t^2/n}$, provided that $t \leq \frac{n}{4}$. For $a > \frac{n}{2}$, we get the same bound by symmetry (consider the coloring $-\chi$). Therefore, for any of the possible 2^n colorings χ, we have

$$\mathrm{P}\left[\operatorname{disc}(\mathcal{F}, \chi) \leq t\right] \leq \left(1 - \tfrac{1}{15}\, e^{-16t^2/n}\right)^m \leq e^{-m e^{-16t^2/n}/15}.$$

For $t = \frac{1}{4}\sqrt{n \ln(m/15n)}$ (which is below $\frac{n}{4}$ for $m \leq 2^n$), the last expression becomes $e^{-n} < 2^{-n}$, and we can conclude that with a positive probability, the discrepancy of our random $\mathcal{F}$ is at least $\frac{1}{4}\sqrt{n \ln(m/15n)}$ under *any* coloring χ. □

A deterministic bound using Hadamard matrices. Proposition 7.3.3 allows us to conclude the existence of n sets on n points with discrepancy at least $c\sqrt{n}$ for some constant $c > 0$ (can you see how?). Here we show a beautiful deterministic argument proving this result.

We first recall the notion of an *Hadamard matrix*. This is an $n \times n$ matrix H with entries $+1$ and -1 such that any two distinct columns are orthogonal; in other words, we have $H^T H = nI$, where I stands for the $n \times n$ identity matrix. Moreover, we assume that the first row and the first column consist of all 1's.

Hadamard matrices do not exist for every n. For example, it is clear that for $n \geq 2$, n has to be even, and with a little more effort one can see that n must be divisible by 4 for $n \geq 4$. The existence problem for Hadamard matrices is not yet fully solved, but various constructions are known. We recall only one simple recursive construction, providing a $2^k \times 2^k$ Hadamard matrix for all natural numbers k. Begin with the 1×1 matrix $H_0 = (1)$, and, having defined a $2^{k-1} \times 2^{k-1}$ matrix H_{k-1}, construct H_k from four blocks as follows:

$$\begin{pmatrix} H_{k-1} & H_{k-1} \\ H_{k-1} & -H_{k-1} \end{pmatrix}.$$

Thus, we have

$$H_1 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \quad H_2 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}.$$

The orthogonality is easy to verify by induction.

Let H be a $4n \times 4n$ Hadamard matrix. Each column except for the first one is orthogonal to the column of all 1's, and so the number of 1's in it is $2n$, as well as the number of -1's. Moreover, the ith and jth columns, $1 < i < j$, are orthogonal too, and it follows that they have exactly n common 1's, n common -1's, and $2n$ positions where one of them has 1 and the other has -1 (check).

Let A be the $(4n-1) \times (4n-1)$ matrix arising from H by deleting the first row and first column and changing the -1's to 0's. By the above, we find that $A^T A = nI + (n-1)J$, where I is the $(4n-1) \times (4n-1)$ identity matrix and J is the $(4n-1) \times (4n-1)$ matrix of all 1's.

Consider the system of sets $S_1, S_2, \ldots, S_{4n-1}$ on $[4n-1]$, where S_i has the ith row of A as the characteristic vector. Let $\chi\colon [4n-1] \to \{-1,+1\}$ be any coloring of the ground set, and let $x \in \{-1,+1\}^n$ be χ interpreted as the column vector, i.e. $x_i = \chi(i)$. By the definition of matrix multiplication, we have

$$Ax = \Big(\chi(S_1), \chi(S_2), \ldots, \chi(S_{4n-1})\Big)^T.$$

Therefore,

$$\begin{aligned}
\sum_{i=1}^{4n-1} \chi(S_i)^2 &= \|Ax\|^2 = (Ax)^T(Ax) = x^T(A^T A)x \\
&= x^T(nI + (n-1)J)x = nx^T Ix + (n-1)x^T Jx \\
&= n\|x\|^2 + (n-1)\Big(\sum_{i=1}^{4n-1} x_i\Big)^2 \geq n(4n-1).
\end{aligned}$$

So for any χ, the average $\chi(S_i)^2$ is at least n, and there exists an i with $|\chi(S_i)| \geq \sqrt{n}$. We have proved that the discrepancy of the set system $\{S_1, \ldots, S_{4n-1}\}$ is at least $\sqrt{n}$. □

7.4 Sums of Moderately Dependent Indicator Variables

Here we present, without a proof, a powerful tail estimate for a sum $X = X_1 + \cdots + X_n$, where X_i attains values 0 and 1 and where some of the X_i may be dependent, but the amount of dependence is suitably bounded.

We will need the notion of a *dependency graph* for a family of random variables. Note that it is slightly different from the one used in Section 6.1 where we considered only random events and the dependency graph was directed!

7.4.1 Definition. *Families of real random variables* $\{X_i\colon i \in A\}$ *and* $\{X_i\colon i \in B\}$ *are* **mutually independent** *if for any choice of* $a_i \in \mathbf{R}$, $i \in A \cup B$,

$$\mathrm{P}[\forall i \in A \cup B\colon X_i \le a_i] = \mathrm{P}[\forall i \in A\colon X_i \le a_i]\,\mathrm{P}[\forall i \in B\colon X_i \le a_i]\,.$$

7.4.2 Definition. *A graph* G *is a* **dependency graph** *for a family of random variables* $\{X_i\colon i \in I\}$ *if* $V(G) = I$, *and for any two disjoint sets* $A, B \subset V$ *with no edges between* A *and* B, *the families* $\{X_i\colon i \in A\}$ *and* $\{X_i\colon i \in B\}$ *are mutually independent.*

7.4.3 Theorem (Janson–Suen inequality). *Let* $X = X_1 + \cdots + X_n$, *where the* X_i *are random variables with* $\mathrm{P}[X_i = 1] = p_i$ *and* $\mathrm{P}[X_i = 0] = 1 - p_i$. *Let* E *be the edge set of a dependency graph of the* X_i, *and define*

$$\Delta = \mathbf{E}[X] + \sum_{\{i,j\}\in E} p_i p_j, \quad \delta = \max_{i\in[n]} \sum_{j\colon \{j,i\}\in E} p_j.$$

Then for any $t \ge 0$, *we have*

$$\mathrm{P}[X \le \mathbf{E}[X] - t] \le e^{-\min(t^2/4\Delta, t/6\delta)}.$$

Remarks. Note that the tail estimate is only one-sided; an exponentially small upper bound for $\mathrm{P}[X \ge \mathbf{E}[X] + t]$ need not hold in general. The theorem is mostly used for showing that $\mathrm{P}[X = 0]$ is very small, i.e. with $t = \mathbf{E}[X]$.

The quantity Δ is an upper bound for $\mathrm{Var}[X]$: We have

$$\mathrm{Var}[X] = \sum_{i=1}^{n} \mathrm{Var}[X_i] + \sum_{\{i,j\}\in E} \mathrm{Cov}[X_i, X_j]\,,$$

and $\mathrm{Var}[X_i] \le \mathbf{E}[X_i]$ and $\mathrm{Cov}[X_i, X_j] \le \mathbf{E}[X_i X_j]$ since $X_i \in \{0,1\}$. Such estimates for $\mathrm{Var}[X]$ were calculated in Section 5.3 in showing that $G(n,p)$ almost surely contains a copy of a given graph H. Theorem 7.4.3, too, has been developed with this application in mind.

Example. Let $H = K_3$ be the triangle. We know from Section 5.3 that if $p = \frac{\varphi}{(}n)n$ with $\varphi(n) \to \infty$, then $\mathrm{P}[K_3 \not\subseteq G(n,p)] \to 0$ as $n \to \infty$. Theorem 7.4.3 shows that this probability is even exponentially small in $\varphi(n)$. To see this, let $\left(X_T\colon T \in \binom{[n]}{3}\right)$ be the indicators of all possible triangles that can appear in $G(n,p)$, and let $X = \sum_T X_T$. We have $p_T = \mathrm{P}[X_T = 1] = p^3$. The edges of a dependency digraph connect triangles T and T' sharing at least two vertices. The same calculations as in Section 5.3 gives $\mathbf{E}[X] \sim n^3p^3 = \varphi(n)^3$ and $\Delta << n^3p^3 + n^4p^5 \sim \varphi(n)^3$. A simple calculation also shows that $\delta \sim np^3 \sim \varphi(n)^3/n^2$, which is very small. For $t = \mathbf{E}[X] \sim \varphi(n)^3$, we have $\min(t^2/4\Delta, t/6\delta) \sim \min(\varphi(n)^3, n^2)$, and so

$$\mathrm{P}[X = 0] \le e^{-\Omega(\min(\varphi(n)^3, n^2))}.$$

A similar bound can be derived for the containment of any fixed balanced graph H in $G(n,p)$. Such results have been obtained earlier with the aid of less powerful tools (Janson's inequality dealing with the probabilities of monotone events). But Theorem 7.4.3 yields similar bounds for containment of balanced graphs H in $G(n,p)$ in the *induced* sense, with calculation very similar to the non-induced case. Such a result appears considerably harder than the non-induced case, because of non-monotonicity, and illustrates the strength of Theorem 7.4.3.

Balls in urns: hypergeometric distribution. In conclusion, we mention another useful concentration result without a proof. We have N urns, labeled 1 through N, and we put m balls into m different urns at random (draws without replacement). Some n of the urns are "distinguished", and we let X denote the number of balls in the distinguished urns ($n, m \leq N$).

We have $\mathbf{E}[X] = \frac{nm}{N}$ and $\sigma^2 = \mathrm{Var}[X] = \frac{nm(N-n)(N-m)}{N^2(N-1)} \leq \frac{nm}{N} = \mathbf{E}[X]$. This X can obviously be written as the sum of n indicator variables ($X_i = 1$ if the ith distinguished urn receives a ball), but these are *not* independent. Nevertheless, it is known that the tail estimates as in Theorems 7.2.1 and 7.3.1 hold for this particular X (with σ and n as above). Knowing this can save many desperate calculations.

8

Concentration of Lipschitz Functions

8.1 Concentration on Product Spaces

We have seen that if X is a sum of many "small" independent random variables $X_1, X_2, \ldots, X_n$, then X is strongly concentrated around its expectation. In this chapter we present more general results, of the following type: If $(\Omega, \Sigma, \mathrm{P})$ is a "suitable" probability space and $f: \Omega \to \mathbf{R}$ is a "nice" random variable on it, then f is tightly concentrated around $\mathbf{E}[f]$.

For example, the basic Chernoff inequality for sums of independent uniform ± 1 variables (Theorem 7.1.1) can be recast as follows in this setting: We consider the probability space $\{-1,1\}^n$ with the uniform probability measure, and f is given by $f(\omega) = \omega_1 + \cdots + \omega_n$, where $\omega = (\omega_1, \ldots, \omega_n) \in \{-1,1\}^n$. Then $\mathrm{P}[f \geq \mathbf{E}[f] + t] < e^{-t^2/2n}$ and $\mathrm{P}[f \leq \mathbf{E}[f] - t] < e^{-t^2/2n}$.

Two essential features of this example will appear in the main theorem of this section.

- First, our probability space is a *product* of many (n) probability spaces; in our case, the factors are the spaces $\{-1,1\}$ with the uniform measure. (Having n independent random variables always implicitly entails a product space with n factors.)

- And second, the *effect* of each component ω_i on the value of f is relatively small: by changing the value of ω_i (and keeping the values of all the other ω_j), the value of f changes by at most 2.

What is the product of probability spaces $(\Omega_1, \Sigma_1, \mathrm{P}_1)$,..., $(\Omega_n, \Sigma_n, \mathrm{P}_n)$? The elementary events of the product have the form $\omega = (\omega_1, \omega_2, \ldots, \omega_n)$, where $\omega_i \in \Omega_i$, and so the ground set of the product is $\Omega = \Omega_1 \times \Omega_2 \times \cdots \times \Omega_n$. Intuitively, a random $\omega \in \Omega$ is selected by choosing each ω_i at random from Ω_i, all these choices being mutually independent. If all the Ω_i are finite, we can define the product measure P on Ω simply by $\mathrm{P}[\{(\omega_1, \ldots, \omega_n)\}] = \prod_{i=1}^n \mathrm{P}_i[\{\omega_i\}]$. For infinite Ω_i, the formal construction of the product measure is more sophisticated, and it is usually considered in courses on measure and integration. In our applications, we will mostly consider finite Ω_i.

Now we formulate a condition on the function f. Let $\Omega = \Omega_1 \times \cdots \times \Omega_n$ and let $f: \Omega \to \mathbf{R}$ be a (measurable) function, i.e. a real random variable on Ω. We say that *the ith coordinate has effect at most c_i on f* if $|f(\omega) - f(\omega')| \le c_i$ for all $\omega, \omega' \in \Omega$ that differ only in the ith coordinate. Here is the promised concentration result:

8.1.1 Theorem (Concentration on product spaces). *Let $(\Omega, \Sigma, \mathrm{P})$ be the product of probability spaces $(\Omega_i, \Sigma_i, \mathrm{P}_i)$, $i = 1, 2, \ldots, n$, and let $f\colon \Omega \to \mathbf{R}$ be a function such that the ith coordinate has effect at most c_i. Then*

$$\mathrm{P}[f \ge \mathbf{E}[f] + t] \le e^{-t^2/2\sigma^2} \quad \text{and} \quad \mathrm{P}[f \le \mathbf{E}[f] - t] \le e^{-t^2/2\sigma^2},$$

where $\sigma^2 = \sum_{i=1}^n c_i^2$. In particular, if each coordinate has effect at most 1, then

$$\mathrm{P}[f \ge \mathbf{E}[f] + t] \le e^{-t^2/2n} \quad \text{and} \quad \mathrm{P}[f \le \mathbf{E}[f] - t] \le e^{-t^2/2n}.$$

Thus, if no coordinate has effect more than 1, then f is concentrated at least as much as the sum of n independent random ± 1 variables.

Before we consider a more general version with Lipschitz functions and a proof, let us see a few applications of this powerful result.

The size of the image of a random function. Let $g\colon [n] \to [n]$ be a random function, all the n^n possible functions being equally likely, and let X be the number of elements in the image, $X = |g([n])|$. By the method of indicators, one can calculate that $\mathbf{E}[X] = n - n(1 - \frac{1}{n})^n \approx n(1 - \frac{1}{e})$, but we do not need to know $\mathbf{E}[X]$ in order to derive a strong concentration result for X.

Our X is a function on the product space $[n]^n$ (the ith coordinate is the value $g(i)$). By changing $g(i)$ and keeping all other $g(j)$ fixed, the size of the image changes by at most 1. Theorem 8.1.1 thus implies that X is strongly concentrated around $\mathbf{E}[X]$: $\mathrm{P}[|X - \mathbf{E}[X]| \ge t] \le 2e^{-t^2/2n}$.

Concentration of the chromatic number. Let us consider the probability space $G(n,p)$ of n-vertex random graphs, for some given n and p. Let χ be the function on this probability space assigning to each graph its chromatic number. It is not at all easy to determine $\mathbf{E}[\chi]$ (it is known quite precisely for a wide range of p, but the proofs are fairly sophisticated). But we do not need to know the expectation in order to apply Theorem 8.1.1!

To use Theorem 8.1.1, we need to consider $G(n,p)$ as a product space. There is a natural product structure corresponding to the potential edges; there are $\binom{n}{2}$ factors Ω_e, where each Ω_e has two elements corresponding to the absence and presence of the edge e in the graph. Clearly, adding or deleting an edge influences the chromatic number by at most 1, and so each of the $\binom{n}{2}$ coordinates has effect at most 1 on χ. Theorem 8.1.1 applies, but it doesn't yield anything interesting: the n in it would be $\binom{n}{2}$ here, and since χ is in the range $[1, n]$, the concentration result is rather useless (the bound is $e^{-t^2/2\binom{n}{2}} \ge e^{-n^2/n(n-1)} \ge e^{-2}$, so it never tends to 0).

The trick is to group the edges into larger chunks. Let $v_1, v_2, \ldots, v_n$ be the vertices enumerated in a fixed order, and let Ω_i be the probability space corresponding to the independent random choice of the edges going forward from v_i, i.e. $\{v_i, v_{i+1}\}$, $\{v_i, v_{i+2}\}$, $\ldots$, $\{v_i, v_n\}$. Then $G(n,p)$ is the product of these Ω_i, $i = 1, 2, \ldots, n-1$. Since changing the edges incident to a single vertex changes the chromatic number of a graph by at most 1, the effect of each coordinate on χ is at most 1. Theorem 8.1.1 now gives:

8.1.2 Theorem (Shamir–Spencer). *Let $n \geq 2$ and $p \in (0,1)$ be arbitrary, and let $c = c(n,p) = \mathbf{E}[\chi(G_{n,p})]$. Then*

$$\mathrm{P}[|\chi(G_{n,p}) - c| \geq t] \leq 2e^{-t^2/2(n-1)}.$$

So the chromatic number is almost always concentrated on about $\sqrt{n}$ values. By an ingenious argument (due to Bollobás), it can even be shown that for sparse random graphs, one of at most 4 values is attained most of the time:

8.1.3 Theorem (Four-value concentration). *Let $\alpha > \frac{5}{6}$ be fixed, and let $p = n^{-\alpha}$. Then for any n, there is an integer $u = u_\alpha(n)$ such that $\chi(G(n,p)) \in \{u, u+1, u+2, u+3\}$ almost surely; i.e.*

$$\lim_{n\to\infty} \mathrm{P}[u(n) \leq \chi(G(n,p)) \leq u(n)+3] = 1.$$

The key additional idea is that, typically, each subgraph of $G(n,p)$ on about $\sqrt{n}$ vertices can be 3-colored, and so deviations with about $\sqrt{n}$ harmful vertices can be fixed using 3 extra colors.

8.1.4 Lemma. *Let $\alpha > \frac{5}{6}$ be fixed, and let $p = n^{-\alpha}$. Then, almost surely, $G(n,p)$ has no subgraph H on at most $\sqrt{8n\ln n}$ vertices with $\chi(H) > 3$.*

Proof. What we really calculate is: almost surely, there is no subgraph on $t \leq \sqrt{8n\ln n}$ vertices with average degree at least 3. This suffices: We consider an inclusion-minimal subset of vertices such that the subgraph induced by it has chromatic number 4; as is easy to check, this subgraph must have all degrees at least 3.

First, let $t \geq 4$ be even. The probability that at least $\frac{3}{2}t$ edges live on some fixed set T of t vertices of $G(n,p)$ is at most (using $\binom{n}{k} \leq (en/k)^k$)

$$\binom{\binom{t}{2}}{3t/2} p^{3t/2} \leq \left(\frac{et^2/2}{3t/2}\right)^{3t/2} p^{3t/2} = \left(\frac{te}{3}\right)^{3t/2} n^{-3\alpha t/2}.$$

There are $\binom{n}{t} \leq (ne/t)^t$ choices of T, and so the probability of existence of at least one T with at least $\frac{3}{2}t$ edges is at most

$$\left[\frac{ne}{t} \cdot \frac{t^{3/2}e^{3/2}}{3^{3/2}} n^{-3\alpha/2}\right]^t.$$

The expression in brackets is at most

$$O(t^{1/2}n^{1-3\alpha/2}) = O(n^{5/4-3\alpha/2}(\ln n)^{1/4}),$$

which goes to 0 as $n \to \infty$ since $\alpha > \frac{5}{6}$. For t odd, the calculation is technically a little more complicated since we need to deal with the integer part, as we have

$\lceil \frac{3}{2}t \rceil$ edges, but the resulting probability is also bounded by $o(1)^t$. The proof is finished by summing over all $t \in [4, \sqrt{8n \ln n}\,]$. □

Proof of Theorem 8.1.3. Let u be the smallest integer such that

$$\mathrm{P}[\chi(G(n,p)) \le u] > \frac{1}{n}.$$

Let X be the minimum number of vertices whose deletion makes $G(n,p)$ u-colorable. When X is viewed as a function on the product space $\prod_{i=1}^{n-1} \Omega_i$ as in the proof of the Shamir–Spencer theorem 8.1.2, each of the n coordinates has effect at most 1 on it (right?). We thus have the tail estimates from Theorem 8.1.1:

$$\mathrm{P}[X \ge \mathbf{E}[X] + t] \le e^{-t^2/2(n-1)}, \quad \mathrm{P}[X \le \mathbf{E}[X] - t] \le e^{-t^2/2(n-1)}.$$

Let us set $t = \sqrt{2(n-1)\ln n}$, so that the right-hand sides become $\frac{1}{n}$. By the definition of u, $G(n,p)$ is u-colorable with probability greater than $\frac{1}{n}$, and so $\frac{1}{n} < \mathrm{P}[X = 0] = \mathrm{P}[X \le \mathbf{E}[X] - \mathbf{E}[X]]$. Combined with the second tail estimate, this shows that $\mathbf{E}[X] < t$, and the first tail estimate then gives $\mathrm{P}[X \ge 2t] \le \mathrm{P}[X \ge \mathbf{E}[X] + t] \le \frac{1}{n}$. So with probability at least $1 - \frac{1}{n}$, $G(n,p)$ with some $2t$ vertices removed can be u-colored. By Lemma 8.1.4, the subgraph on the removed $2t$ vertices is 3-colorable almost surely, and so all of $G(n,p)$ can be colored with at most $u+3$ colors almost surely. On the other hand, by the definition of u, $\chi(G(n,p)) \ge u$ almost surely as well. □

8.2 Concentration of Lipschitz Functions, With a Proof

There are several ways of proving Theorem 8.1.1 (concentration in product spaces). Here we present one of the conceptually simplest proofs. A natural formulation needs a somewhat more general setting, with Lipschitz functions.

Let M_1 be a metric space with a metric ρ_1, M_2 a metric space with a metric ρ_2, and $K > 0$ a real number. We recall that a mapping $\varphi\colon M_1 \to M_2$ is called *K-Lipschitz* if it expands no distance in ratio larger than K; that is, if $\rho_2(\varphi(x), \varphi(y)) \le K\rho_1(x,y)$ for all $x, y \in M_1$.

We consider spaces equipped with both a probability measure and a metric. A *metric probability space* is a four-tuple $(\Omega, \Sigma, \mathrm{P}, \rho)$, where $(\Omega, \Sigma, \mathrm{P})$ is a probability space and ρ is a metric on Ω.

Let us consider the situation as in Theorem 8.1.1 with $c_1 = c_2 = \cdots = c_n = 1$ (each coordinate has effect at most 1). Let us view each factor $(\Omega_i, \Sigma_i, \mathrm{P}_i)$ as a metric probability space with the "discrete" metric ρ_i given by $\rho_i(\omega_i, \omega_i') = 1$ for every two distinct elements $\omega_i, \omega_i' \in \Omega_i$. A metric ρ on the product space $(\Omega, \Sigma, \mathrm{P})$ is defined by $\rho(\omega, \omega') = \sum_{i=1}^n \rho_i(\omega_i, \omega_i')$, where $\omega = (\omega_1, \ldots, \omega_n), \omega' = (\omega_1', \ldots, \omega_n') \in \Omega$. For our specific choice of the metrics ρ_i, the resulting ρ is the *Hamming metric*; the distance of two vectors ω, ω' is the number of coordinates where they differ. If $f\colon \Omega \to \mathbf{R}$ is a function, then, as is easy to check, each coordinate has effect at most 1 if and only if f is 1-Lipschitz, where Ω is considered with the just introduced metric ρ, and $\mathbf{R}$ with the usual metric.

The definition of ρ on the product space makes sense for arbitrary metrics ρ_i on the factors. We write $\rho = \rho_1 + \rho_2 + \cdots + \rho_n$ and we call ρ the *ℓ_1-sum* of the ρ_i. We prove the following generalization of Theorem 8.1.1:

8.2.1 Theorem. *For $i = 1, 2, \ldots, n$, let $(\Omega_i, \Sigma_i, \mathrm{P}_i, \rho_i)$ be a metric probability space, and suppose that the diameter (maximum distance) of (Ω_i, ρ_i) is at most c_i. Let $M = (\Omega, \Sigma, \mathrm{P}, \rho)$ be the product space with $\rho = \rho_1 + \rho_2 + \cdots + \rho_n$. Then for any 1-Lipschitz (and measurable) function $f\colon \Omega \to \mathbf{R}$ and for all $t \geq 0$, we have*

$$\mathrm{P}[f \geq \mathbf{E}[f] + t] \leq e^{-t^2/2\sigma^2} \quad \textit{and} \quad \mathrm{P}[f \leq \mathbf{E}[f] - t] \leq e^{-t^2/2\sigma^2},$$

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

The proof resembles the proof of the basic Chernoff inequality (Theorem 7.1.1) in many features. In that proof, we estimated the expectation $\mathbf{E}\left[e^{uX}\right]$, where X was the considered random variable. Here we define a similar quantity for a general metric probability space $M = (\Omega, \Sigma, \mathrm{P}, \rho)$: the *Laplace functional* of M is a function $E_M\colon (0, \infty) \to \mathbf{R}$ given by

$$E_M(u) = \sup\left\{\mathbf{E}\left[e^{uf}\right] : f\colon \Omega \to \mathbf{R} \text{ is 1-Lipschitz and } \mathbf{E}[f] = 0\right\}.$$

First we show that a bound on E_M implies concentration of Lipschitz functions; this is exactly as in the proof of Chernoff's inequality. Assume that $E_M(u) \leq e^{au^2/2}$ for some $a > 0$ and all $u > 0$, and let $f\colon \Omega \to \mathbf{R}$ be 1-Lipschitz. We may suppose that $\mathbf{E}[f] = 0$. Using Markov's inequality for the random variable $Y = e^{uf}$, we have $\mathrm{P}[f \geq t] = \mathrm{P}[Y \geq e^{tu}] \leq \mathbf{E}[Y]/e^{tu} \leq E_M(u)/e^{tu} \leq e^{au^2/2 - ut}$, and setting $u = \frac{t}{a}$ yields $\mathrm{P}[f \geq t] \leq e^{-t^2/2a}$. So it suffices to show that under the assumptions of Theorem 8.2.1, $E_M(u) \leq e^{\sigma^2 u^2/2}$.

Next, crucially, we prove that the Laplace functional is submultiplicative.

8.2.2 Lemma. *Let $M_1 = (\Omega_1, \Sigma_1, \mathrm{P}_1, \rho_1)$ and $M_2 = (\Omega_2, \Sigma_2, \mathrm{P}_2, \rho_2)$ be metric probability spaces, and let $M = (\Omega, \Sigma, \mathrm{P}, \rho)$ be their product with $\rho = \rho_1 + \rho_2$. Then $E_M(u) \leq E_{M_1}(u) \cdot E_{M_2}(u)$ for all $u > 0$.*

Proof. Let $f\colon \Omega \to \mathbf{R}$ be 1-Lipschitz with $\mathbf{E}[f] = 0$. We set $g(y) = \mathbf{E}_x[f(x,y)] = \int_{\Omega_1} f(x,y)\, \mathrm{dP}_1(x)$ (the expectation of $f(x,y)$ with y fixed and x random). We rewrite

$$\begin{aligned}
\mathbf{E}\left[e^{uf}\right] &= \int_{\Omega_2}\int_{\Omega_1} e^{uf(x,y)}\, \mathrm{dP}_1(x)\, \mathrm{dP}_2(y) \\
&= \int_{\Omega_2} e^{ug(y)} \left(\int_{\Omega_1} e^{u(f(x,y)-g(y))}\, \mathrm{dP}_1(x)\right) \mathrm{dP}_2(y).
\end{aligned}$$

For every y, the function $x \mapsto f(x,y) - g(y)$ has zero expectation and it is clearly 1-Lipschitz, and so the inner integral is at most $E_{M_1}(u)$. Next, we have

$\mathbf{E}[g] = 0$ and we claim that g is also 1-Lipschitz. Indeed,

$$\begin{aligned} |g(y) - g(y')| &= \left| \int_{\Omega_1} f(x,y) - f(x,y')\,\mathrm{dP}_1(x) \right| \\ &\leq \int_{\Omega_1} |f(x,y) - f(x,y')|\,\mathrm{dP}_1(x) \\ &\leq \int_{\Omega_1} \rho_2(y,y')\,\mathrm{dP}_1(x) = \rho_2(y,y'). \end{aligned}$$

So $\int_{\Omega_2} e^{ug(y)}\,\mathrm{dP}_2(y) \leq E_{M_2}(u)$ and we are done. □

Finally, to prove Theorem 8.2.1, it remains to bound the Laplace functional of the factors.

8.2.3 Lemma. *Let $M = (\Omega, \Sigma, \mathrm{P}, \rho)$ be a metric probability space of diameter at most c. Then $E_M(u) \leq e^{c^2u^2/2}$ for all $u > 0$.*

Proof. For simplicity, we give the proof with $c = 1$. If $f\colon \Omega \to \mathbf{R}$ is 1-Lipschitz with $\mathbf{E}[f] = 0$, then its range is contained in $[-1,1]$. Let h be the linear function given by $h(x) = x\sinh u + \cosh u$, where $\cosh u = \frac{1}{2}(e^u + e^{-u})$ and $\sinh u = \frac{1}{2}(e^u - e^{-u})$. Elementary calculus shows that $h(x) \geq e^{ux}$ holds for all $x \in [-1,1]$ (use Taylor series). So

$$\mathbf{E}\left[e^{uf}\right] \leq \mathbf{E}[h \circ f] = \mathbf{E}[f]\sinh u + \cosh u = \cosh u \leq e^{u^2/2}.$$

This proves the lemma, and Theorem 8.2.1 follows. □

Variations. If we can prove better bounds for the Laplace functionals of the factors than the general Lemma 8.2.3, the above proof method yields an improvement over Theorem 8.2.1. One such possible improvement is similar to the passage from the basic Chernoff inequality for sums of independent uniform ± 1 random variables to the more general form, Theorem 7.2.1, dealing with sums of arbitrary independent bounded random variables. Here, for simplicity, we consider only an illustrative special case.

We suppose that each factor $M_i = (\Omega_i, \Sigma_i, \mathrm{P}_i, \rho_i)$ consists of two points, say $\Omega_i = \{0,1\}$, with probabilities $1-p$ and p, and with $\rho_i(0,1) = 1$ (so that the product space models n tosses of a biased coin). Let $f\colon \Omega_i \to \mathbf{R}$ be a 1-Lipschitz function on M_i with $\mathbf{E}[f] = 0$. That is, $|f(0) - f(1)| \leq 1$ and $(1-p)f(0) + pf(1) = 0$. Then $\mathbf{E}\left[e^{uf}\right] = (1-p)e^{uf(0)} + pe^{uf(1)}$, and elementary calculus shows that this expression is maximized, under the above two conditions on $f(0)$ and $f(1)$, for $f(0) = -p$ and $f(1) = 1-p$. Therefore, $E_{M_i}(u) \leq pe^{(1-p)u} + (1-p)e^{-pu} = e^{-pu}(1 - p + pe^u)$, and for the product space M, we have $E_M(u) \leq e^{-npu}(1 - p + pe^u)^n$. Using Markov's inequality as usual and performing some heroic calculations (for which we refer to the book of Janson, Łuczak, and Ruciński, Theorem 2.1), one can arrive at the following counterpart of Theorem 7.2.1:

8.2.4 Proposition. *Let M_i be the two-point probability spaces as above, let $M = (\Omega, \Sigma, \mathrm{P}, \rho)$ be their product, and let $f: \Omega \to \mathbf{R}$ be a 1-Lipschitz function. Then, for all $t > 0$,*

$$\mathrm{P}[f \geq \mathbf{E}[f] + t] < e^{-t^2/2(\sigma^2+t/3)} \quad \textit{and} \quad \mathrm{P}[f \leq \mathbf{E}[f] - t] < e^{-t^2/2(\sigma^2+t/3)},$$

where $\sigma^2 = np$.

More results of this type can be found in

> D. A. Grable: A large deviation inequality for functions of independent, multi-way choices, *Combinatorics, Probability and Computing* 7,1(1998) 57–63.

The proofs in that paper use martingales; this notion will be briefly discussed later.

Another strengthening of Theorem 8.1.1 is based on the observation that the Lipschitz condition for f need not be used in full in the proof. The idea, introduced by Alon, Kim, and Spencer, is to imagine that we are trying to find the value of f by making queries about the values of the ω_i to a truthful oracle (such as "what is the value of ω_7?"). Sometimes we can perhaps infer the value of f by querying the values of only some of the variables. Or sometimes, having learned the values of some of the variables, we know that some other variable cannot influence the value of f by much (although that variable may have much greater influence in other situations). By devising a clever querying strategy, the bound for σ^2 can again be reduced in some applications; see Grable's paper cited above.

8.3 Martingales, Azuma's Inequality, and Concentration on Permutations

Theorem 8.2.1 has been generalized in many ways; we will indicate some of them. One direction of generalizations replaces the assumption that we deal with a product space with many factors by weaker assumptions. The essential fact about the considered metric probability space is not the product structure, but some kind of "high dimensionality".

In this section, we consider the (rather sophisticated) probabilistic notion of a martingale, which leads to quite general concentration results. Currently it seems that in practically all applications of this kind, martingales can be replaced by other, even more powerful, tools. But martingales are often encountered in proofs in the literature, and so we introduce them at least briefly.

Let (Ω, Σ, P) be a probability space, and let $\Xi_0 = \{\emptyset, \Omega\} \subset \Xi_1 \subset \Xi_2 \subset \cdots \subseteq \Sigma$ be a sequence of σ-algebras[1] on Ω. In the case of a finite Ω, one can think of the Ξ_i as successively finer and finer partitions of Ω. (Formally, in this case, Ξ_i is the σ-algebra generated by some partition Π_i of Ω; i.e. $\Xi_i = \{C_1 \cup C_2 \cup \cdots \cup C_k \colon k = 0, 1, \ldots, |\Pi_i|, C_1, \ldots, C_k \in \Pi_i\}$.)

[1] We recall that a σ-algebra is a set system closed under complements, countable unions, and countable intersections. Every measure is defined on some σ-algebra.

For example, if $\Omega = \{0,1\}^n$, we can let Ξ_i be the σ-algebra generated by the partition Π_i of Ω induced by the first i coordinates. Each class of Π_i has the form $\{\omega \in \Omega\colon \omega_j = x_j \text{ for } j = 1, 2, \ldots, i\}$ for some $x_1, x_2, \ldots, x_i \in \{0,1\}$.

Next, we need the notion of *conditional expectation*. In the discrete case, if Ξ is a σ-algebra generated by a partition Π, the conditional expectation of a random variable X with respect to Ξ is a random variable that is constant on each class C of Π, and whose value on C equals the average of X over C. For a general, possibly infinite, Ξ, the definition is more complicated.

8.3.1 Definition. *Let (Ω, Σ, P) be a probability space, $\Xi \subset \Sigma$ a σ-algebra and X a random variable on Ω. The conditional expectation of X with respect to Ξ is a random variable Y (usually denoted by $\mathbf{E}[X \mid \Xi]$) that satisfies*

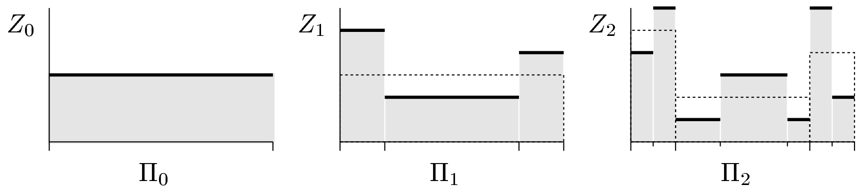
1. *Y is Ξ-measurable.*
2. *For every $B \in \Xi$ with $\mathrm{P}[B] \neq 0$, we have $\mathbf{E}[Y \mid B] = \mathbf{E}[X \mid B]$. Here, for any random variable Z and any event B with $\mathrm{P}[B] > 0$, we write $\mathbf{E}[Z \mid B]$ for $\frac{1}{\mathrm{P}[B]} \int_B Z(\omega)\, \mathrm{dP}(\omega)$.*

In general it is not obvious that Y exists and that it is unique. In our discrete case, though, it is exactly the random variable obtained by averaging over the classes as described above.

Finally, we define a *martingale*. Let $Z_0, Z_1, \ldots$ be a sequence of random variables on Ω, where Z_i is Ξ_i-measurable. In our example with $\{0,1\}^n$, this means that Z_i does not depend on the coordinates $i+1$ through n. The (finite or infinite) sequence $Z_0, Z_1, Z_2, \ldots$ is called a martingale if we have

$$\mathbf{E}[Z_i \mid \Xi_{i-1}] = Z_{i-1}, \quad i = 1, 2, 3, \ldots \quad (8.1)$$

If Ξ_{i-1} and Ξ_i are given by partitions Π_{i-1} and Π_i, respectively, where Π_i refines Π_{i-1}, then Z_i is constant on each class of Π_i and Z_{i-1} is constant on each class of Π_{i-1}. The martingale condition (8.1) means that on each class C of the coarser partition Π_{i-1}, Z_{i-1} is the average of Z_i over all the classes of Π_i that are contained in C. The martingale condition is schematically illustrated below:



The space Ω is indicated as an interval, and the partitions $\Pi_0, \Pi_1, \ldots$ are drawn as partitions into subintervals. The values of Z_i are indicated by the thick lines, and the martingale condition means that the area of each dashed rectangle should equal the total area of the corresponding gray rectangles.

Here is the basic result about concentration of martingales:

8.3.2 Theorem (Azuma's inequality). *Let $Z_0, Z_1, \dots, Z_n = f$ be a martingale on some probability space, and suppose that $|Z_i - Z_{i-1}| \le c_i$ for $i = 1, 2, \dots, n$. Then*

$$\mathrm{P}[f \ge \mathbf{E}[f] + t] \le e^{-t^2/2\sigma^2} \quad \textit{and} \quad \mathrm{P}[X \le \mathbf{E}[f] - t] \le e^{-t^2/2\sigma^2},$$

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

That is, if one can "interpolate" between f and the constant function $\mathbf{E}[f]$ by a martingale with bounded differences, then f is strongly concentrated.

The proof of Azuma's inequality is conceptually similar to that of Theorem 8.2.1, and we omit it (it can be found in the book of Alon and Spencer, for example).

Random variables on product spaces give rise to examples (somewhat trivial) of martingales, as follows. Let $(\Omega_i, \Sigma_i, \mathrm{P}_i)$ be probability spaces, $i = 1, 2, \dots, n$, let $(\Omega, \Sigma, \mathrm{P})$ be their product, and let $f\colon \Omega \to \mathbf{R}$ be a random variable on the product. Let us define a random variable Z_i on Ω: it depends only on the first i coordinates, and for every choice of $x_1 \in \Omega_1, \dots, x_i \in \Omega_i$, we have $Z_i(x_1, \dots, x_i) = \mathbf{E}_{\omega_{i+1},\dots,\omega_n}[f(x_1, \dots, x_n, \omega_{i+1}, \dots, \omega_n)]$. In words, $Z_i(x_1, \dots, x_i)$ is the expectation of $f(\omega)$ when the first i coordinates are fixed to $x_1, \dots, x_i$ and the others are chosen at random. So Z_0 is simply the number $\mathbf{E}[f]$, while $Z_n = f$.

Since Z_i depends only on the first i variables, the appropriate σ-algebra Ξ_i is the one generated by the product of Σ_1 through Σ_i. In order to get used to the notion of a martingale, the reader may want to verify that the Z_i thus defined satisfy the martingale condition $\mathbf{E}[Z_i \,|\, \Xi_{i-1}] = Z_{i-1}$ and that, moreover, if the effect of the ith variable on f is at most c_i, then $|Z_i - Z_{i-1}| \le c_i$. Once this is checked, it becomes clear that Azuma's inequality generalizes Theorem 8.1.1.

A more general example of a martingale, and practically the only type of martingales encountered in combinatorial applications, is obtained as follows. We have some probability space $(\Omega, \Sigma, \mathrm{P})$ (not necessarily a product space), a random variable $f\colon \Omega \to \mathbf{R}$, and a sequence $\Xi_0 = \{\emptyset, \Omega\} \subset \Xi_1 \subset \Xi_2 \cdots \subseteq \Sigma$ of σ-algebras, and we set $Z_i = \mathbf{E}[f \,|\, \Xi_i]$. Such a martingale is used in the next example.

Concentration of Lipschitz functions of a random permutation. Here we illustrate on an important and concrete example how Azuma's inequality allows us to deal with Lipschitz functions on a metric probability space that does not "quite" have a product structure but it is "high-dimensional" in a suitable sense. In other words, we consider Lipschitz functions of many moderately dependent random variables.

Let S_n denote the set of all permutations of $[n]$ (i.e. bijections $[n] \to [n]$). We consider the uniform probability measure on S_n, and we define the distance of two permutations $\pi_1, \pi_2 \in S_n$ as $\rho(\pi_1, \pi_2) = |\{i \in [n]\colon \pi_1(i) \ne \pi_2(i)\}|$.

8.3.3 Theorem. *Let $f\colon S_n \to \mathbf{R}$ be a 1-Lipschitz function. For $\pi \in S_n$ chosen at random and for all $t \ge 0$, we have*

$$\mathrm{P}[f(\pi) \ge \mathbf{E}[f] + t] \le e^{-t^2/8n} \quad \textit{and} \quad \mathrm{P}[f(\pi) \le \mathbf{E}[f] - t] \le e^{-t^2/8n}.$$

Example. Let $I(\pi)$ be the number of *inversions* of a permutation $\pi \in S_n$; i.e. $I(\pi) = |\{(i,j) \in [n]^2 \colon i < j,\ \pi(i) > \pi(j)\}|$. The number of inversions determines the complexity of some sorting algorithms (such as insert-sort), for example. It is easy to check that I is n-Lipschitz. By applying Theorem 8.3.3 on $f(\pi) = \frac{1}{n} I(\pi)$, we get that $I(\pi)$ is concentrated in an interval of length $O(n^{3/2})$ around $\mathbf{E}[I] = \frac{1}{2}\binom{n}{2} \approx \frac{n^2}{4}$.

Proof of Theorem 8.3.3. We define a sequence $\Pi_0, \Pi_1, \ldots, \Pi_{n-1}$ of partitions of S_n, where Π_i is the partition according to the values at $1, 2, \ldots, i$. That is, each class C of Π_i has the form $C = C(a_1, \ldots, a_i) = \{\pi \in S_n \colon \pi(1) = a_1, \ldots, \pi(i) = a_i\}$ for some (distinct) $a_1, \ldots, a_i \in [n]$. In particular, Π_0 has the single class S_n, and Π_{n-1} is the partition into singletons.

Let Ξ_i be the σ-algebra generated by Π_i, and let Z_i be the random variable given by

$$Z_i = \mathbf{E}\left[f(\pi) \mid \Xi_i\right].$$

More explicitly, if π lies in a class C of Π_i, then

$$Z_i(\pi) = \mathrm{ave}_{\sigma \in C}\, f(\sigma) := \frac{1}{|C|} \sum_{\sigma \in C} f(\sigma).$$

The sequence $Z_0, Z_1, \ldots, Z_n$ satisfies the martingale condition (8.1). We want to apply Azuma's inequality 8.3.2, and so we need to bound the differences: we will prove that

$$|Z_i - Z_{i-1}| \le 2. \tag{8.2}$$

We consider a permutation π in some class $C = C(a_1, \ldots, a_{i-1})$ of Π_{i-1}. The value $Z_{i-1}(\pi)$ is the average of f over C. In the partition Π_i, the class C is further partitioned into several classes $C_1, \ldots, C_k$ (in fact, we have $k = n-i+1$), π lies in one of them, say in C_1, and $Z_i(\pi)$ is the average of f over C_1. We thus ask, by how much the average over C_1 can differ from the average over C.

The average over C is the average of the averages over the C_j, $j = 1, 2, \ldots, k$. Thus, it suffices to show that the average over C_{j_1} and the average over C_{j_2} cannot differ by more than 2 (for all j_1, j_2). The reason is that there is a bijection $\varphi \colon C_{j_1} \to C_{j_2}$ such that $\rho(\pi, \varphi(\pi)) \le 2$ for all $\pi \in C_{j_1}$. Indeed, let $C_{j_1} = C(a_1, \ldots, a_{i-1}, b_1)$ and $C_{j_2} = C(a_1, \ldots, a_{i-1}, b_2)$, where b_1 and b_2 are distinct and also different from all of $a_1, \ldots, a_{i-1}$. The bijection φ is defined by the transposition of the values b_1 and b_2: For $\pi \in C_{j_1}$, we set $\varphi(\pi) = \pi'$, where $\pi'(i) = b_2$, $\pi'(\pi^{-1}(b_2)) = b_1$, and $\pi'(j) = \pi(j)$ for $\pi(j) \notin \{b_1, b_2\}$. We have

$$\begin{aligned} |\,\mathrm{ave}_{C_{j_1}} f - \mathrm{ave}_{C_{j_2}} f| &= \left|\mathrm{ave}_{\pi \in C_{j_1}} [f(\pi) - f(\varphi(\pi))]\right| \\ &\le \mathrm{ave}_{\pi \in C_{j_1}} |f(\pi) - f(\varphi(\pi))| \\ &\le 2, \end{aligned}$$

because $\rho(\pi, \varphi(\pi)) \le 2$ and f is 1-Lipschitz.

We have established the bound (8.2) for the martingale differences, and Azuma's inequality 8.3.2 yields Theorem 8.3.3. □

The proof of Theorem 8.3.3 can be generalized to yield concentration results for more general discrete metric probability spaces. The key condition is that such spaces have a suitable sequence of partitions. Some such results can be found, for instance, in

> B. Bollobás: Martingales, isoperimetric inequalities and random graphs, in: *52. Combinatorics, Eger (Hungary)*, Colloq. Math. Soc. J. Bolyai, 1987, pages 113–139.

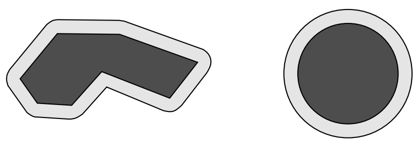
8.4 Isoperimetric Inequalities and Concentration on the Sphere

The method of proof of Theorem 8.2.1 (concentration of Lipschitz functions on product spaces) is suitable for dealing with Hamming-type metrics (or ℓ_1-sums of metrics). To some extent, this is also true for Azuma's inequality and other martingale-based results. Sometimes we need to deal with other "high-dimensional" metric spaces, where the metric is not of a Hamming type; a notable example is various subspaces of $\mathbf{R}^n$ with the Euclidean metric. Here concentration of measure can sometimes be proved by geometric methods. We will consider just one example: measure concentration on the Euclidean sphere.

The Euclidean sphere. Let $S^{n-1} = \{x \in \mathbf{R}^n\colon \|x\| = 1\}$ denote the unit sphere in $\mathbf{R}^n$. We consider it with the Euclidean metric inherited from $\mathbf{R}^n$, and the probability measure P on S^{n-1} is the usual surface measure scaled so that the whole S^{n-1} has measure 1. More formally, for a set $A \subseteq S^{n-1}$, we let $\tilde{A} = \{\alpha x\colon x \in A, \alpha \in [0,1]\}$ be the union of all segments connecting points of A to the center of S^{n-1}, and we set $\mathrm{P}[A] = \lambda^n(\tilde{A})/\lambda^n(B^n)$, where λ^n is the Lebesgue measure in $\mathbf{R}^n$ and B^n denotes the unit ball.

A result about concentration of Lipschitz functions on S^{n-1}, called *Lévy's lemma*, is usually proved via a geometric result, an *isoperimetric inequality.*

Isoperimetric inequalities. The mother of all isoperimetric inequalities states that among all planar geometric figures with a given perimeter, the circular disc has the largest possible area. (This is well-known but not easy to prove rigorously.) In the sense considered here, isoperimetric inequalities claim that among all sets of a given volume in some metric space under consideration, a ball of that volume has the smallest volume of the t-neighborhood:



(In the picture, assuming that the dark areas are the same, then the light gray area is the smallest for the disc.) Letting $t \to 0$, one can get a statement involving the perimeter or surface area. But the formulation with t-neighborhood makes sense even in spaces where "surface area" is not defined.

We note that a ball in the Euclidean metric on S^{n-1} is a spherical cap, that is, an intersection of S^{n-1} with a halfspace. The isoperimetric inequality for the sphere states that for all measurable sets $A \subseteq S^{n-1}$ and all $t \geq 0$, we have

$\mathrm{P}[A_t] \geq \mathrm{P}[C_t]$, where A_t denotes the set of all points of S^{n-1} of distance at most t from A, and where C is a spherical cap with $\mathrm{P}[C] = \mathrm{P}[A]$. This is a rather difficult geometric result; a proof can be found, for example, in

T. Figiel, J. Lindenstrauss, and V. D. Milman: The dimension of almost spherical sections of convex bodies, *Acta Math.*, 139:53–94, 1977.

Let C be a cap of measure $\frac{1}{2}$, that is, a hemisphere. Then C_t is the complement of a cap of height $1-t$, and some calculation (which we omit here) shows that $1 - \mathrm{P}[C_t] \leq 2e^{-t^2 n/2}$. Consequently, by the isoperimetric inequality, we obtain:

8.4.1 Theorem (Measure concentration for the sphere). *Let $A \subseteq S^{n-1}$ be a measurable set with $\mathrm{P}[A] \geq \frac{1}{2}$, and let A_t denote the t-neighborhood of A (in the Euclidean metric). Then*

$$1 - \mathrm{P}[A_t] \leq 2e^{-t^2 n/2}.$$

Thus, if A occupies half of the sphere, almost all points of the sphere lie at distance at most $O(n^{-1/2})$ from A.

We should stress that measure concentration is an exclusively high-dimensional phenomenon; the inequality is practically meaningless for S^2 or S^3, and it becomes interesting only when the dimension is large.

Theorem 8.4.1 speaks about the neighborhoods of sets, while in probabilistic applications, one often needs concentration of Lipschitz functions. The passage to Lipschitz functions is not too difficult. First we need to introduce the median of a function.

Let f be a real random variable (on any probability space; in the discussion below, f is a 1-Lipschitz function $S^{n-1} \to \mathbf{R}$). We define the number $\mathrm{med}(f)$, called the *median* of f, by

$$\mathrm{med}(f) = \sup\{t \in \mathbf{R}\colon \mathrm{P}[f \leq t] \leq \tfrac{1}{2}\}$$

We have $\mathrm{P}[f < \mathrm{med}(f)] \leq \frac{1}{2}$ and $\mathrm{P}[f > \mathrm{med}(f)] \leq \frac{1}{2}$. This is perhaps less obvious than it might seem at first sight. The first inequality can be derived from the σ-additivity of the measure P:

$$\begin{aligned} \mathrm{P}[f < \mathrm{med}(f)] &= \sum_{k=1}^{\infty} \mathrm{P}\left[\mathrm{med}(f) - \tfrac{1}{k-1} < f \leq \mathrm{med}(f) - \tfrac{1}{k}\right] \\ &= \sup_{k \geq 1} \mathrm{P}\left[f \leq \mathrm{med}(f) - \tfrac{1}{k}\right] \leq \tfrac{1}{2}. \end{aligned}$$

The second inequality follows similarly.

Here is the promised result about concentration of Lipschitz functions on the sphere:

8.4.2 Theorem (Lévy's Lemma). *Let $f\colon S^{n-1} \to \mathbf{R}$ be 1-Lipschitz. Then for $t \in [0,1]$,*

$$\mathrm{P}[f \geq \mathrm{med}(f) + t] \leq 2e^{-t^2 n/2} \quad \textit{and} \quad \mathrm{P}[f \leq \mathrm{med}(f) - t] \leq 2e^{-t^2 n/2}.$$

Proof. We prove only the first inequality. Let $A = \{x \in S^{n-1}\colon f(x) \leq \mathrm{med}(f)\}$. By the properties of the median, $\mathrm{P}[A] \geq \frac{1}{2}$. Since f is 1-Lipschitz, we have $f(x) \leq \mathrm{med}(f) + t$ for all $x \in A_t$. Therefore, by Theorem 8.4.1, we get $\mathrm{P}[f(x) \geq \mathrm{med}(f) + t] \leq \mathrm{P}[S^{n-1} \setminus A_t] \leq 2e^{-t^2 n/2}$. □

The median is generally difficult to compute. But for a 1-Lipschitz function, it cannot be too far from the expectation:

8.4.3 Proposition. *Let $f\colon S^{n-1} \to \mathbf{R}$ be 1-Lipschitz. Then*

$$|\,\mathrm{med}(f) - \mathbf{E}[f]\,| \leq 12n^{-1/2}.$$

Proof.

$$\begin{aligned}
|\,\mathrm{med}(f) - \mathbf{E}[f]\,| &\leq \mathbf{E}[|f - \mathrm{med}(f)|] \\
&\leq \sum_{k=0}^{\infty} \tfrac{k+1}{\sqrt{n}} \mathrm{P}\left[|f - \mathrm{med}(f)| \geq \tfrac{k}{\sqrt{n}}\right] \\
&\leq n^{-1/2} \sum_{k=0}^{\infty} (k+1) \cdot 4e^{-k^2/2} \\
&\leq 12n^{-1/2}.
\end{aligned}$$

□

Other important spaces with concentration similar to Theorem 8.4.2 include the n-dimensional torus (the n-fold Cartesian product $S^1 \times \cdots \times S^1 \subset \mathbf{R}^{2n}$) and the group $\mathrm{SO}(n)$ of all rotations around the origin in $\mathbf{R}^n$. Let us remark that results similar to Theorem 8.1.1 (concentration on product spaces) can also be derived from suitable isoperimetric inequalities. For example, if our space is the product $\{0,1\}^n$ with the uniform probability measure, the Hamming cube, then an isoperimetric inequality holds (Harper's inequality, again stating that the ball has the smallest t-neighborhood among all sets of a given measure), and the special case of Theorem 8.1.1 can be derived from it, with a little worse estimate.

Much information about these results and their applications can be found in

J. Lindenstrauss and V. D. Milman: The local theory of normed spaces and its applications to convexity, in *Handbook of Convex Geometry (P.M. Gruber and J. M. Wills eds)*, North-Holland, Amsterdam, 1993, pages 1149–1220.

Let us remark that "functional-theoretic" methods, as opposed to geometric ones, have recently been prominent in new developments in this direction. A thorough treatment of concentration phenomena is the recent book

M. Ledoux: *The Concentration of Measure Phenomenon*, volume 89 of *Mathematical Surveys and Monographs*, Amer. Math. Soc., Providence, RI, 2001.

9

Concentration: Beyond the Lipschitz Condition

9.1 Talagrand's Inequality

Here we enrich our collection of results about concentration by a remarkable result of Talagrand. We begin with a special case, which is easier to state. The setting is similar to that in Theorem 8.1.1: f is a function on a product space $(\Omega, \Sigma, \mathrm{P})$ such that the ith coordinate has effect at most 1.

We say that f *has certificates of size s for exceeding value r* if the following holds. For any $\omega = (\omega_1, \ldots, \omega_n) \in \Omega$ with $f(\omega) \geq r$, there is a subset $I \subseteq [n]$ of at most s indices such that these coordinates alone force the value of f to be at least r: whenever $\omega' \in \Omega$ satisfies $\omega'_i = \omega_i$ for all $i \in I$, we have $f(\omega') \geq r$ as well.

Example: nondecreasing subsequences. Let $(\Omega, \Sigma, \mathrm{P})$ be the product of n intervals $[0,1]$ with the uniform probability measure. For $\omega \in \Omega$, let $f(\omega)$ be the length of a longest nondecreasing subsequence of the sequence $(\omega_1, \omega_2, \ldots, \omega_n)$, i.e. the maximum k such that there are indices $i_1 < i_2 < \cdots < i_k$ with $\omega_{i_1} \leq \omega_{i_2} \leq \cdots \leq \omega_{i_k}$. Clearly, each coordinate has effect at most 1. Moreover, for each $r \geq 0$, f has certificates of size at most r for exceeding the value r (just fix the nondecreasing subsequence).

The following theorem asserts that if f possesses small certificates for exceeding certain values, then it is even more concentrated than an arbitrary 1-Lipschitz function.

9.1.1 Theorem (Talagrand's inequality, special case). *Let $(\Omega_i, \Sigma_i, \mathrm{P}_i)$ be probability spaces, $i = 1, 2, \ldots, n$, let $(\Omega, \Sigma, \mathrm{P})$ be their product, and let $f\colon \Omega \to \mathbf{R}$ be a (measurable) function such that each coordinate has effect at most 1. Let $m = \mathrm{med}(f)$ and let $t \geq 0$. Supposing that f has certificates of size at most s_1 for exceeding the value m, we have*

$$\mathrm{P}[f \leq m - t] \leq 2e^{-t^2/4s_1}.$$

If f has certificates of size at most s_2 for exceeding the value $m + t$, we have

$$\mathrm{P}[f \geq m + t] \leq 2e^{-t^2/4s_2}$$

(note the asymmetry in the lower and upper tail estimates!).

The theorem speaks about deviations from the median, rather than from the expectation. But under suitable conditions, one can show that the median is close to the expectation, by a calculation similar to the proof of Proposition 8.4.3. For example, if $m \geq 1$ and f has certificates of size $O(r)$ for exceeding the value r, for all $r \geq 1$, we get $|\operatorname{med}(f) - \mathbf{E}[f]| = O(\sqrt{\mathbf{E}[f]})$.

Nondecreasing subsequences continued. The length of a longest nondecreasing subsequence satisfies the assumption of Theorem 9.1.1, and we get that it is concentrated around the median m in an interval of length about $\sqrt{m}$. As we will show next, m is about $\sqrt{n}$, and so f is typically concentrated on about $n^{1/4}$ values. Note the power of Talagrand's inequality: for example, Theorem 8.1.1 would give only about $\sqrt{n}$!

For a k-tuple of indices $i_1 < \cdots < i_k$, we have $\mathrm{P}[\omega_{i_1} \leq \cdots \leq \omega_{i_k}] = \frac{1}{k!}$ (by symmetry, all the $k!$ permutations are equally probable). Thus, $\mathrm{P}[f \geq k] \leq \binom{n}{k}\frac{1}{k!} \leq \left(\frac{en}{k}\right)^k \left(\frac{e}{k}\right)^k = \left(\frac{e\sqrt{n}}{k}\right)^{2k}$. So $m \leq 3\sqrt{n}$, say.

To derive a lower bound for m, let $g(\omega)$ be the length of a longest *nonincreasing* subsequence in ω. By symmetry, $\operatorname{med}(g) = \operatorname{med}(f)$. By the Erdős–Szekeres lemma, we always have $f(\omega)g(\omega) \geq n$. Since we know that $\mathrm{P}[f \leq 3\sqrt{n}] \geq \frac{1}{2}$, we get $\mathrm{P}\left[g \geq \frac{1}{3}\sqrt{n}\right] \geq \frac{1}{2}$, and so $m \geq \frac{1}{3}\sqrt{n}$.

A more general form of Talagrand's inequality. Let $(\Omega, \Sigma, \mathrm{P})$ be a product probability space as in Theorem 9.1.1 above. The form of Talagrand's inequality we are going to state next looks like a kind of isoperimetric inequality for this space, but with a little unusual notion of distance $d(\omega, A)$ of a point $\omega \in \Omega$ from a set $A \subseteq \Omega$. We say that a unit vector $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbf{R}^n$, $\|\alpha\| = 1$, with $\alpha_i \geq 0$ for all i, is a *witness* for $d(\omega, A) \geq \tau$ if we have $\sum_{i:\, \omega_i \neq \omega_i'} \alpha_i \geq \tau$ for all $\omega' \in A$. We define $d(\omega, A)$ as the supremum of $\tau \geq 0$ possessing a witness for $d(\omega, A) \geq \tau$.

This definition apparently needs some time to be digested. A helpful example is with $\Omega = \{0,1\}^n$, the cube: here $d(\omega, A)$ turns out to be the distance of ω to the convex hull of A ($\{0,1\}^n$ is interpreted as a subset of $\mathbf{R}^n$).

9.1.2 Theorem (Talagrand's inequality). *Let $A, B \subseteq \Omega$ be two (measurable) sets such that $d(\omega, A) \geq \tau$ for all $\omega \in B$. Then*

$$\mathrm{P}[A]\,\mathrm{P}[B] \leq e^{-\tau^2/4}.$$

The proof is (clever but) not impossibly complicated, but we choose to omit it. It can be found, e.g., in the second edition of the book of Alon and Spencer.

In order to get used to this result, let us derive Theorem 9.1.1 from it.

Proof of Theorem 9.1.1. Let f be as in Theorem 9.1.1, and let $r \geq 0$ be such that f has certificates of size at most s for exceeding the value r. For all $t \geq 0$, we prove

$$\mathrm{P}[f \leq r - t]\,\mathrm{P}[f \geq r] \leq e^{-t^2/4s}; \tag{9.1}$$

this will give both the inequalities in Theorem 9.1.1. Indeed, using it with $r = m$, we obtain $\mathrm{P}[f \leq m - t]\,\mathrm{P}[f \geq m] \leq e^{-t^2/4s_1}$, and the first inequality in the theorem follows using $\mathrm{P}[f \geq m] \geq \frac{1}{2}$. Similarly, the second inequality follows by substituting $r = m + t$.

In order to prove (9.1), we set, not surprisingly, $A = \{\omega \in \Omega\colon f(\omega) \leq r - t\}$ and $B = \{\omega \in \Omega\colon f(\omega) \geq r\}$, and we want to show that for all $\omega \in B$, $d(\omega, A) \geq \tau = \frac{t}{\sqrt{s}}$. Once we succeed in this, we are done.

Fix $\omega \in B$, and let $I \subseteq [n]$, $|I| \leq s$, be the set of indices of a certificate for $f(\omega) \geq r$: any ω' sharing with ω the coordinates indexed by I satisfies $f(\omega') \geq r$. We may assume $I \neq \emptyset$, for otherwise, $f \geq r$ always and $\mathrm{P}[f \leq r - t] = 0$. Let $\alpha \in \mathbf{R}^n$ be the unit vector with $\alpha_i = |I|^{-1/2}$ for $i \in I$ and $\alpha_i = 0$ for $i \notin I$. For $\omega' \in A$, define $\omega'' \in \Omega$ by

$$\omega''_i = \begin{cases} \omega_i & \text{for } i \in I \\ \omega'_i & \text{for } i \notin I. \end{cases}$$

Then $f(\omega'') \geq r$, while $f(\omega') \leq r - t$ since $\omega' \in A$, and so $|f(\omega'') - f(\omega')| \geq t$. Since the effect of each coordinate is at most 1, ω'' and ω' differ in at least t positions (all of which are indexed by I), and ω and ω' also differ in at least t positions indexed by I. So $\sum_{i\colon \omega_i \neq \omega'_i} \alpha_i \geq t|I|^{-1/2} \geq t/\sqrt{s}$. Therefore α witnesses that $d(A, \omega) \geq \frac{t}{\sqrt{s}}$, and (9.1) follows from Theorem 9.1.2. $\square$

Concentration of the largest eigenvalue. This is a neat application of the more general version of Talagrand's inequality (Theorem 9.1.2). Let $\mathcal{M}$ denote the probability space of all symmetric matrices $M = (m_{ij})_{i,j=1}^n$, where $m_{ii} = 0$ for all i, the entries m_{ij} for $1 \leq i < j \leq n$ are chosen independently and uniformly at random in the interval $[0, 1]$, and those with $i > j$ are defined by symmetry. Formally, $\mathcal{M}$ can be identified with the product space $[0, 1]^m$, where $m = \binom{n}{2}$ is the number of entries of M above the diagonal. (The argument below works, with small changes, for many other distributions of the m_{ij}; the selected example gives particularly simple calculations.) As linear algebra teaches us, each $M \in \mathcal{M}$ has n real eigenvalues $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$. We derive a very strong concentration result for λ_n. (Eigenvalues of random matrices, significant in many applications, are usually quite difficult to handle.)

We use the following well-known characterization of λ_n:

$$\lambda_n = \max\{x^T M x\colon x \in \mathbf{R}^n,\ \|x\| = 1\}.$$

First we determine the order of magnitude of $\mathbf{E}[\lambda_n]$. On the one hand, setting $u = (\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}})$, we have

$$\mathbf{E}[\lambda_n] \geq \mathbf{E}\left[u^T M u\right] = \frac{1}{n} \sum_{i,j=1}^{n} \mathbf{E}[m_{ij}] = \frac{1}{2n}(n^2 - n) = \frac{1}{2}(n - 1).$$

On the other hand, for any M and any unit vector x, we have, by the Cauchy–Schwarz inequality,

$$\begin{aligned} x^T M x &= \sum_{i,j} x_i x_j m_{ij} \\ &\leq \Bigl(\sum_{i,j} x_i^2 x_j^2\Bigr)^{1/2} \Bigl(\sum_{i,j} m_{ij}^2\Bigr)^{1/2} \end{aligned}$$

$$
\begin{aligned}
&= \left(\sum_{i,j} m_{ij}^2\right)^{1/2} \\
&= \|M\|_2.
\end{aligned}
$$

We estimate $\mathbf{E}\left[\|M\|_2\right]^2 \leq \mathbf{E}\left[\|M\|_2^2\right] = \sum_{ij} \mathbf{E}\left[m_{ij}^2\right] = \frac{1}{3}(n^2 - n)$, and so

$$\tfrac{1}{2}(n-1) \leq \mathbf{E}\left[\lambda_n\right] \leq \tfrac{1}{\sqrt{3}} n.$$

Now we start with the concentration result. For numbers r and $t \geq 0$, let $A \subseteq \mathcal{M}$ be the set of all matrices with $\lambda_n \leq r$, and let $B \subseteq \mathcal{M}$ consist of those matrices with $\lambda_n \geq r + t$. We want to show that for all $M \in B$, we have $d(M, A) \geq \frac{t}{2}$, where $d(\omega, A)$ is as in Talagrand's inequality. Since $M \in B$, there is a unit vector $x = x(M)$ with $x^T M x \geq r + t$. On the other hand, for any $N \in A$, we have $x^T N x \leq r$. We calculate

$$t \leq x^T M x - x^T N x = \sum_{1 \leq i < j \leq n} 2 x_i x_j (m_{ij} - n_{ij}) \leq \sum_{\substack{1 \leq i < j \leq n \\ m_{ij} \neq n_{ij}}} 2|x_i x_j|.$$

This suggests an appropriate choice for a vector $\alpha = (\alpha_{ij})_{1 \leq i < j \leq n}$ witnessing $d(M, A) \geq \frac{t}{2}$. Namely, letting $\beta_{ij} = 2|x_i x_j|$, we find

$$\|\beta\|^2 = 4 \sum_{i<j} x_i^2 x_j^2 \leq 2 \left(\sum_{i=1}^n x_i^2 \right)^2 = 2,$$

, and so for $\alpha = \frac{\beta}{\|\beta\|}$, we have

$$\sum_{i<j:\, m_{ij} \neq n_{ij}} \alpha_{ij} \geq \tfrac{1}{2} \sum_{i<j:\, m_{ij} \neq n_{ij}} 2|x_i x_j| \geq \tfrac{t}{2}.$$

The assumptions of Talagrand's inequality 9.1.2 are satisfied for A and B with $\tau = \frac{t}{2}$, and we obtain

$$\mathrm{P}[A]\, \mathrm{P}[B] \leq e^{-t^2/16}.$$

Setting r to the median $m = \mathrm{med}(\lambda_n)$, we have $\mathrm{P}[A] = \frac{1}{2}$, and so

$$\mathrm{P}[\lambda_n \geq m + t] \leq 2e^{-t^2/16}.$$

Letting $r = m - t$, we get $\mathrm{P}[B] = \frac{1}{2}$ and so

$$\mathrm{P}[\lambda_n \leq m - t] \leq 2e^{-t^2/16}.$$

Thus, λ_n is concentrated in an interval of length only $O(1)$ around the median! Further calculation, similar to the proof of Proposition 8.4.3, shows that $|m - \mathbf{E}[\lambda_n]| = O(1)$.

9.2 The Vu–Kim Inequality

Even sophisticated concentration inequalities for Lipschitz functions are useless if the investigated function is not Lipschitz enough. Of course, often this may be simply because the function is not concentrated, and this possibility should not be overlooked. But sometimes there still is a concentration result, and the rather complicated-looking inequality presented in this section may help.

As a running example, let T be the number of triangles in the random graph $G(v,p)$. (We use v instead of the usual n, since n will be reserved for the number of variables in the considered function). We already studied this random variable in Section 5.3, where we showed that for $p >> \frac{1}{v}$ (we recall that this notation means $\frac{p}{v} \to \infty$), $G(v,p)$ almost surely contains a triangle. Here we let $p = \frac{\varphi(v)}{v}$, where $\varphi(v) \to \infty$ as $v \to \infty$ but not very fast (say, $\varphi(v) = v^{1/9}$).

Formally, T is a real function on the space $\{0,1\}^n$ with $n = \binom{v}{2}$ and with the appropriate product measure. By adding a single edge we can create as many as $v-2$ new triangles, and so the effect of each variable on T is at least $v-2$ (in fact, it equals $v-2$). We have $\mathbf{E}[T] = \binom{v}{3}p^3 = \Theta(\varphi(v)^3)$, while $\sigma = (\sum_{i=1}^n c_i^2)^{1/2}$ in Theorem 8.1.1 is $\Theta(v^2)$. If we want the bound $e^{-t^2/2\sigma^2}$ in that theorem to be meaningful for deviations t comparable to $\mathbf{E}[T]$ or smaller, we would need $\varphi(v)$ as large as $v^{2/3}$! Neither Talagrand's inequality seems to be helpful in this situation.

Yet T is much more concentrated than these results indicate. The intuitive reason is that the situation where one edge is contained in very many triangles is extremely rare. For instance, the *expected* number of triangles containing a given edge is only $(v-2)p^2 = \Theta(\varphi(v)^2/v)$, which is quite small. Formalizing this intuition is not so easy. The expected effect of each variable being small is generally not sufficient for concentration. This is illustrated by the next example, which also introduces us to the realm of multivariate polynomials, where we will stay for the rest of this section.

9.2.1 Example. Let $n = 4k$, and for $t = (t_1,\ldots,t_n) \in \{0,1\}^n$, let us define

$$f(t) = (t_1t_2 + t_3t_4 + \cdots + t_{2k-1}t_{2k})(t_{2k+1} + t_{2k+2} + \cdots + t_{4k}).$$

Suppose that each t_i independently attains value 1 with probability $p = n^{-1/2}$ and value 0 with probability $1-p$ (in other words, f is considered on $\{0,1\}^n$ with a suitable product probability measure). By multiplying the parentheses as polynomials and using the linearity of expectation, we find $\mathbf{E}[f] = 2k^2p^3 = n^{1/2}/8$. What is the expected effect of t_i? If, for example, t_2 through t_n are chosen at random, then the expected effect of changing t_1 from 0 to 1 or back is $\mathbf{E}[t_2(t_{2k+1} + \cdots + t_{4k})] = 2kp^2 = \frac{1}{2}$, and similarly for $t_2,\ldots,t_{2k}$. The expected effect of t_{2k+1} through t_{4k} is even much smaller.

Yet f is not concentrated at all! Indeed, using a Chernoff-type inequality (such as Theorem 7.2.1), we see that the sum $t_{2k+1} + \cdots + t_{4k}$ in the second parenthesis is close to $n^{1/2}/4$ with high probability. The first parenthesis, $(t_1t_2 + t_3t_4 + \cdots + t_{2k-1}t_{2k})$, is always an integer, and so with high probability, f is either 0 or at least about $2\mathbf{E}[f]$.

Vu and Kim have developed a machinery for proving concentration of functions f that are "mostly" Lipschitz but not quite, such as T in our running example. We briefly describe the setting and state one of their concentration inequalities, reasonably general but not the most general available.

To apply the result, we need to suppose that f is defined on the product of some probability spaces $(\Omega_i, \Sigma_i, \mathrm{P}_i)$, $i = 1, 2, \ldots, n$, where each Ω_i is a subset of the interval $[0, 1]$. A typical example is $\Omega_i = \{0, 1\}$. We also need f to be expressible (or approximable) by a suitable polynomial. More precisely, we assume that there is a polynomial $\bar{f} = \bar{f}(t_1, t_2, \ldots, t_n) \in \mathbf{R}[t_1, \ldots, t_n]$ with *all coefficients lying in* $[0, 1]$ such that $f(t) = \bar{f}(t)$ for all $t = (t_1, \ldots, t_n) \in \Omega$.

Exotic as this condition might sound, it is often naturally fulfilled in combinatorial applications. In our running example with the number of triangles in $G(v, p)$, we have one indicator variable $t_{ij} \in \{0, 1\}$ for each pair $\{i, j\} \in \binom{[v]}{2}$ of vertices, and

$$T = \sum_{\{i,j,k\} \in \binom{[v]}{3}} t_{ij} t_{jk} t_{ik}. \tag{9.2}$$

If f cannot be written as a suitable polynomial, it is sometimes possible to choose another function $\tilde{f}$ that can be so expressed and approximates f. Then one can apply the result below to show concentration for $\tilde{f}$, and infer that f, being close to $\tilde{f}$, is concentrated as well. (Let us remark that some of the results below can also be directly extended to some functions other than polynomials; see the reference given below.)

In the sequel, we will not formally distinguish between f (which is defined on Ω) and the polynomial $\bar{f}$ that extends f to the whole $[0, 1]^n$. We thus assume that f is a real polynomial defined on $[0, 1]^n$. However, all random choices of the variables t_i are according to the distribution given by Ω. In particular, values of t_i not lying in Ω_i have zero probability.

The Vu–Kim inequality asserts that an f as above is concentrated provided that the expectation of each partial derivative of f up to some fixed order $\ell - 1$ is sufficiently small, and the maximum of all partial derivatives of order ℓ or larger is small as well.

Namely, for the polynomial f as above and an j-term sequence $I = (i_1, i_2, \ldots, i_j)$ of indices, let

$$\partial_I f = \frac{\partial^j f}{\partial t_{i_1} \partial t_{i_2} \cdots \partial t_{i_j}}$$

(this is again a real function on $[0, 1]^n$). Further we let

$$M_\ell = M_\ell(f) = \sup_{t \in \Omega,\, |I| \geq \ell} \partial_I f(t),$$

where $|I|$ is the length of the sequence I, and

$$E_j = E_j(f) = \max_{|I| = j} \mathbf{E}\left[\partial_I f\right]$$

The expectation is with respect to a random $t \in \Omega$; in particular, $E_0 = \mathbf{E}[f]$. Heuristically, $E_j(f)$ can be interpreted as the maximum average effect on f of any group of j variables, and $M_\ell(f)$ corresponds to the maximum effect of any group of ℓ variables.

In our running example, with the polynomial T given by (9.2), the degree of T in each variable is 1, and so it suffices to consider sequences I with at most 3 terms, all distinct. We have $\partial T/\partial t_{12} = \sum_{i>1} t_{1i}t_{2i}$, and so $E_1(T) = (v-2)p^2 = \Theta(\varphi(v)^2/v)$ (exactly what we calculated before!). Further, $\partial T/\partial t_{12}\partial t_{23} = t_{13}$, and similarly for all other pairs of edges sharing a vertex, while all the other partial derivatives of order 2 are 0. Therefore, $E_2(T) = 1 = M_2(T)$. Finally, $E_3(T) = M_3(T) = 1$; note that $M_\ell(f) \le 1$ for any polynomial f of degree at most ℓ with all coefficients in $[0,1]$.

Here is the promised inequality.

9.2.2 Theorem (Vu–Kim inequality). *Let* $\mathrm{P}_1, \mathrm{P}_2, \ldots, \mathrm{P}_n$ *be probability measures on* $[0,1]$*, and let* P *be the product measure on* $[0,1]^n$*. Let* $f\colon [0,1]^n \to \mathbf{R}$ *be a function given by an* n*-variate real polynomial with all coefficients lying in* $[0,1]$*. Let* $\ell \ge 1$ *be a fixed integer, suppose that* $M_\ell(f) \le 1$*, and for* $j = 1, 2, \ldots, \ell-1$*, let* $E_j = E_j(f)$ *be as above. Let* $\tau \ge \sqrt{\log n}$ *be a parameter, and set*

$$\begin{aligned}\mathcal{E}_1 &= \max\left(E_1, \tau^2 E_2, \tau^4 E_3, \ldots, \tau^{2(\ell-2)} E_{\ell-1}, \tau^{2(\ell-1)}\right),\\ \mathcal{E}_0 &= \max\left(E_0, \tau^2 \mathcal{E}_1\right).\end{aligned}$$

Then

$$\mathrm{P}\left[\left|f - \mathbf{E}\left[f\right]\right| \ge a\tau\sqrt{\mathcal{E}_0\mathcal{E}_1}\right] \le be^{-\tau^2},$$

where a *and* b *are suitable positive constants depending only on* ℓ*.*

If the quantities E_j decrease sufficiently fast, namely, if $E_j/E_{j+1} \ge \tau^2$ for all $j = 0, 1, \ldots, \ell-1$, then $\mathcal{E}_0 = E_0$, $\mathcal{E}_1 = E_1$, and $\sqrt{\mathcal{E}_0\mathcal{E}_1}$ is independent of τ (in the appropriate range of τ). In such a case, the concentration is of the usual Gaussian type (as in most of the inequalities mentioned earlier). But often we get only weaker bounds; this is the case for our running example.

In that example, we have $M_2(T) \le 1$, and so we can choose $\ell = 2$. As was noted above, $E_0(T) = \Theta(\varphi(v)^3)$ and $E_1(T) = \Theta(\varphi(v)^2/v)$. Since we assume $\varphi(v) << v^{1/2}$, we obtain $\mathcal{E}_1 = \max(E_1, \tau^2) = \tau^2$ and $\mathcal{E}_0 = \max(E_0, \tau^2 E_1) = \max(\varphi(v)^3, \tau^4)$. If we use the concrete value $\varphi(v) = v^{1/9}$ and consider only the τ with $\tau^4 \le \varphi(v)^3$, the resulting inequality is

$$\mathrm{P}\left[\left|T - \mathbf{E}\left[T\right]\right| \ge a\tau^4\right] \le be^{-\tau^2}, \qquad \sqrt{2\log v} \le \tau \le v^{1/12}.$$

Rewritten in the parameterization by the deviation t used in the inequalities in the preceding sections, this becomes

$$\mathrm{P}\left[\left|T - \mathbf{E}\left[T\right]\right| \ge t\right] \le e^{-\alpha\sqrt{t}}, \quad c_1(\log v)^2 \le t \le c_2 v^{1/3}$$

for suitable positive constants α, c_1, c_2. Such kind of result is typical for applications of the Vu–Kim inequality; in some range of deviations, from logarithmically small to a small power of n, we obtain an exponentially decreasing bound. The exact values of the exponents seldom matter much.

Let us remark that the "obvious" first choice of ℓ in this example is 3, the degree of T. Then $M_3 \le 1$ is automatic, but the resulting bound is quantitatively worse, as the reader may want to check.

There are other techniques that yield concentration results for the quantity T and in some similar situations (for example, the Janson–Suen inequality—Theorem 7.4.3). But the Vu–Kim inequality currently appears as the most general and flexible tool, and in several applications it provides the only known path to the goal.

Theorem 9.2.2 does not cover deviations of logarithmic order, and so it typically does not work very well for functions with logarithmic or smaller expectations (for $\varphi(v)$ much smaller than $\log v$ in our running example, say). There are more precise versions covering such situations as well. These and other variations, as well as a proof of Theorem 9.2.2 and further applications of it, can be found in the survey paper

V. H. Vu: Concentration of non-Lipschitz functions and applications, *Random Structures & Algorithms*, 2002, in press.