The Probabilistic Method

Spring School 2000

Jiří Matoušek Jan Vondrák

Contents

Preface

This is an extract from "Lecture notes on the probabilistic method" which is a text accompanying the lecture taught by J. Matoušek at Charles University. Only the basic techniques and examples are described here. For more information, the reader is invited to refer to

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

which an extensive and modern book on this subject which served as the basis for both the lecture and this text. A large part of the material here is taken directly from this book, sometimes with a little different presentation.

The techniques are 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, MATFYZPRESS 1996).

A more advanced source is

> S, Janson, T. Łuczak, A. Ruciński: *Topics in random graphs*, J. Wiley & Sons, 2000.

Finally, 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.

1

The Probabilistic Method

The probabilistic method is a remarkable technique based on the theory of probability which, surprisingly, serves as a tool in proofs of theorems which 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, the 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.

1.1 Ramsey numbers

The Ramsey theorem states that any large enough graph contains either a clique or an independent set of a given size. (A *clique* is a set of vertices inducing a complete subgraph and an *independent set* is a set of vertices inducing an empty subgraph.)

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

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

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 $R(k,k)$.

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

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

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 6.1.3), and we get

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

If we choose $n = \lfloor 2^{k/2} \rfloor$, we have

$$2\binom{n}{k}2^{-\binom{k}{2}} \leq 2\frac{n^k}{k!}2^{k/2-k^2/2} = \left(\frac{n}{2^{k/2}}\right)^k \frac{2^{k/2+1}}{k!} \leq \frac{2^{k/2+1}}{k!}.$$

The last fraction decreases asymptotically to zero, and as the reader can check, for $k = 3$ it is already less than 1. Thus for $k \geq 3$, the probability that a random graph on n vertices contains either a clique or an independent set of size k is strictly less than 1. This implies that in some graphs on n vertices neither of the two appears, i.e.

$$R(k,k) > n = \lfloor 2^{k/2} \rfloor.$$

□

One might object that the use of a probability space is artificial here and the same proof can be formulated in terms of good and bad 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 non-empty. In simple cases, it is indeed possible to phrase the proof 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

Linearity of Expectation

2.1 Computing expectation using indicators

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

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

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

Proof.

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

□

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

$$\mathbf{E}\left[X\right] = \mathbf{E}\left[X_1\right] + \mathbf{E}\left[X_2\right] + \ldots + \mathbf{E}\left[X_n\right].$$

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

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

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

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

Proof.

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

□

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 $\{1, \dots, n\}$. 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

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

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

2.2 Splitting Graphs

We can always 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 $\omega \in \Omega$.

2.2.1 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 $p = \frac{1}{2}$. 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

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

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

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

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. $\square$

3

The Second Moment

3.1 Variance and the Chebyshev Inequality

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

3.1.1 Definition (Variance). *The variance of 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 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.

3.1.2 Definition. *The covariance of two random variables is*

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

3.1.3 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.

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

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

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

□

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* to estimate the probability that a random variable deviates from its expectation at least by a given number.

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

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

Proof.

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

□

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 4, we will examine stronger methods giving better bounds for certain classes of random variables. In this section, though, the Chebyshev inequality will be sufficient.

3.2 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.

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

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 $m+k, |k| < \sqrt{m}$, 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

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

and the proposition follows. □

3.3 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 $p(n) << \frac{1}{n}$. 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 large enough, the random graph contains *some* triangle almost surely.

3.3.1 Lemma. *Consider a sequence of non-negative random variables $X_1, X_2, \ldots$ 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 $t = \frac{1}{2}\mathbf{E}[X_n]$ in the Chebyshev inequality:

$$\mathrm{P}\left[|X_n - \mathbf{E}[X_n]| \geq \tfrac{1}{2}\mathbf{E}[X_n]\right] \leq \frac{4\,\mathrm{Var}[X_n]}{(\mathbf{E}[X_n])^2}$$

and we get

$$\lim_{n\to\infty} \mathrm{P}\left[X_n = 0\right] \leq \lim_{n\to\infty} \mathrm{P}\left[X_n \leq \tfrac{1}{2}\mathbf{E}\left[X_n\right]\right] \leq \lim_{n\to\infty} \frac{4\,\mathrm{Var}\left[X_n\right]}{(\mathbf{E}\left[X_n\right])^2} = 0.$$

□

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

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

For every triangle

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

and for a pair of triangles sharing an edge

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

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

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

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

which tends to zero if $p(n) >> \frac{1}{n}$. Lemma 3.3.1 implies that in such a case, the probability that $G_{n,p}$ contains a triangle approaches 1 as n tends to infinity.

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*.

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

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

(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 *, too).*

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 would be much more difficult). It turns out that our approach can be extended to any subgraph H that is *balanced.*

3.3.3 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 that H itself.*

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

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

is the 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 clearly 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:

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

First,

$$\sum_\beta \mathrm{Var}[X_\beta] \leq O(n^v)\mathbf{E}\left[X_\beta^2\right] = O(n^v p^e).$$

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

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

The number of pairs β, γ sharing t vertices is $O(n^{2v-t})$ because we can choose the base 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 base set. For a fixed t, we get

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

For the variance of X, we get

$$\mathrm{Var}\left[X\right] \leq O(n^v p^e) + \sum_{t=2}^{v-1} O((n^v p^e)^{2-t/v})$$

and

$$\lim_{n\to\infty} \frac{\mathrm{Var}\left[X\right]}{(\mathbf{E}\left[X\right])^2} \leq \lim_{n\to\infty} \left(O((n^v p^e)^{-1}) + \sum_{t=2}^{v-1} O((n^v p^e)^{-t/v}) \right) = 0$$

since $\lim_{n\to\infty} n^v p^e = \infty$. This completes the second part of the proof because by Lemma 3.3.1,

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

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 maximal density $\rho(H')$. We give here only the result without a proof.

3.3.5 Theorem. *Let H be a graph and $H' \subset H$ its subgraph with maximal density $\rho(H')$. Then*

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

is the threshold function for the event that H is a subgraph of $G_{n,p}$.

4

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*. 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 $\frac{n}{2} + O(\sqrt{n \log n})$.

4.1 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.

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

We have $\mathrm{P}[X \geq t] = \mathrm{P}[Y \geq e^{ut}]$. Markov's inequality tells us that $\mathrm{P}[Y \geq q] \leq \mathbf{E}[Y]/q$. We have

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

(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

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

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 4.1.1 is proved. $\square$

Combinatorial discrepancy. We show a nice application. Let X be an n-point set, and let $\mathcal{S}$ 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 $\mathcal{S}$ contains approximately the same number of red and blue points (we want a "balanced" coloring). The *discrepancy* of the set system $\mathcal{S}$ measures how well this can be done. 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 $S \in \mathcal{S}$ is just $\chi(S) = \sum_{x\in S} \chi(x)$. The discrepancy $\mathrm{disc}(\mathcal{S},\chi)$ of $\mathcal{S}$ under the coloring χ is $\max_{S\in\mathcal{S}} |\chi(S)|$, and the discrepancy of $\mathcal{S}$ is the minimum of $\mathrm{disc}(\mathcal{S},\chi)$ over all χ.

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

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

Proof. 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 4.1.1 tells us that

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

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 $S \in \mathcal{S}$ simultaneously. □

5

Concentration of Lipschitz functions

5.1 Lipschitz functions of independent variables

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. Here we present a strong concentration result for random variables of the more general form $f(X_1, X_2, \ldots, X_n)$ for a "nice" function f of n variables. The condition that none of the X_i be too big that was needed in concentration results for sums is replaced by requiring that changing any single X_i cannot influence the value of f by too much.

Recall that a function f from a metric space M_1 with metric ρ_1 into a metric space M_2 with metric ρ_2 is called *K-Lipschitz* if $\rho_2(f(x), f(y)) \leq K\rho_1(x,y)$ for all $x, y \in M_1$. In our particular case, suppose that the random variable X_i attains values in a set R_i, and so f is a function $R_1 \times R_2 \times \cdots \times R_n \to \mathbf{R}$. On $\mathbf{R}$, we consider the usual metric, and on $R_1 \times \cdots \times R_n$, the distance of two vectors $x = (x_1, x_2, \ldots, x_n)$ and $y = (y_1, y_2, \ldots, y_n)$ is the number of coordinates in which they differ, i.e. $|\{i \in [n] \colon x_i \neq y_i\}|$ (the *Hamming distance* of x and y). Thus, f is K-Lipschitz if $|f(x) - f(y)| \leq K$ for all x, y that differ in a single coordinate. Sometimes one of the coordinates has greater influence on f than the others; then it is useful to measure the maximum possible effect of each coordinate separately. We thus say that *the ith coordinate has effect at most c_i* for f if $|f(x) - f(y)| \leq c_i$ for all x, y that differ only in the ith coordinate.

5.1.1 Theorem. *Let $X_1, X_2, \ldots, X_n$ be independent random variables, X_i attaining values in a set R_i, and let $f \colon R_1 \times \cdots \times R_n \to \mathbf{R}$ be a function such that the ith coordinate has effect at most c_i, $i = 1, 2, \ldots, n$. Then the random variable $X = f(X_1, X_2, \ldots, X_n)$ satisfies, for any $t > 0$,*

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

where $\sigma^2 = \sum_{i=1}^n c_i^2$. In particular, if f is 1-Lipschitz, then

$$\mathrm{P}[X \geq \mathbf{E}[X] + t] \leq e^{-t^2/2n} \quad \text{and} \quad \mathrm{P}[X \leq \mathbf{E}[X] - t] \leq e^{-t^2/2n}.$$

Thus, a 1-Lipschitz function of n independent random variables is concentrated at least as much as the sum of n independent ± 1 random variables.

We postpone the proof of Theorem 5.1.1, which uses a sophisticated and useful notion from probability theory, to the next part, and we now show applications of this very 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. Theorem 5.1.1 implies that X is strongly concentrated around $\mathbf{E}[X]$: $\mathrm{P}[|X - \mathbf{E}[X]| \geq t] \leq 2e^{-t^2/2n}$. Indeed, let $R_i = [n]$ and $X_i = g(i)$. Changing the value of $g(i)$ changes the size of the image of g by at most 1, and so X is 1-Lipschitz and Theorem 5.1.1 applies.

Concentration of the chromatic number. Let X be the chromatic number of the random graph $G_{n,p}$ (for some given n and p). It is not at all easy to determine $\mathbf{E}[X]$ (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 5.1.1!

In order to apply Theorem 5.1.1, we need to express X as a function of independent random variables. A first natural attempt might be to consider, for each potential edge $e = \{u, v\}$, the indicator random variable X_e for the presence of e in $G_{n,p}$. Our X is a 1-Lipschitz function of these X_e, but their number is too large: the n in Theorem 5.1.1 would be $\binom{n}{2}$ and, since X is in the range $[1, n]$, the concentration result would be useless. The trick is to group the X_e into larger chunks. Namely, let $v_1, v_2, \ldots, v_n$ be the vertices of $G_{n,p}$ enumerated in a fixed order, and let X_i be the *vector* of $n - i$ zeros and ones, indicating the presence or absence of the edges going from v_i to $v_{i+1}, v_{i+2}, \ldots, v_n$. That is, $R_i = \{0,1\}^{n-i}$ and

$$X_i = (X_{\{v_i,v_{i+1}\}}, X_{\{v_i,v_{i+2}\}}, \ldots, X_{\{v_i,v_n\}}).$$

These X_i, $i = 1, 2, \ldots, n-1$, are independent, and the chromatic number is a 1-Lipschitz function in the X_i, because changing the edges incident to a single vertex changes the chromatic number of a graph by at most 1. Theorem 5.1.1 now gives

5.1.2 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 p not too large, one of at most 4 values is attained most of the time:

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

$$\mathrm{P}[\chi(G_{n,p}) \notin \{u, u+1, u+2, u+3\}] \to 0$$

as $n \to \infty$.

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.

5.1.4 Lemma. *Let $\alpha > \frac{5}{6}$, $p = n^{-\alpha}$. 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 $t \le \sqrt{8n \ln n}$ vertices with average degree at least 3. This suffices since a vertex-minimal 4-chromatic subgraph must have all degrees at least 3. First, let $t \ge 4$ 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 $\binom{n}{k} \le (en/k)^k$)

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

There are $\binom{n}{t} \le (ne/t)^t$ 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

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

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 5.1.3. 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. This X is a 1-Lipschitz function of the independent random variables $X_1, X_2, \ldots, X_{n-1}$ as in the proof of the Shamir–Spencer theorem 5.1.2 above (right?). We thus have the tail estimates from Theorem 5.1.1:

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

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 $\mathrm{P}[X \ge 2t] \le \mathrm{P}[X \le \mathbf{E}[X] + t] \le \frac{1}{n}$. So with probability at least $1 - \frac{1}{n}$, $G_{n,p}$ with some $2t$ vertices removed can be u-colored. By Lemma 5.1.4, 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. □

5.2 Proof and martingales

Here we prove Theorem 5.1.1. The proof resembles the proof of Theorem 4.1.1. There we needed the inequality $\mathbf{E}\left[e^{uX_i}\right] \le e^{u^2/2}$, where u is a real parameter and X_i attains values -1 and $+1$ with probability $\frac{1}{2}$. Here we will use an analogous estimate for a more general random variable.

5.2.1 Lemma. *Let Y be a random variable attaining values in the interval $[-1,1]$ and with $\mathbf{E}[Y]=0$. Then for any real parameter $u \ge 0$ we have*

$$\mathbf{E}\left[e^{uY}\right] \le e^{u^2/2}.$$

Proof. 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 $h(x) \ge e^{ux}$ holds for all $x \in [-1,1]$ (use Taylor series). So

$$\mathbf{E}\left[e^{uY}\right] \le \mathbf{E}[h(Y)] = \mathbf{E}[Y]\sinh u + \cosh u = \cosh u \le e^{u^2/2}$$

□

Proof of Theorem 5.1.1. For simpler notation, we prove the theorem with $c_1 = c_2 = \cdots = 1$ (the general case is similar). Let $X_1, X_2, \ldots, X_n$ be the independent random variables as in the theorem, X_i attaining values in R_i, and let $f\colon R_1 \times \cdots \times R_n \to \mathbf{R}$ be 1-Lipschitz.

We define a sequence $Z_0, Z_1, \ldots, Z_n$ of random variables, where Z_i depends on $X_1, X_2, \ldots, X_i$. First we define functions $z_i\colon R_1 \times \cdots \times R_i \to \mathbf{R}$ by

$$z_i(x_1, \ldots, x_i) = \mathbf{E}_{X_{i+1},\ldots,X_n}\left[f(x_1, \ldots, x_i, X_{i+1}, \ldots, X_n)\right].$$

The expectation on the right-hand side is with respect to X_{i+1} through X_n chosen independently at random (while the first i variables are fixed to the specific values $x_1, \ldots, x_i$); this is indicated by the subscript at the expectation operator $\mathbf{E}[.]$. Now if $X_1, \ldots, X_i$ are random, $Z_i = z_i(X_1, \ldots, X_i)$ becomes a random variable. In particular, Z_n is the same as X and Z_0 is simply the expectation $\mathbf{E}[X]$ (a single number). Without loss of generality, we may assume $Z_0 = \mathbf{E}[X] = 0$ (for otherwise we work with the new variable $X - \mathbf{E}[X]$).

We need the following property of the Z_i:

$$|Z_i - Z_{i-1}| \le 1, \quad i = 1, 2, \ldots, n \tag{5.1}$$

(recall that Z_i depends on $X_1, \ldots, X_i$ and Z_{i-1} on $X_1, \ldots, X_{i-1}$, and the inequality holds for all possible values of these variables). Roughly speaking, this is because Z_{i-1} is Z_i averaged over all values of X_i, and changing X_i while keeping the other variables fixed changes the value of f by at most 1. For a more precise proof, we need the following property of independent random variables: If U and V are some indepenent random variables and $g(U,V)$ is a function of U and V, we have

$$\mathbf{E}_{U,V}[g(U,V)] = \mathbf{E}_U\left[\mathbf{E}_V[g(U,V)]\right]. \tag{5.2}$$

On the left-hand side, we take the expectation for both U and V chosen at random, while on the right-hand side, we first take expectation with respect to random V for each fixed value of U, obtaining a function of U, and then we take its expectation with respect to a random U. The proof is simple, and for a finite probability space, it is very similar to the proof of Lemma 6.1.9. Returning to the proof of (5.1), we have

$$\begin{aligned}
& z_i(x_1,\dots,x_i) - z_{i-1}(x_1,\dots,x_{i-1}) \\
= \; & \mathbf{E}_{X_{i+1},\dots,X_n}\left[f(x_1,\dots,x_i,X_{i+1},\dots,X_n)\right] \\
& -\mathbf{E}_{X_i,\dots,X_n}\left[f(x_1,\dots,x_{i-1},X_i,\dots,X_n)\right] \\
= \; & \mathbf{E}_{X_{i+1},\dots,X_n}\Big[f(x_1,\dots,x_i,X_{i+1},\dots,X_n) \\
& -\mathbf{E}_{X_i}\left[f(x_1,\dots,x_{i-1},X_i,X_{i+1},\dots,X_n)\right]\Big]
\end{aligned}$$

by (5.2). For any choice of values of $X_{i+1},\dots,X_n$, the value of $f(x_1,\dots,x_i,X_{i+1},\dots,X_n)$ is fixed, while $\mathbf{E}_{X_i}\left[f(x_1,\dots,x_{i-1},X_i,X_{i+1},\dots,X_n)\right]$ is an average over all choices of X_i with the values of all the other variables fixed. Since the effect of the ith variable is at most 1, this average is no more than 1 away from $f(x_1,\dots,x_i,X_{i+1},\dots,X_n)$, and (5.1) follows.

As in the proof of Theorem 4.1.1, we want to estimate $\mathbf{E}\left[e^{uX}\right]$ (this will give one of the tail estimates, and the other one follows by considering $-X$ instead of X). By induction on i, we prove that

$$\mathbf{E}\left[e^{uZ_i}\right] \le e^{iu^2/2}.$$

Suppose that this has been proved up to $i-1$, and put $Y = Z_i - Z_{i-1}$. By (5.1), Y attains values in the interval $[-1,1]$. Recalling that Z_i only depends on the variables $X_1, X_2, \dots, X_i$, we have, using (5.2) again,

$$\mathbf{E}\left[e^{uZ_i}\right] = \mathbf{E}\left[e^{uZ_{i-1}}e^{uY}\right] = \mathbf{E}_{X_1,\dots,X_{i-1}}\left[e^{uZ_{i-1}}\mathbf{E}_{X_i}\left[e^{uY}\right]\right].$$

By Lemma 5.2.1, we have $\mathbf{E}_{X_i}\left[e^{uY}\right] \le e^{u^2/2}$ (for any values of $X_1,\dots,X_{i-1}$). So

$$\mathbf{E}\left[e^{uZ_i}\right] \le \mathbf{E}_{X_1,\dots,X_{i-1}}\left[e^{uZ_{i-1}}e^{u^2/2}\right] = e^{u^2/2}\,\mathbf{E}\left[e^{uZ_{i-1}}\right] \le e^{iu^2/2}$$

by the inductive hypothesis.

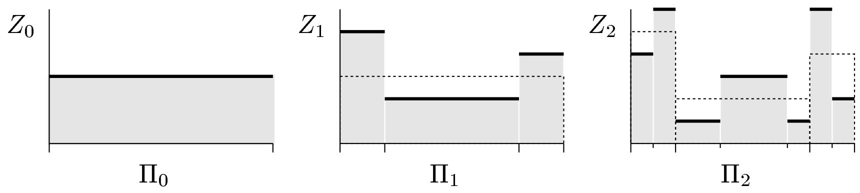
We have derived $\mathbf{E}\left[e^{uX}\right] \le e^{nu^2/2}$ for all u. The desired inequality $\mathrm{P}[X \ge t] < e^{-t^2/2n}$ now follows by applying Markov's inequality for the random variable e^{uX}, exactly as in the proof of Theorem 4.1.1. □

Martingales and Azuma's inequality. We introduce the (rather sophisticated) notion of martingale, which allows us to state the result of the proof above in greater generality.

Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let $\mathcal{F}_0 = \{\emptyset, \Omega\} \subset \mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots \subseteq \mathcal{F}$ be a sequence of σ-algebras on Ω. In the case of a finite Ω, one can think of the $\mathcal{F}_i$ as successively finer and finer partitions of Ω (formally, in this case, $\mathcal{F}_i$ is the σ-algebra generated by some partition Π_i of Ω, i.e. its members are all disjoint unions of some of the classes of Π_i). For example, if $\Omega = \{0,1\}^n$, then $\mathcal{F}_i$ can be generated by the partition of Ω induced by the first i coordinates. Let $Z_0, Z_1, \ldots$ be a sequence of random variables on Ω, where each Z_i is $\mathcal{F}_i$-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

$$\mathbf{E}\left[Z_i | \mathcal{F}_{i-1}\right] = Z_{i-1}, \quad i = 1, 2, \ldots, n. \tag{5.3}$$

The expression $\mathbf{E}\left[Z_i|\mathcal{F}_{i-1}\right]$ means the *conditional expectation* of Z_i with respect to $\mathcal{F}_{i-1}$. If $\mathcal{F}_{i-1}$ and $\mathcal{F}_i$ 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, Z_{i-1} is constant on each class of Π_{i-1}, and the martingale condition (5.3) 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.

It is not difficult to check that the sequence $Z_0, Z_1, \ldots, Z_n$ in the above proof is a martingale. Moreover, almost the same proof gives the tail estimates as in Theorem 5.1.1 for the nth term of an arbitrary martingale $Z_0, Z_1, \ldots, Z_n$ with $|Z_i - Z_{i+1}| \le c_i$:

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

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

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

Sketch of proof. Suppose that $\mathbf{E}\left[X\right] = 0$ and $c_i = 1$. We again prove $\mathbf{E}\left[e^{uZ_i}\right] \le e^{iu^2/2}$ by induction on i (and all the rest of the proof is exactly as above). This time we have

$$\mathbf{E}\left[e^{uZ_{i-1}}e^{uY}\right] = \mathbf{E}\left[\mathbf{E}\left[e^{uZ_{i-1}}e^{uY} \,\middle|\, \mathcal{F}_{i-1}\right]\right] = \mathbf{E}\left[e^{uZ_{i-1}}\mathbf{E}\left[e^{uY} \,\middle|\, \mathcal{F}_{i-1}\right]\right].$$

The martingale condition (5.3) now guarantees that $\mathbf{E}[Y \mid \mathcal{F}_{i-1}] = 0$, and Lemma 5.2.1 implies that $\mathbf{E}\left[e^{uY} \mid \mathcal{F}_{i-1}\right] \leq e^{u^2/2}$. (If $\mathcal{F}_{i-1}$ is given by a partiton Π_{i-1}, then for each class C of Π_{i-1}, we consider the random variable Y_C defined as Y restricted to the probability space C. The martingale condition gives $\mathbf{E}[Y_C] = 0$ and we apply Lemma 5.2.1 for Y_C.) □

Remark: two strengthenings. Theorem 5.1.1 and Azuma's inequality can be strengthened in several ways, which allows one to deal with some applications where the original versions are too weak. Here we will briefly mention two directions of such strengthenings.

Suppose that $f(X_1, \ldots, X_n)$ is 1-Lipschitz. If X_1 attains value 0 with probability $1-p$ and value 1 with probability p and p is small, one would expect that the contribution of X_i to the "total variance", i.e. to the quantity denoted by σ^2 in Theorem 5.1.1, should be considerably smaller than 1. A result of this type indeed holds, also with variables X_i attaining more than two values, and a precise formulation 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.

Another strengthening is based on the observation that the Lipschitz condition for f need not be used in full in the proof of Theorem 5.1.1. 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 X_i to a truthful oracle (such as "what is the value of X_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 the paper cited above.

5.3 Lipschitz functions on discrete metric spaces

Here we consider generalizations of Theorem 5.1.1, where we want a concentration result for a Lipschitz function $f(X_1, X_2, \ldots, X_n)$, but the random variables $X_1, \ldots, X_n$ are not independent anymore. Clearly, we have to require something of the X_i in order to get a concentration result; for example, if all the X_i were equal to X_1 and $f(X_1, \ldots, X_n) = X_1 + \cdots + X_n$, then there is no concentration at all. The framework to be examined here is when the vector $(X_1, X_2, \ldots, X_n)$ is a random point of a suitable "high-dimensional" metric space. The concentration results are closely related to interesting geometric properties of the considered metric spaces: the so-called *isoperimetric inequalities*.

Concentration of Lipschitz functions of a random permutation. We prove one concrete result in the direction indicated above. 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 $\rho(\pi_1, \pi_2) = |\{i \in [n]\colon \pi_1(i) \neq \pi_2(i)\}|$.

5.3.1 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 $t \geq 0$, we have*

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

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 5.3.1 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 5.3.1. 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 (pairwise 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 $\mathcal{F}_i$ be the σ-algebra generated by Π_i, and let Z_i be the random variable given by

$$Z_i = \mathbf{E}[f(\pi)|\mathcal{F}_i].$$

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 (5.3). We want to apply Azuma's inequality 5.2.2, and so we need to bound the differences: we will prove that

$$|Z_i - Z_{i-1}| \leq 2. \tag{5.4}$$

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.

Now 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 $\rho(\sigma, \varphi(\sigma)) \leq 2$ for all $\sigma \in C_{j_1}$. 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 $\sigma \in C_{j_1}$, we set $\varphi(\sigma) = \sigma'$, where $\sigma'(i) = b_2$, $\sigma'(\sigma^{-1}(b_2)) = b_1$, and $\sigma'(j) = \sigma(j)$ for $\sigma(j) \notin \{b_1, b_2\}$. We have

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

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

We have established the bound (5.4) for the martingale differences, and Azuma's inequality 5.2.2 yields Theorem 5.3.1. □

The proof of Theorem 5.3.1 can be generalized to yield concentration results for more general discrete metric spaces. The key condition is that such spaces have a suitable sequence of partitions. Some results of this kind 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.

6

Appendix

6.1 Probability theory

This section summarizes the fundamental notions of probability theory and some results which are used in the text. In no way is it intended to serve as a substitute for a course in probability theory.

6.1.1 Definition (Probability space). *A probability space 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 on* Σ *with* $\mathrm{P}[\Omega] = 1$. *The elements of* Ω *are called* events *and the elements of* Σ *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* 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) is equally likely.

6.1.2 Definition (Random graphs). *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 the probability of a graph with* m *edges is*

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

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 for each pair $\{u, v\}$ of vertices and connect them by an edge if the outcome is heads.

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

6.1.3 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,

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

□

6.1.4 Definition (Independence). *Events A, B are independent 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 $I \subseteq \{1, 2, \dots, n\}$*

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

This is not equivalent to all the pairs A_i, A_j being independent. Exercise: find three events A_1, A_2 and A_3 which 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.

6.1.5 Definition (Conditional probability). *For events A, B where $\mathrm{P}[B] > 0$, we define the conditional probability as*

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

Note that if A and B are independent, the conditional probability $\mathrm{P}[A|B]$ is equal to $\mathrm{P}[A]$.

6.1.6 Definition (Random variables). *A real random variable on a probability space $(\Omega, \Sigma, \mathrm{P})$ is a function $X : \Omega \to \mathbf{R}$ 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 only real random variables.

6.1.7 Definition (Expectation). *The* expectation *of a (real) random variable X is*

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

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

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

6.1.8 Definition (Independence of variables). *Random variables X, Y are independent if*

$$\forall a, b \in \mathbf{R} : \mathrm{P}[X \le a \text{ and } Y \le b] = \mathrm{P}[X \le a]\,\mathrm{P}[Y \le b]$$

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

As we will check in Chapter 2, $\mathbf{E}[X+Y] = \mathbf{E}[X] + \mathbf{E}[Y]$ 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

6.1.9 Lemma. *If X and Y are independent random variables, then $\mathbf{E}[XY] = \mathbf{E}[X] \cdot \mathbf{E}[Y]$.*

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

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

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

6.2 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 $n! \leq n^n$. More refined bounds are

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

(where e 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 $\binom{n}{k} \leq n^k$, and sharper ones are

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

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

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

(see also Section 3.2 for a derivation of a slightly weaker lower bound).

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

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

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 $0 \leq p \leq \frac{1}{2}$.