Math 130B

Asaf Ferber

June 9, 2020

Abstract

This document served as my lecture notes for Math 130B in UC Irvine, during the spring quarter 2020. Most of the material is taken from the two lovely books: "Probability and computing: Randomization and probabilistic techniques in algorithms and data analysis" by Mitzenmacher and Upfal [2], and "A first course in probability" by Ross [3]. Moreover, I also found the lecture notes of Gravner [1] very useful. I have no claims for originality of the arguments and most of the math is taken from other sources. I do take full responsibility for all the math mistakes appearing in this document.

Contents

1 A general course introduction

In this course we will advance our knowledge in probability – both in theory and in applications. In particular, as a side goal, we will be exposed to some basic (and not-so-basic) ideas from randomized algorithms.

1.1 Randomized algorithms: Verifying polynomial identities

Suppose that we have written a program to multiply monomials. For example, the following can be an output of our program:

$$(x-1)(x-2)(x-3)(x-10) = x^4 - 16x^3 + 70x^2 - 116x + 60.$$

Before you continue reading, in order to appreciate the following arguments, please try to check whether the above identity is correct or not.

As you have all noticed, the above identity is wrong as the coefficient of x^2 should be 71. Moreover, if you indeed try to figure it our by yourself, then it should be easy to convince you that, for a similar identity, if we take a product of 10000 binomials, it will take you a lot of time to do the same verification.

This raises the following problem:

Problem 1.1. *Suppose that $F(x)$ and $G(x)$ are two polynomials. Can we find an "efficient" way to check whether the polynomial identity $F(x) = G(x)$ is correct?*

As this class is about probability and not algorithms, we will not try to formally define/explain what the word "efficient" stands for. Instead, we will always try to compare how many basic operations we need to make using a naive approach versus how many basic steps our (sometimes also quite naive) algorithm makes.

Naive approach Clearly, in order to convince ourselves that $F(x) = G(x)$, we can write both polynomials in their canonical forms as: $F(x) = \sum_{k=0}^{d} a_k x^k$, and $G(x) = \sum_{k=0}^{d} b_k x^k$, and check whether $a_k = b_k$ for all k. Indeed, by a theorem from calculus/linear algebra, we know that $F(x) = G(x)$ if and only if $a_k = b_k$ for all k (try to prove this!).

Now let us analyze the number of basic operations that we need to perform in order to write $F(x) = \prod_{i=1}^{d}(x - r_i)$ in its canonical form. Imagine that we expand this expression in d steps, where in each round j we compute $F_{j+1} := (x - r_j) \cdot F_j(x)$, where $F_j(x)$ is the canonical form of the product of the first $j-1$ binomials $\prod_{i=1}^{j-1}(x - r_i)$.

Now, since $F_j(x)$ is a polynomial of degree $j-1$, it has exactly j terms in its canonical form (some may be zero). Therefore, to compute $F_{j+1}(x)$, we need to do at most (say) $3j$ basic operations (multiplying all coefficients by r_j, multiplying by x, and then adding up coefficients of the same power of x). All in all, since $j = 1, \ldots, d$, and since $1 + 2 + \ldots + d = \binom{d+1}{2}$ (WHY?), we can compute the canonical form, starting from a product of monomials, by doing at most $3\binom{d+1}{2}$ basic operations. Note that if we assume that d is large, the multiplicative constants in this expression don't matter much in terms of number of computations – we consider $3d^2$ or $100d^2$ the same. We say that the number of operations is *of order at most* d^2, and denote it by $O(d^2)$. This means that there exists $C > 0$ such that for all d, the number of operations is at most Cd^2.

Observe that the naive approach has two main disadvantages:

1. It is basically the same approach like we originally tried to do. In particular, if the same person writes the code, then assuming that the original code has an error, it is very likely that the same error occurs in the second code.

2. In order to double check something that we believe is likely to be true, it would be better if we could find a simpler (meaning – faster) way to check for correctness, even if there is some small chance for making an error. This is where probability comes into the game!

A Probabilistic Approach Assume that the degree of $F(x)$ and $G(x)$ is d. Now, if $F(x) = G(x)$, then in particular, it means that $F(r) = G(r)$ for all $r \in \mathbb{R}$. Moreover, if $F(x) \neq G(x)$, then in particular we have that $F(x) - G(x)$ is a polynomial (and not the 0 polynomial) of degree at most d. Therefore, by the fundamental theorem of algebra, it has at most d roots.

Consider the set (say) $\{1, \ldots, 100d\}$. We know that if $F(x) \neq G(x)$ then there are at most d numbers r in this set that satisfy $F(r) - G(r) = 0$. So, we can pick a random element r from this set and compute

$F(r) - G(r)$. If the answer is 0, then we conclude that $F(x)$ is probably equal to $G(x)$. Otherwise, we conclude that $F(x) \neq G(x)$. Observe that this test has a 1-*sided error*. That is, if $F(r) \neq G(r)$ for the randomly chosen r, then we are absolutely certain that $F(x) \neq G(x)$, and the only error can come if we conclude that $F(x) = G(x)$ (WHY?). What is the probability for such an error? Clearly it's at most $1/100$, which we consider to be a "reasonable" margin of error. How efficient is this test? In order to compute the value of $F(r) - G(r)$ we have to do only $O(d)$ basic operations (WHY?), so we saved quite a lot of time in the case where d is large.

In later sections, we will use more sophisticated tools to analyse more complicated algorithms.

1.2 Probabilistic counting: number of triangles in dense graphs

A *graph* $G = (V, E)$ is a pair consisting of a set V whose elements are called *vertices*, and a collection of pairs of vertices $E \subseteq \binom{V}{2} = \{\{u, v\} : u, v \in V\}$ whose elements are called *edges*.

A well known theorem in Graph Theory is the following theorem by Mantel:

Theorem 1.2 (Mantel's theorem)**.** *Suppose that $G = (V, E)$ is a graph on n vertices that contains no triangles. Then, $|E| \leq \frac{n^2}{4}$.*

Claim 1.3. *Let $G = (V, E)$ be any graph. Then,*

$$\sum_{x \in V} d(x)^2 = \sum_{xy \in E} (d(x) + d(y)),$$

where $d(x)$ is the degree *of the vertex x: the number of vertices adjacent to x.*

Proof. This is left as an exercise for the reader. □

Proof of Mantel's. Since G is triangle-free, we know that any two adjacent vertices x and y have no common neighbor. Therefore, for each $xy \in E$ we have $d(x) + d(y) \leq n$. Now, by the above claim we obtain that

$$\sum_{x \in V} d(x)^2 = \sum_{xy \in E} (d(x) + d(y)) \leq |E|n.$$

Moreover, since $\sum_{x \in V} d(x) = 2|E|$ (try proving this!), by Cauchy-Schwarz we obtain that

$$4|E|^2 \leq n \sum_{x \in V} d(x)^2.$$

Combining the above bounds we obtain

$$4|E|^2 \leq n^2|E|,$$

which gives us the desired inequality. □

Following Mantel's theorem one can naturally ask the following question:

Question 1.4. *Suppose that $G = (V, E)$ is a graph on n vertices and with $|E| \geq n^2/4 + \epsilon n^2$. How many triangles does it have?*

Basically, what if we have just a few more than $\frac{n^2}{4}$ edges? Clearly one cannot hope for some closed formula because there are many such graphs and it is quite easy to form many such graphs with distinct numbers of triangles. BUT what we are interested is in its asymptotics: how the answer behaves as $n \to \infty$. Observe that there are $\binom{n}{3}$ triangles in the complete graph on n vertices so any lower bound of the form cn^3 is good for us. To obtain such a lower bound we will use some (very) useful probabilistic tricks.

Theorem 1.5. *Suppose that $G = (V, E)$ is a graph on n vertices and with $|E| \geq n^2/4 + \epsilon n^2$. Then, G has at least $f(\epsilon)n^3$ many triangles for some function f.*

Proof. First, we need to make the following rather trivial observation:

Claim 1.6. *Suppose that $G' = (V', E')$ is a graph on m vertices with $|E'| \geq m^2/4 + k$. Then G' has at least k triangles.*

Proof. Take such a G'. By Mantel's theorem it must have a triangle. Let $e \in E'$ be an edge from this triangle, and consider the graph $G'' = (V', E' - \{e\})$, the graph obtained by deleting the edge e. This graph still has more than $m^2/4$ edges if $k > 1$ and therefore must also contain a triangle (which is not the previous triangle!). Repeat this k times. □

Now, let $X \subseteq V$ be a random subset of vertices, where each vertex is chosen with probability p (to be determined later), independently at random. Let $e(X)$ be the random variable counting the number of edges induced in X. Since for every edge $xy \in E$, the probability that both x and y are in X is p^2, by the linearity of expectation we have that

$$\mathbb{E}[e(X)] = |E|p^2 \geq n^2p^2/4 + \epsilon n^2p^2.$$

Now, let $T(G)$ denote the number of triangles in G. Observe that we don't know $T(G)$ as this is the parameter we are trying to bound. Since each triangle appears in X with probability p^3, by the linearity of expectation we have that

$$\mathbb{E}[T(X)] = T(G)p^3.$$

Moreover, by the above claim, we have the *deterministic (not random) bound*

$$T(X) \geq e(X) - |X|^2/4.$$

Taking the expectation of both sides and using the linearity of expectation gives

$$\mathbb{E}[T(X)] = T(G)p^3 \geq |E|p^2 - (np)^2/4 \geq \epsilon n^2p^2.$$

You may have noticed that we're cheating a little bit here. $\mathbb{E}[|X|^2] \neq \mathbb{E}[|X|]^2$, but these quantities are close to one another in a precise sense that we will make formal later. If you suspend your disbelief and accept this cheat, we have that

$$T(G) \geq \epsilon n^2/p.$$

In particular, we can choose $p = c/n$ for some large enough c, and obtain the desired result. This completes the proof. □

If you don't completely understand all the steps in this proof yet, don't be too worried. This was one of many examples illustrating the power of probabilistic arguments in different areas. We will now turn to give a quick review of knowledge that we expect to have from Math 130A.

2 The basics

2.1 Axioms of probability

Recall that a *probability space* consists of a *sample space* Ω, which is the set of all possible outcomes of our "experiment", a family $\mathcal{F}$ of subsets of Ω representing all the allowable events, and a *probability function* $\Pr : \mathcal{F} \to \mathbb{R}$ satisfying Definition (2.1).

Definition 2.1. *A* probability mass function *(or a probability function)* $\Pr : \mathcal{F} \to \mathbb{R}$ *is any function that satisfies the following three axioms:*

1. *For all* $E \in \mathcal{F}$, $0 \leq \Pr[E] \leq 1$, *and*

2. $\Pr[\Omega] = 1$, *and*

3. *For any sequence of pairwise mutually disjoint events* $E_1, E_2, \ldots$ *we have*

$$\Pr[\cup_i E_i] = \sum_i \Pr[E_i].$$

Suppose $A, B \in \mathcal{F}$ are two events in the sample space Ω, it is helpful to represent the relations between them (intersection, union, complement etc) in a Venn diagram.

For convenience, we will recall some basic properties of the operations $\cap$ and $\cup$:

Commutative laws For all A, B we have $A \cup B = B \cup A$ and $A \cap B = B \cap A$.

Associative laws For all A, B, C we have $(A \cup B) \cup C = A \cup (B \cup C)$ and $(A \cap B) \cap C = A \cap (B \cap C)$.

Distributive laws For all A, B, C we have $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$ and $(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$.

DeMorgan's laws For all $n \in \mathbb{N}$ and all $A_1, \ldots, A_n$ we have

$$\left(\bigcup_{i=1}^{n} A_i\right)^c = \bigcap_{i=1}^{n} A_i^c \text{ and } \left(\bigcap_{i=1}^{n} A_i\right)^c = \bigcup_{i=1}^{n} A_i^c.$$

Exercise 2.2. *Prove that Axiom* 3 *from Definition 2.1 also implies its finite version. That is, prove that if* Pr *satisfies Axioms 1-3 then, in particular, for any sequence of mutually disjoint events* $A_1, A_2, \ldots, A_N$ *we have*

$$\Pr\left[\bigcup_{n=1}^{N} A_n\right] = \sum_{n=1}^{N} \Pr[A_n].$$

The following propositions are simple consequences of the axioms in Definition 2.1.

Proposition 2.3. $\Pr[A^c] = 1 - \Pr[A]$.

Proposition 2.4. *If* $A \subseteq B$ *then* $\Pr[A] \leq \Pr[B]$.

Proposition 2.5. $\Pr[A \cup B] = \Pr[A] + \Pr[B] - \Pr[A \cap B]$

A generalization of the last proposition is called the *inclusion-exclusion* principle and is summarized in the following theorem.

Theorem 2.6 (Inclusion-Exclusion formula)**.** *Let* $A_1, A_2, \ldots$ *be events in a sample space* Ω. *Then,*

$$\Pr\left[\bigcup_{n=1}^{N} A_n\right] = \sum_{i=1}^{N} \Pr[A_i] - \sum_{1 \leq i_1 < i_2 < N} \Pr[A_{i_1} \cap A_{i_2}] + \sum_{1 \leq i_1 < i_2 < i_3 < N} \Pr[A_{i_1} \cap A_{i_2} \cap A_{i_3}] \pm \ldots$$

Proof. First, note that if an outcome is not in $\cup A_n$, then it contributes nothing to $\Pr[\cup A_n]$. Indeed, it belongs to none of the A_n's and therefore, the right hand side (RHS for short) doesn't count it at all.

Second, suppose that an element $x \in \cup A_n$ belongs to exactly m events. Then. x appears in any summand that involves intersections only of subsets that contain x. Therefore, it contributes

$$\sum_{k=1}^{m} (-1)^{k+1} \binom{m}{k} = 1.$$

This completes the proof. □

As a simple exercise one can also prove the following useful estimates that follow almost immediately from the inclusion-exclusion formula:

Theorem 2.7. *Let $A_1, \ldots, A_n \subseteq \Omega$ events in a sample space Ω. Then,*

- **The union bound** $\Pr[\cup_{i=1}^{n} A_i] \leq \sum_{i=1}^{n} \Pr[A_i]$, *and*
- **Simple lower bound** $\Pr[\cup A_i] \geq \sum \Pr[A_i] - \sum \Pr[A_i \cap A_j]$, *and*
- **Three-terms upper bound** $\Pr[\cup A_i] \leq \sum \Pr[A_i] - \sum \Pr[A_i \cap A_j] + \sum \Pr[A_i \cap A_j \cap A_k]$, *etc.*

In general, the above theorem asserts that the inclusion-exclusion formula, as an alternating sum, always "sandwiches" the probability we are trying to estimate.

2.2 Independence and conditional probability

In this section we introduce two of the main concepts in probability, namely, conditional probabilities, and independence of events. We start with the latter.

To motivate the concept a bit, let us go back to the "verifying polynomial identities algorithm" for a moment. Suppose that we want to "boost" the error probability from 1/100 to something smaller. One way to do this is to sample a random point from a larger set, say, $\{1, \ldots, 1000d\}$. This will decrease the error to be at most 1/1000. Another way is to perform the algorithm again *independently* (whatever that means...). The probability to have an error twice is at most $1/100^2 = 1/10000$ – not that bad! If we repeat the algorithm k times, then the error probability becomes $1/100^k$ which is exponentially small in the number of trials. To get a better feeling for the word "exponential," try to write down the number $1/100^k$ even for small k (for example, take $k = 10$), and observe that this is a super tiny constant. In order to make the above arguement rigorous, we need to turn the word "indepedent" into a mathematical term:

Definition 2.8. *Two events $E, F \in \mathcal{F}$ are called* independent *if*

$$\Pr[E \cap F] = \Pr[E] \Pr[F].$$

More generally, events $E_1, E_2, \ldots, E_n$ are called mutually independent *if for any subset $I \subset [n]$ we have*

$$\Pr[\bigcap_{i \in I} E_i] = \prod_{i \in I} \Pr[E_i].$$

That is, if E_i is the event "the ith time that we run our algorithm is a failure", and we run the algorithm k times independently, then the probability to fail in all of them is precisely

$$\Pr[\bigcap_{i=1}^{k} E_i] = \prod_{i=1}^{k} \Pr[E_i].$$

Next, we discuss *conditional probability*. As an illustrating and nice example, consider the following problem: suppose that a couple has two kids and that the genders and DOB of the kids are independent. Now, suppose we are told that the older kid is a boy. What is the probability that the younger one is a boy as well? Clearly, as the genders are independent, the answer is 1/2.

Next, suppose we are told that at least one of the kids is a boy. What is the probability that both of them are boys? In this case, if we examine all the options, it is not hard to convince ourselves that the answer is 1/3 (DO IT!).

Lastly, suppose we are told that one of the kids is a boy that was born on a Monday. What is the probability that both of them are boys? In particular, does the additional information affect the probability that we've just computed?

In order to be able to compute these probabilities, we need to introduce the concept of *conditional probabilities*.

Definition 2.9. *Let $E, F \in \mathcal{F}$. Then, the probability that the event E occurs given that the event F occurs is*

$$\Pr[E \mid F] = \frac{\Pr[E \cap F]}{\Pr[F]}.$$

Observe that if E and F are independent and $\Pr[F] \neq 0$, then we have

$$\Pr[E \mid F] = \frac{\Pr[E \cap F]}{\Pr[F]} = \frac{\Pr[E]\Pr[F]}{\Pr[F]} = \Pr[E].$$

That is, at least on an intuitive level, if E and F are independent, then any information about one of them does not affect the probability of the other event.

As an almost immediate corollary from Definition 2.9, we obtain the following *multiplication rule*:

Theorem 2.10 (The multiplication rule)**.** *Let $E_1, \ldots, E_n \subseteq \Omega$ be any events. Then,*

$$\Pr[\cap_{i=1}^{n} E_i] = \Pr[E_1] \cdot \Pr[E_2 \mid E_1] \cdot \Pr[E_3 \mid E_1 \cap E_2] \cdots \Pr[E_n \mid \cap_{i=1}^{n-1} E_i].$$

Next, suppose that $E, F \subseteq \Omega$ are events. We can separate E into the parts that live in F and those that live outside of F, i.e.

$$E = (E \cap F) \cup (E \cap F^c),$$

which is a union of two disjoint events. Therefore, by Axiom 3 we have that

$$\Pr[E] = \Pr[E \cap F] + \Pr[E \cap F^c],$$

which by the multiplication rule gives us

$$\Pr[E] = \Pr[E \mid F]\Pr[F] + \Pr[E \mid F^c]\Pr[F^c].$$

This simple formula can be generalized to arbitrary events as follows:

Theorem 2.11 (The law of total probability)**.** *Suppose that $F_1, \ldots, F_n \subseteq \Omega$ are disjoint events for which $\Omega = \cup F_i$, and let $E \subseteq \Omega$ be any event. Then,*

$$\Pr[E] = \sum_{i=1}^{n} \Pr[E \mid F_i]\Pr[F_i]. \tag{1}$$

It might not look like it, but the law of total probability is immensely powerful. Effectively, it lets you compute the probability of *any* event E by conditioning on *any* partition $\cup F_i$ that you want. The subtlety is in choosing the "correct" partition for the problem at hand, as the next example illustrates.

Exercise 2.12. *A man possesses five coins, two of which are double-headed, one is double-tailed, and two are normal. He shuts his eyes, picks a coin at random, and tosses it.*

1. *What is the probability that the lower face of the coin is a head?*
2. *He opens his eyes and see that the coin is showing heads. What is the probability that the lower face is heads as well?*
3. *He shuts his eyes again, and tosses the coin again. What is the probability that the lower face is head?*
4. *He opens his eyes and sees that the coin is showing heads. What is the probability that the lower face is a head?*
5. *He discards this coin, picks another one at random (but not the same coin), and tosses it. What is the probability that it shows heads?*

Exercise 2.13 (The prosecutor's fallacy)**.** *Let G be the event that an accused person is guilty, and let T be the event that some testimony is true. Some lawyers have argued on the assumption that* $\Pr[G \mid T] = \Pr[T \mid G]$. *Show that this holds if and only if* $\Pr[G] = \Pr[T]$.

Exercise 2.14 (Galton's paradox)**.** *You flip three fair coins. At least flips come up the same and it is an even chance that the third is heads or tails. Therefore,* $\Pr[\textit{all alike}] = \frac{1}{2}$. *Do you agree? If not, what should this probability be instead?*

2.3 Application: upper bounding an "atom-probability"

We now give a simple application of conditional probability to *random sums*. In particular, we will learn two tricks which will be very useful (and won't be explained again) in later sections.

Suppose that $\bar{a} = (a_i)_{i=1}^n \in \mathbb{R}^n$ is not the all-zero vector (that is, $a_j \neq 0$ for at least one $1 \leq j \leq n$). Let $\bar{x} = (x_i)_{i=1}^n \in \{0,1\}^n$ be a 0/1 vector chosen according to a uniform distribution (meaning, each of the 2^n possible vectors can be chosen with the same probability 2^{-n}). The parameter (or more formally, the random variable) that we are interested in is the inner product of the two vectors, that is,

$$S_n := \bar{a} \cdot \bar{x} = \sum_{i=1}^{n} a_i x_i.$$

The random variable S_n is referred to as a "random sum". Now, we define its *atom probability* as follows:

$$\rho(S_n) = \max_{m \in \mathbb{R}} \Pr[S_n = m].$$

(exercise: why could we write max and not sup?) We will revisit this random variable many times in these notes, but for now, let us prove the following simple upper bound on $\rho(S_n)$.

Proposition 2.15. *For all $m \in \mathbb{R}$ we have that*

$$\Pr[\bar{a} \cdot x = \sum_i a_i x_i = m] \leq \frac{1}{2}.$$

Proof. The first trick that we need to take away from this is that a randomly chosen vector $\bar{x} \in \{0,1\}^n$ can be seen as $\bar{x} = (x_1, \ldots, x_n)$, where the x_i's are independent, and identically distributed (from now on *i.i.d*), random variables, with $\Pr[x_1 = 1] = \Pr[x_1 = 0] = \frac{1}{2}$.

Exercise 2.16. *Prove it!*

The second trick that we need to understand is how to use conditional probability here. Observe that since $\bar{a} \neq \bar{0}$, there must exist a coordinate $1 \leq j \leq n$ for which $a_j \neq 0$. By relabelling if necessary, we may assume that $a_1 \neq 0$. Now, suppose we already know the values of the partial vector $\bar{x}^{(1)} = (x_2, \ldots, x_n)$. In particular, we already know the value $s = \sum_{i=2}^{n} a_i x_i$, and we have that $S_n = a_1 x_1 + s$. The important observation here is that, no matter what s is, there is at most one value of x_1 for which $S_n = m$, for a fixed m (WHY?). Since x_1 takes one of two values, each occurring with probability $1/2$, we have $\Pr[S_n = m \mid \bar{x}^{(1)} = \bar{b}^{(1)}] \leq 1/2$ for any $\bar{b}^{(1)}$. Now we use the law of total probability, conditioning on the possible values of the partial vector $\bar{x}^{(1)}$:

$$\Pr[S_n = m] = \sum_{\bar{b}^{(1)} \in \{0,1\}^{n-1}} \Pr[S_n = m \mid \bar{x}^{(1)} = \bar{b}^{(1)}] \Pr[\bar{x}^{(1)} = \bar{b}^{(1)}] \leq \frac{1}{2} \sum_{\bar{b}^{(1)} \in \{0,1\}^{n-1}} \Pr[\bar{x}^{(1)} = \bar{b}^{(1)}] = \frac{1}{2},$$

where the last equality follows from the fact that Pr is a probability mass function. This completes the proof. □

Next we show how to obtain a much stronger bound using a beautiful combinatorial argument.

2.4 Improved atom probability: Sperner's theorem and a problem of Littlewood and Offord

In this section we will give a simple solution to a problem that we will revisit many times.

Exercise 2.17. *Let Ω be the set of all perumtations of $\{1, \ldots, n\}$. What is the probability that the elements of a fixed subset $S \subseteq \{1, \ldots, n\}$ of size k will be an initial segment of a randomly chosen permutation?*

If you solved the above exercise correctly, then you obtained that the answer is $\frac{1}{\binom{n}{k}}$. This fact will be the key in the proof of the following theorem due to Sperner:

Theorem 2.18 (Sperner's theorem)**.** *Suppose that $\mathcal{F}$ is a collection of subsets of $[N]$ such that there are no two subsets $A \neq B \in \mathcal{F}$ for which $A \subset B$. Then, $|\mathcal{F}| \leq \binom{n}{n/2}$.*

Proof. Let Ω be the set of all permutations of $\{1, \ldots, n\}$, equipped with a uniform distribution. For every $S \in \mathcal{F}$, let E_S be the event "S is an initial segment of the chosen permutation". Since $\mathcal{F}$ contains no two sets $S, T \in \mathcal{F}$ with $S \subset T$, it follows that the events $(E_S)_{S \in \mathcal{F}}$ are mutually disjoint. In particular, it follows that

$$\sum_{S \in \mathcal{F}} \Pr[E_S] \leq 1.$$

Now, since $\Pr[E_S] = \frac{1}{\binom{n}{|S|}}$ (this follows from the above exercise), and since $\binom{n}{|S|} \leq \binom{n}{n/2}$ for all $|S|$ (the middle binomial coefficient is always the biggest one. Prove this!), we conclude that

$$|\mathcal{F}| / \binom{n}{n/2} \leq 1,$$

which completes the proof. □

We will now use Sperner's theorem to solve the following, seemingly unrelated, problem, posed by Littlewood and Offord in 1938, considering the distribution of zeroes of random polynomials. Suppose that $a_1, \ldots, a_n$ are given real numbers, with absolute value at least 1. How many sums of the form $\sum_i \varepsilon_i a_i$, where $\varepsilon_i \in \{-1, 1\}$ can lie within an open unit interval? They proved that this number is at most $\frac{c \log n}{\sqrt{n}} 2^n$ for some fixed constant c. Later on, Erdős found an elegant way to obtain an optimal bound using Sperner's theorem. This result is now known as the Erdős-Littlewood-Offord inequality and has a tremendous amount of applications and extensions.

Theorem 2.19 (Erdős, 1945)**.** *Let $a_1, \ldots, a_n$ be real numbers of absolute value at least one. For all open unit intervals I, there are at most $\binom{n}{\lfloor n/2 \rfloor}$ vectors $(\varepsilon_i)_{i=1}^n \in \{-1, 1\}^n$ such that $\sum \varepsilon_i a_i \in I$.*

Proof. Note that by changing signs of the a_is we do not change the distribution, and therefore we are allowed to assume that they are all positive. Now, fix an open, unit interval I, and let $S_I = \{(\varepsilon_i)_{i=1}^n \in \{\pm 1\}^n : \sum_i \varepsilon_i a_i \in I\}$. For each vector $\varepsilon = (\varepsilon_i) \in S$, let $A_\varepsilon \subseteq [n]$ be the set of all indices i for which $\varepsilon_i = 1$, and let $\mathcal{A} := \{A_\varepsilon : \varepsilon \in S_I\}$. In order to complete the proof, it is enough to claim that $\mathcal{A}$ contains no sets $S, T \in \mathcal{A}$ for which $S \subset T$, and then apply Sperner's theorem. Indeed, suppose that there are $\varepsilon \neq \varepsilon'$ with $A_\varepsilon \subset A_{\varepsilon'}$. As all the a_is have absolute value at least 1, it follows that

$$\left| \sum \varepsilon_i a_i - \sum \varepsilon_i' a_i \right| \geq 1$$

and therefore they cannot both lie in I. This completes the proof. □

Note that the above theorem gives the best possible bound as the sequence $a_i = 1$ for all i shows. Clearly, the number of vectors ε can be large only if there are many cancelations. That is, intuitively, it means that the sequence a_i has some 'nice' additive properties. What if, for example, we enforce all the a_i's to be distinct integers? can we do better?

2.5 Bayes's Formula

Let $B_1, \ldots, B_n$ be a set of disjoint events for which $\cup B_i = \Omega$. Suppose now that and event A has occurred and we are interested in determining which one of the B_j's also occurred. Then, by (1) we have the following proposition:

Proposition 2.20 (Bayes's formula)**.**

$$\Pr[B_j \mid A] = \frac{\Pr[A \cap B_j]}{\Pr[A]} = \frac{\Pr[A \mid B_j] \Pr[B_j]}{\sum_{i=1}^n \Pr[A \mid B_i] \Pr[B_i]}.$$

Exercise 2.21. *Prove it!*

As a simple application to Bayes's formula, consider the following example: suppose that we have 3 coins. One of them is biased with a probability 2/3 for H, and the other two are unbiased. We don't know which of them is biased and we would like to identify it. We permute the three coins at random and then flip all of them. Suppose that the first and second coins turned up H and the third is T. What is the probability that the first coin is the biased one?

Clearly, before flipping them, the probability is 1/3 as we permuted the coins at random and each coin has equal probability to be the first. Now, let E_i be the event "the ith coin is biased" (where $i = 1, 2, 3$), and let B be the event "the outcome is HHT". We wish to compute the probability $\Pr[E_1 \mid B]$.

First, we compute

$$\Pr[B \mid E_1] = \Pr[B \mid E_2] = \frac{2}{3} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{6},$$

and

$$\Pr[B \mid E_3] = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{12}.$$

Now, using Bayes's formula we have that

$$\Pr[E_1 \mid B] = \frac{\Pr[B \mid E_1] \Pr[E_1]}{\sum_{i=1}^{3} \Pr[B \mid E_i] \Pr[E_i]} = \frac{\frac{1}{6}}{\frac{1}{6} + \frac{1}{6} + \frac{1}{12}} = \frac{2}{5}.$$

That is, based on the outcome, it is more likely that coin 1 is biased.

2.6 Application: Karger's algorithm for finding a min-cut

Let $G = (V, E)$ be any graph. A *cut* of G is a partition of its vertices into two non-empty sets $V = A \cup B$. Every cut (A, B) of G determines the *cut-set* $E(A, B)$, which is the set of all edges in E with one endpoint in A and the other endpoint in B. The *size* of a cut (A, B) is defined as $|E(A, B)|$, that is, as the number of edges in the corresponding cut-set. The *min-cut* problem is about finding a cut with a cut-set of a minimum size.

Observe that in order to find a min-cut using a naive approach, one could go over all 2^n possible partitions, counting the number of cut edges and then take the minimum. In particular, the running time is exponential since there are 2^n possible partitions. Here we will present a beautiful and simple randomized algorithm due to Karger that can find a min-cut with probability (say) 0.9999 in running time which is only polynomial in n.

In order to describe the algorithm we need to set up some notation. The main operation in the algorithm is *edge contraction*. That is, when we say "contracting the edge uv" we mean that we merge the two vertices u and v into one vertex, eliminate all edges connecting u and v, and retain all other edges in the graph. The new graph may have parallel edges but no self-loops (WHY?).

The algorithm consists of $n-2$ *rounds*, where in each round i, we pick an edge uv uniformly at random and contract it. Note that after each round, the number of vertices in the new graph reduces by exactly one, and therefore, after the $(n-2)$nd round we are left with only two vertices. The algorithm then outputs the set of all the edges connecting these two remaining vertices.

It is not hard to convince yourself that every cut-set during the execution of the algorithm is also a cut-set of the original graph (actually convince yourself!). On the other hand, clearly not every cut-set appears as a cut-set in an intermediate stage of the algorithm. Therefore, even though the algorithm outputs a cut-set of the original graph, there is no guarantee that it is indeed a min-cut. The main result that we are going to prove is the following:

Theorem 2.22 (Karger)**.** *The algorithm outputs a min-cut with probability at least* $\frac{1}{\binom{n}{2}} = \frac{2}{n(n-1)}$.

Before proving the theorem, let us explain how we can indeed find a min-cut with high probability in polynomial time. Clearly, the algorithm itself is a polynomial time algorithm (there are roughly n rounds, and in each round we contract one edge). Now, suppose we run this algorithm Cn^2 many times for some constant $C > 0$, independently at random, and then declare the the minimal output among the Cn^2 outputs is a min-cut. What is the probability that it is not?

Since by Theorem 2.22 we know that an output is a min-cut with probability at least $2/n(n-1)$, the probability that we missed a min-cut Cn^2 times is at most

$$(1 - 2/n(n-1))^{Cn^2} \leq e^{-2C},$$

where in the last inequality we used the simple inequality $1 - x \leq e^{-x}$ which is valid for all $x \in \mathbb{R}$.

Finally, by taking C to be (say) 100, we obtain a very small probability for an error.

Now we turn to the proof of Theorem 2.22.

Proof. Let k be the size of the min-cut set of G. The graph may have several cut-sets of minimum size, and we compute the probability of finding one specific such set C. Since C is a cut-set in the graph, removal of the set C partitions the set of vertices into two sets, S and $V - S$, such that there are no edges connecting vertices in S to vertices in $V - S$.

Assume that, throughout an execution of the algorithm, we contract only edges that connect two vertices in S or two vertices in $V - S$, but not edges in C. In that case, all the edges eliminated throughout the execution will be edges connecting vertices in S or vertices in $V - S$, and after $n - 2$ iterations the algorithm returns a graph with two vertices connected by the edges in C. We may therefore conclude that, if the algorithm never chooses an edge of C in its $n - 2$ iterations, then the algorithm returns C as the minimum cut-set.

This argument gives some intuition for why we choose the edge at each iteration uniformly at random from the remaining existing edges: if the size of the cut C is small and if the algorithm chooses the edge uniformly at each step, then the probability that the algorithm chooses an edge of C is small - at least when the number of edges remaining is large compared to the size of C.

Now, let E_i be the event that the edge contracted in the i-th round of the algorithm is not in C, and let $F_j = \cap_{i=1}^{j} E_i$ be the event that none of the first j contracted edges belong to C. Our goal is to lower bound $\Pr[F_{n-2}]$. To this end, we use the multiplication rule

$$\Pr[F_{n-2}] = \Pr[E_{n-2} \mid F_{n-3}] \Pr[E_{n-3} \mid F_{n-4}] \cdots \Pr[E_2 \mid F_1] \Pr[E_1], \tag{2}$$

and observe that if F_j holds, then the current graph has $n - j$ vertices, and since the size of the min-cut is k, we obtain that the graph has at least $k(n - j)/2$ edges. Indeed, every vertex must have degree at least k (WHY?). Therefore, by picking an edge at random, the probability that it does not belong to C is at least

$$\Pr[E_{j+1} \mid F_j] \geq 1 - \frac{k}{k(n-j)/2} = 1 - \frac{2}{n-j} = \frac{n-j-2}{n-j}.$$

Plugging it into (2) we obtain that

$$\Pr[F_{n-2}] \geq \prod_{j=0}^{n-3} \left(\frac{n-j-2}{n-j} \right) = \frac{2}{n(n-1)}.$$

This completes the proof. □

3 Discrete random variables

3.1 Random variables and expectation

So far we have dealt with events and their probabilities. Another very important concept in probability is that of a random variable. A *random variable* is simply a function X defined on the points of our sample space Ω. That is, associated with every $\omega \in \Omega$, there is a value $X(\omega)$. For the time being, we will only consider functions that take values over the reals $\mathbb{R}$, but the range of a random variable can be any set.

Examples 3.1. *1. Experiment: roll two dice. Clearly,* $\Omega = \{(a,b) \mid 1 \le a,b \le 6\}$ *(and hence* $|\Omega| = 36$*). Define a random variable* $X : \Omega \to \mathbb{R}$ *as follows.* $X((a,b)) = a+b$ *(that is, we don't care about the actual outcome of the experiment but only about the sum of the numbers on the dice). What can we say about* $\Pr[X = x]$*? trivially,* $\Pr[X = x] = 0$ *for all* $x \le 1$ *or* $x \ge 13$*. For all the other values, there is a non-zero probability (compute it!). An important observation is that*

(a) "$X = x$*" is an event and can be formally described as* $E_x := \{(a,b) \in \Omega \mid a+b = x\}$*.*

(b) For all $x \neq y$ *the events* E_x *and* E_y *are disjoint.*

(c) $\Omega = \cup_x E_x$ *(That is, all the events defined by* X *form a partition of the sampling space). In particular, we have that* $\sum_x \Pr[X = x] = 1$*.*

2. Experiment: do n *independent trials each of which has a success probability* p*. Let* X *bethe number of successes.*

3. Experiment: do infinitely many independent trials each of which has a success probability p*, and let* X *be the first time we had a success.*

Given a random variable X, we define its *cumulative distribution function*, or, more simply, its *distribution function of* F_X as follows:

$$F_X(x) = \Pr[X], \quad -\infty < x < \infty.$$

That is, for all real values x, $F_X(x)$ outputs the probability that the random variable X is at most x.

A random variable X which has only finitely or countable many possible values is called a *discrete random variable*. For such a variable, the function $p_X(x_i) := \Pr[X = x_i]$ is called its *probability mass function*. Whenever the random variable we are dealing with is clear from the context, we will delete the subscript X and simply write $p(x)$.

The probability mass function has the following properties:

1. $p(x_i) > 0$ for all i (we don't list outcomes which occur with probability 0).
2. For any subset A we have $\Pr[X \in A] = \sum_{x \in A} p(x)$.
3. $\sum_i p(x_i) = 1$ (that is, X has some value).

The definition of independence of events extends to random variables as follows:

Definition 3.2. *Two random variables* X, Y *are called* independent *if and only if*

$$\Pr[X = x, Y = y] = \Pr[X = x] \cdot \Pr[Y = y]$$

for all x, y*.*

More generally, random variables $X_1, \ldots, X_n$ *are called* mutually independent *if and only if for all* $I \subseteq [n]$ *and for every possible* $(x_i)_{i \in I}$ *we have that*

$$\Pr[\cap_{i \in I} X_i = x_i] = \prod_{i \in I} \Pr[X_i = x_i].$$

Given a random variable X, we define its *expected value* as

$$\mathbb{E}[X] = \sum_{w \in \Omega} \Pr[\{w\}] X(w).$$

Another expression for the expectation (prove its equivalence!) is

$$\mathbb{E}[X] = \sum_{x \in \text{range}(X)} x \Pr[X = x].$$

This is quite straightforward to verify using the fact that

$$\Pr[X = x] \quad = \sum_{w \in \Omega : X(w) = x} \Pr[\{w\}].$$

A key property of expectation that significantly simplifies its computation is the *linearity of expectation*. This property is summarized in the following theorem:

Theorem 3.3. *Let $X_1, \ldots, X_n$ be random variables, and let $a_0, \ldots, a_n$ be any numbers. Then,*

$$\mathbb{E}[a_0 + \sum_{i=1}^{n} a_i X_i] = a_0 + \sum_{i=1}^{n} a_i \mathbb{E}[X_i].$$

As we will see in the next section, given a random variable X, sometimes we are interested in calculating the expectation of some function of X (for example, we are interested at $\mathbb{E}[X^2]$, or $\mathbb{E}[1/X]$ etc.). Observe that for any function g we have that $g(X)$ is also a random variable. Therefore, by the definition of expectation we can write:

$$\mathbb{E}[g(X)] = \sum_{x} g(x) \Pr[X = x].$$

(convince yourselves!)

For example, imagine that X is a r.v with probability mass function $\Pr[X = 1] = \Pr[X = -1] = \frac{1}{2}$, and let $g_1(x) = x^2 + 2$, $g_2(x) = e^x$, and $g_3(x) = e^{2\pi i t x}$ (recall from complex analysis that $e^{it} = \cos t + i \sin t$).

Then,

$$\mathbb{E}[g_1(X)] = ((-1)^2 + 2)\frac{1}{2} + (1^2 + 2)\frac{1}{2} = 3,$$

$$\mathbb{E}[g_2(X)] = e^{-1} \cdot \frac{1}{2} + e^{1} \cdot \frac{1}{2},$$

and

$$\mathbb{E}[g_3(X)] = e^{-2\pi i t} \cdot \frac{1}{2} + e^{2\pi i t} \cdot \frac{1}{2} = \cos 2\pi t.$$

The following theorem asserts that in case that the random variables are independent, we can also compute expectation of products :

Theorem 3.4. *Suppose that $X_1, \ldots, X_n$ are mutually independent random variables and let $g_1, \ldots, g_n$ be functions of one variable. Then,*

$$\mathbb{E}[g_1(X_1) \cdot g(X_2) \cdots g_n(X_n)] = \mathbb{E}[g_1(X_1)] \cdot \mathbb{E}[g_2(X_2)] \cdots \mathbb{E}[g_n(X_n)].$$

Proof. Prove it by induction on n. □

3.2 Application: Expected number of cycles in a random permutation

Recall that a *permutation* is just a bijection $\pi : [n] \to [n]$. Now, given a permutation π, we can define its associated directed graph D_π as follows: The vertices are $V = [n]$, and the *directed edges* (namely, ordered pairs) are $E = \{(j, \pi(j)) \mid j \in [n]\}$. A moment's thought now reveals that this graph must consists of a collection of vertex disjoint cycles (self loops and cycles of length two are allowed), covering the entire vertex-set.

Exercise 3.5. *Pick two arbitrary permutations of length 10 and draw their associates digraphs.*

Exercise 3.6. *Find permutations π_1, π_2, π_3 of length 10 with $10, 1,$ and 4 many cycles, respectively.*

A natural question to ask is:

Question 3.7. *How many cycles a "typical" permutation has? or in other words, what is the expected number of cycles in a randomly chosen permutation?*

We prove the following theorem:

Theorem 3.8. *The expected number of cycles in a randomly chosen permutation of length n is*

$$\sum_{j=1}^{n} \frac{1}{j} = \ln n + C + \epsilon_n,$$

where $\epsilon_n \to 0$ as $n \to \infty$.

Proof. How can we even start? First, let X be a random variable defined on the set of all permutations of length n (this set is usually denoted as S_n) as

$$X(\pi) = \text{the number of cycles in the permutation } \pi.$$

Next, our goal is to "break" this variable into a sum of simpler random variables. The main observation here is the following: suppose we have a cycle of length k, and that we assign each of its members a "weight" $1/k$. Then, the total number of weights along this cycle is exactly 1. This leads us to define the following random variables:

$$\text{For all } 1 \leq i \leq n, \text{ let } Y_i = \text{ the length of the (unique) cycle that element } i \text{ belongs to.}$$

From the above discussion, it is clear that we have

$$X = \sum_{i=1}^{n} \frac{1}{Y_i},$$

and therefore, by linearity of expectation we have

$$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[1/Y_i].$$

Observe that, by symmetry, all the Y_i's are distributed the same (CONVINCE YOURSELF!) and therefore, we have that

$$\mathbb{E}[X] = n\mathbb{E}[1/Y_1].$$

Now we calculate $\mathbb{E}[1/Y_1]$. By definition, we have that

$$\mathbb{E}[1/Y_1] = \sum_{j=1}^{n} \frac{1}{j} \Pr[\text{element 1 is in a cycle of length } j]. \tag{3}$$

Claim 3.9. *For all $1 \leq j \leq n$ we have that*

$$\Pr[\textit{element 1 is in a cycle of length } j] = \frac{1}{n}.$$

Proof. The probability that 1 is on a cycle of length 1 (meaning, a self loop) is precisely $1/n$. The probability that 1 is on a cycle of length 2 is $\frac{n-1}{n} \cdot \frac{1}{n-1} = \frac{1}{n}$ (we have $n-1$ ways out of n to choose $\pi(1)$, then we need to choose $\pi(\pi(1)) = 1$ out of $n-1$ possibilities. Try to complete the proof by writing the formula for a general $1 \leq j \leq n$. □

All in all, plugging the estimate from Claim 3.9 into (3) yields

$$\mathbb{E}[1/Y_1] = \frac{1}{n} \sum_{j=1}^{n} \frac{1}{j},$$

and therefore,

$$\mathbb{E}[X] = n\mathbb{E}[1/Y_1] = \sum_{j=1}^{n} \frac{1}{j} = \ln n + C + \epsilon_n.$$

This complete the proof. □

Example 3.10 (A riddle – Beating the intuition)**.** *Suppose that n prisoners are playing the following game: there are n boxes, and each box contains a card with a name of exactly one prisoner and every name appears exactly once. The game is played in n rounds, where in round i, prisoner i goes into the room with the boxes and can open up to $n/2$ boxes. If one of the opened boxes contains their name, they close all the other boxes and immediately get out of jail, get \$1,000,000 and start a new life (without seeing/talking to their prisoner friends). If they fail in finding their name in one of these boxes, then they and all the other remaining prisoners are immediately executed.*

The rules are that no communication between the prisoners is allowed after the game starts (they can talk before the game starts and agree on some strategy). Moreover, the boxes are labelled and the cards are distributed uniformly at random among the boxes.

Try to come up with a good strategy which can help all the prisoners to start a new life.

We will show how to find such a strategy that with probability at least 0.3 *ensures that all prisoners will be able to start a new life, regardless of the number of prisoners.*

Proof. The strategy is simple – the prisoners assign themselves numbers ahead of time and prisoner i opens box i. They see a name (which corresponds to a number) on the card. If this is their name, then they are happy and the game is over for them. Otherwise, they see some name j. Then they open box j, etc until they open $n/2$ boxes.

Observe that this strategy can only fail if the corresponding permutation has cycles of length larger than $n/2$. Let's count the number of such "bad" permutations. We can upper bound the number of such permutations by (check!)

$$\sum_{k=n/2}^{n} \binom{n}{k} (k-1)!(n-k)! = n! \sum_{k=n/2}^{n} \frac{1}{k},$$

which by the inequality

$$\sum_{k=n/2}^{n} \frac{1}{k} \leq \int_{n/2-1}^{n+1} \frac{1}{x} dx = \ln \frac{n+1}{n/2-1} \approx \ln 2 + o(1)$$

is at most $0.7n!$. In particular, since the permutation (the names in the boxes) is random, the probability that it has no cycles of length larger than $n/2$ is at least 0.3. This completes the proof. □

3.3 Variance of random variables

Given a random variable X, we would like to summarize the essential properties of its distribution function (which is sometimes very complicated to calculate!) by certain measures. One example for such a measure is $\mathbb{E}[X]$, which gives us the weighted average of the possible values of X. The main disadvantage of the expectation is that it does not measure our "risk" (in case that X meausres our gain/loss in some luck game), or more precisely, the "variation" (or spread), of the values of X. To this aim we define the *variance* of X.

Definition 3.11. *Let X be any random variable. Its* variance *is defined as*

$$Var(X) = \mathbb{E}[(X - \mathbb{E}[X])^2].$$

That is, the variance of X is the weighted average of the "distance squere" of X from its expectation. Intuitively, if $Var(X)$ is "large", then the values of X differ from the expectation "a lot".

As an immediate consequence from the definition, since we take an expectation of something squared, we have that $Var(X) \geq 0$, regardless of what X is.

Another consequence, which follows quite easily by combining the definition with linearity of expectation, is the following alternative definition for variance:

$$Var(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 .$$

In technical problems, you will find that this definition of variance is usually easier to use. The reason why I chose to define it via the first definition is that it doesn't hide the idea behind the variance.

Another simple and interesting consequence from both definitions of the variance combined is that

$$(\mathbb{E}[X])^2 \leq \mathbb{E}[X^2],$$

for every r.v X. (prove it!).

To have a little feeling for how to work with the first definition of the variance let us prove the following intuitive proposition:

Proposition 3.12. *Let X be a discrete random variable. Then, $Var(X) = 0$ if and only if X is constant (by X is a constant we mean that there exists some constant $C \in \mathbb{R}$ for which $\Pr[X = C] = 1$).*

Proof. Suppose that X is a constant. Then, we have that with probability 1 we have both $X = C$ and $X^2 = C^2$. In particular, we have that

$$\mathbb{E}[X] = C \text{ and } \mathbb{E}[X^2] = C^2,$$

and therefore, by the second defintion of variance we conclude that

$$Var(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = C^2 - C^2 = 0.$$

For the other direction, suppose that $Var(X) = 0$ and we wish to show that there exists some constant C for which $\Pr[X = C] = 1$. Assume towards a contradiction that there exists two constants $C_1 \neq C_2$ for which $0 < \Pr[X = C_1], \Pr[X = C_2] < 1$. Without loss of generality we may assume that $C_1 \neq \mathbb{E}[X]$. Therefore, from the first definition of variance we conclude that

$$Var(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] \geq (C_1 - \mathbb{E}[X])^2 \Pr[X = C_1] > 0,$$

which contradicts the assumtion $Var(X) = 0$. This completes the proof. □

3.4 Special (discrete) distributions

In this section we review some special discrete distributions that will play a crucial role in later sections.

3.4.1 Bernoulli random variables

A random variable X is said to be *distributed according to Bernoulli distribution with parameter* p, and denoted by $X \sim Ber(p)$ if and only if its probability mass function is

$$\Pr[X = 1] = p, \text{ and } \Pr[X = 0] = 1 - p.$$

We won't over estimate Bernoulli random variables if we say that these are probably the most (or at least in the top 5) important random variables in probability. The reasons for that will be clear in later sections. For now, let us record that $\mathbb{E}[X] = p$ and that $Var(X) = p(1-p)$ (prove it!).

3.4.2 Binomial distribution

We say that a random variable is distributed according to a *binomial distribution* with parameters n and p, and denote it by $X \sim Bin(n, p)$, if and only if its probability mass function is

$$\Pr[X = k] = \binom{n}{k} p^k (1-p)^{n-k}, \text{ where } k = 0, \ldots, n.$$

As a motivating story, one can think about the following problem: Suppose that n independent trials are being performed, each of which has a success probability p. Let X be the random variable counting the number of successes. Then $X \sim Bin(n, p)$. Now, for each trial i we define a random variable X_i which outputs 1 if the ith trial is a success and 0 otherwise. Clearly, $X_i \sim Ber(p)$ for all i, and more importantly, we have

$$X = X_1 + \ldots + X_n.$$

Therefore, by the linearity of expectation and the fact that $\mathbb{E}[Ber(p)] = p$, we obtain that

$$\mathbb{E}[X] = \mathbb{E}[X_1 + \ldots + X_n] = \mathbb{E}[X_1] + \ldots + \mathbb{E}[X_n] = np.$$

There is also a shortcut to calculate the variance (and we will see it in a later section), but for now let us just state, without a proof, that $Var(X) = np(1-p)$.

3.4.3 Geometric distribution

We say that $X \sim Geo(p)$ (that is, X is distributed according to a geometric distribution with parameter p) if and only if its probability mass function is

$$\Pr[X = k] = (1-p)^{k-1} p, \text{ for } k = 1, 2, \ldots$$

A geometric random variable counts the number of trials required to see the first success, when independent trials with success probability p are being performed.

Indeed, $\Pr[X = 1] = p$ is the probability to have a success in the first trial, $\Pr[X = 2] = (1-p)p$ is the probability that the first trial is a failure and the second is a success, $\Pr[X = 3] = (1-p)^2 p$ is precisely the probability to fail in first two trials and have a first success in the third, etc.

The expectation of X is $\mathbb{E}[X] = \frac{1}{p}$ and its variance is $Var(X) = \frac{1-p}{p^2}$ (can you prove it?).

3.4.4 Uniform distribution

Given a set $S = \{x_1, \ldots, x_n\}$, we say that a random variable X is *uniformly distributed over* S, and is denoted by $X \sim Uniform(S)$ if and only if $\Pr[X = x_i] = \frac{1}{n}$ for all $1 \leq i \leq n$.

It is relatively straight forwards to compute (do it!) that

$$\mathbb{E}[X] = \frac{x_1 + \ldots + x_n}{n},$$

and

$$Var(X) = \frac{x_1^2 + \ldots + x_n^2}{n} - \left(\frac{x_1 + \ldots + x_n}{n}\right)^2.$$

3.4.5 Poisson distribution

A random variable X is distributed according to *poisson distribution with parameter* λ, and denoted by $X \sim Poisson(\lambda)$ if and only if its probability mass function is

$$\Pr[X = k] = e^{-}\frac{\lambda^k}{k!}, \text{ where } k = 0, 1, \ldots.$$

Its expectation is $\mathbb{E}[X] = \lambda$ and its variance is $Var(X) = \lambda$. It is a good exercise to calculate these parameters by yourselves (Hint: recall the Taylor expansion of $e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$).

The main takeaway is the following theorem that, roughly speaking, asserts that poisson distribution is a "good" approximation for binomial distribution.

Theorem 3.13. *Suppose that n is "large" and that p is "small". Then for $X \sim Bin(n, p)$, we can "approximate" X by assuming that $X \sim Poisson(\lambda)$, where $\lambda = np$. That is, for all fixed k, if n goes to infinity then*

$$\binom{n}{k} p^k (1-p)^{n-k} \to e^{-\lambda}\frac{\lambda^k}{k!}.$$

Proof. Let us fix a k and calculate

$$\lim_{n\to\infty} \binom{n}{k} p^k (1-p)^{n-k} = \lim_{n\to\infty} \frac{n(n-1)\cdots(n-k+1)}{k!}\frac{\lambda^k}{n^K}\left(1-\frac{\lambda}{n}\right)^n \left(1-\frac{\lambda}{n}\right)^{-k}$$

$$= \lim_{n\to\infty} \frac{\lambda^k}{k!} e^{-\lambda},$$

where in the last equality we used the following fact that we know from calculus:

$$\lim_{n\to\infty} \left(1-\frac{\lambda}{n}\right)^n = e^{-\lambda}.$$

This completes the proof. □

Before we move on to the next topic, let us give the following somehow surprising example:

Example 3.14. *An experiment with success probability p is being performed n times. Let X be the number of heads, and Y be the number of tails. Clearly, X and Y are not independent as $Y = n - X$ (so in particular, knowing X we can compute Y). BUT suppose now that the experiment is being performed a* **random** *number of times N, where $N \sim Poisson(\lambda)$. Show that now X and Y are independent!*

4 Continuous random variables

In this section we learn about *continuous* random variables. A random variable X is *continuous* if there exists a non-negative function f so that, for every interval $I \subseteq \mathbb{R}$, we have that

$$\Pr[X \in I] = \int_I f(x)dx.$$

The function $f = f_X$ is called the *density* of X.

For a random variable X with a density function f, its *distribution function* $F(x) := F_X(x)$ is defined by

$$F(x) = \int_{-\infty}^{x} f(t)dt.$$

Observe that $F(x) = \Pr[X \leq x]$.

Since a density function is not unique (for example, changing finitely many values of f cannot change the integral), we will usually assume that the function f is continuous in all but at most finitely many jumps, and therefore we have that $F'(x) = f(x)$ for all x for which F is differential at.

Density function has the same role as the probability mass function for discrete random variables, and it measures how "probable" certain subsets of $\mathbb{R}$ are. The following proposition summarizes its properties (and is left as an exercise):

Proposition 4.1. *Let X be a random variable with density function f. Then,*

1. $\int_{-\infty}^{\infty} f(x)dx = 1$,
2. $\Pr[X = x] = 0$ *for all* $x \in \mathbb{R}$,
3. $\Pr[a \leq X \leq b] = \int_a^b f(x)dx$.

Example 4.2. *Suppose that X is a random variable with a density function*

$$f_X(x) = \begin{cases} 3x^2 & \text{if } x \in [0,1] \\ 0 & \text{otherwise.} \end{cases}$$

Let $Y = 1 - X^4$ and compute its density function $f_Y(y)$.

Proof. The plan for solving this problem goes as follows: first, we compute Ys distribution function $F_Y(y)$. Then, we define $f_Y(y) = F'(y)$ in all "relevant" points.

Now, observe that $f_Y(y)$ is non-zero only for $y \in [0,1]$ as its values are restricted to that interval (WHY?). Next, let $y \in (0,1)$, and recall that

$$F_Y(y) = \Pr[Y \leq y] = \Pr[1 - y \leq X^4] = \Pr[(1-y)^{1/4} \leq X].$$

Finally, by definition we have that

$$\Pr[(1-y)^{1/4} \leq X] = \int_{(1-y)^{1/4}}^{1} 3x^2 dx.$$

Therefore, we have

$$f_Y(y) = F'(y) = -3(1-y)^{1/2}(-1/4)(1-y)^{-3/4},$$

for all $y \in (0,1)$. For $y \notin (0,1)$ we have $f_Y(y) = 0$. This completes the proof. □

4.1 Expectation and variance

Analogously to the discrete case, we define

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x) dx,$$

and

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f(x) dx. \tag{4}$$

The variance is defined as

$$Var(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2.$$

Example 4.3. *Let* $f(x) = \begin{cases} \lambda e^{-x/20} & \text{if } x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$

1. *Find a value of λ for which $f(x)$ is a density function.*

2. *Let X be a random variable with f as its density function. Compute* $\Pr[X \leq 10]$.

3. *Compute $\mathbb{E}[X]$, $\mathbb{E}[X^2]$, and $Var(X)$.*

Proof. By definition, all $\lambda \geq 0$ the function f is non-negative. Therefore, in order to make it a density function, we need to choose λ for which

$$\int_{-\infty}^{\infty} f(x)dx = 1.$$

Observe that

$$\int_{-\infty}^{\infty} f(x)dx = \int_{0}^{\infty} \lambda e^{-x/20}dx = \frac{-\lambda}{20} \cdot e^{-x/20} \mid_0^\infty = \lambda/20.$$

Therefore, for $\lambda = 20$ the function $f(x)$ is a density function.

Now, to calculate $\Pr[X \leq 10]$, we need to calculate the integral

$$\int_{-\infty}^{10} f(x)dx.$$

(do it yourselves!).

For its expectation, we need to calculate

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} 20xe^{-x/20}dx,$$

which can be done by the integration-by-parts principle.

Similarly, we compute

$$\mathbb{E}[X^2] = \int_{-\infty}^{\infty} 20x^2e^{-x/20}dx,$$

and then combine the results to compute the variance as

$$Var(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2.$$

Please complete the missing details by yourselves! □

For non-negative random variables we can find the following proposition useful in calculating the expectation:

Proposition 4.4. *Let X be a random variable with density function f. Suppose that $f(x) = 0$ for all $x < 0$. Then,*

$$\mathbb{E}[X] = \int_{0}^{\infty} \Pr[X \geq x]dx.$$

Proof. From the definition of a density function, we know that

$$\Pr[X \geq x] = \int_{y=x}^{\infty} f(y)dy.$$

Therefore, we obtain that

$$\int_{0}^{\infty} \Pr[X \geq x]dx = \int_{0}^{\infty} \left(\int_{y=x}^{\infty} f(y)dy \right) dx.$$

Finally, by interchanging the order of integration, we obtain that the above quantity equals

$$\int_{0}^{\infty} \left(\int_{x=0}^{y} f(y)dx \right) dy = \int_{0}^{\infty} f(y) \left(\int_{x=0}^{y} 1dx \right) dy \int_{0}^{\infty} yf(y)dy = \mathbb{E}[X].$$

This completes the proof. □

Exercise 4.5. *Prove the following:*

1. *Let $X_1, X_2, \ldots, X_n$ be iid random variables for which $\mathbb{E}[X_i^{-1}]$ exists. Show that if $m \leq n$, then*

$$\mathbb{E}[S_m/S_n] = m/n,$$

where $S_m = X_1 + \ldots + X_m$.

2. *Let X be a non-negative random variable with density function f. Show that*

$$\mathbb{E}[X^r] = \int_0^\infty r x^{r-1} \Pr[X > x] dx.$$

4.2 Special (continuous) distributions

As in the discrete case, we will now summarize some famous distributions for continuous random variables.

4.2.1 Uniform distribution

We say that a continuous random variable X is distributed according to a *uniform* distribution over the interval $[a, b]$, and denote it by $X \sim U[a, b]$, if and only if its density function is

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a, b] \\ 0 & \text{otherwise.} \end{cases}$$

That is, all the points in the interval $[a, b]$ have the same "mass", and points outside this interval have "mass zero".

It is a good (and relatively easy) exercise to prove the following:

- $\mathbb{E}[X] = \frac{b+a}{2}$.
- $Var(X) = \frac{(b-a)^2}{12}$.

Exercise 4.6. *Calculate the distribution function $F(x)$ of $X \sim U[a, b]$.*

Example 4.7. *A stick of length 1 is split at a point U that is uniformly distributed over $(0, 1)$. Determine the expected length of the piece that contains the point p, where p is some arbitrary point $0 < p < 1$.*

Proof. Fix some $0 < p < 1$, and let X be a random variable that measures the length of the piece that contains the point p. Observe that

$$X = \begin{cases} 1 - U & \text{if } U < p \\ U & \text{if } p < U. \end{cases}$$

Therefore, by (4) we have that

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} X(u) f(u) du = \int_0^p (1 - u) du + \int_p^1 u du = p(1 - p) + \frac{1}{2}.$$

□

Exercise 4.8. *Suppose that $X \sim U[0, 1]$. What is the probability that $X \in \mathbb{Q}$?*

Example 4.9. *Suppose that $X \sim U[0, 1]$. Then X divides the interval $[0, 1]$ into two intervals $[0, x]$ and $[x, 1]$. Let R be the ratio between the smaller and the larger intervals. That is, if $x \leq 1/2$, then $R = \frac{x}{1-x}$, and if $x \geq 1/2$ then $R = \frac{1-x}{x}$. Compute the density function of R.*

Proof. First, observe that $0 \leq R \leq 1$ always, and therefore we have that $f_R(r) = 0$ for all $r \notin [0,1]$. Now, suppose that $r \in [0,1]$ and we wish to compute Rs distribution function $F_R(r)$. Note that

$$F_R(r) = \Pr[R \leq r] = \Pr\left[X \leq 1/2, \frac{X}{1-X} \leq r\right] + \Pr\left[X \geq 1/2, \frac{1-X}{X} \leq r\right] = \Pr\left[X \leq 1/2, X \leq \frac{r}{r+1}\right] + \Pr\left[X \right. \tag{5}$$

In particular, we have that

$$f_R(r) = F'_R(r) = \frac{2}{(r+1)^2}.$$

This completes the proof. □

4.2.2 Exponential distribution

A random variable X is distributed according to an *exponential distribution with parameter* $\lambda > 0$, and is denoted by $X \sim Exp(\lambda)$ if it has the following density function:

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0. \end{cases}$$

This is a distribution for the waiting time for some random event, for example, for a lightbulb to burn out or for the next earthquake of at

Its distribution function is

$$F(x) = \begin{cases} 1 - e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0. \end{cases}$$

(prove it!)

These are good exercises to prove that

1. $\mathbb{E}[X] = \frac{1}{\lambda}$, and
2. $Var(X) = \frac{1}{\lambda^2}$, and
3. (Memoryless property) $\Pr[X \geq x + y \mid X \geq y] = \Pr[X \geq x] = e^{-\lambda x}$.

The memoryless property means that, starting at any time step, assuming that the event hasn't occured yet, the distribution of the remaining waiting time is the same as it was at the beginning.

Exercise 4.10. *Which of the discrete distributions that we learnt are memoryless?*

Here we'll only prove the expectation's formula, but it is highly recommended to prove the rest by yourselves:

Since we always have $X \geq 0$, we can use the formula from Proposition 4.4 to obtain

$$\mathbb{E}[X] = \int_0^\infty \Pr[X \geq x]dx = \int_0^\infty (1 - F(x))dx = \int_0^\infty e^{-\lambda x}dx = \frac{1}{\lambda}.$$

This completes the proof.

In practice, exponential distribution often arises as the distribution of the amount of time until some specific event occurs.

4.2.3 Normal distribution

We say that a random variable X is distributed according to a *normal (or Gaussian) distribution with parameters* μ and σ^2, and write $X \sim N(\mu, \sigma^2)$ for short, if its density function is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

Arguably, normal distribution is the most important continuous distribution and it arises in many areas. In particular, it can be obtained as a continuous limit of the binomial distribution with p fixed and $n \to \infty$ (we will prove it later in this section). This result is a special case of the so-called Central Limit Theorem (CLT for short) that will be discussed towards the end of this course. Roughly speaking, the CLT asserts that the sum of a large number of independent random variables is approximately normally distributed.

Let us now list the basic parameters that we need to know. Suppose that $X \sim N(\mu, \sigma^2)$, then:

1. Its density function is $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$, for all $-\infty < x < \infty$.
2. $\mathbb{E}[X] = \mu$.
3. $Var(X) = \sigma^2$.

Proving that $\int_{-\infty}^{\infty} f(x)dx = 1$ is a slightly tricky exercise in integration (if you are willing for a challenge, do it!). Also proving the formula for $Var(X)$ is not that simple. But, for the expectation, there is a neat way to prove it by the symmetric nature of the density function as follows:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} xf(x)dx = \int_{-\infty}^{\infty} (x-\mu)f(x)dx + \mu \cdot \int_{-\infty}^{\infty} f(x)dx$$

which, by the change of variable $x - \mu \mapsto z$, equals

$$\int_{-\infty}^{\infty} z \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{z}{2\sigma^2}\right) dz + \mu.$$

Now, observe that the function $g(z) = z\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{z}{2\sigma^2}\right)$ is *odd* (that is, $g(z) = -g(-z)$ for all z) and therefore the integral in the right hand side equals 0. This gives us

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} xf(x)dx = \mu$$

as desired.

Another important property of normal random variables is the following property, which is left as a simple (and very recommended!) exercise:

Exercise 4.11. *Suppose that $X \sim N(\mu, \sigma^2)$. Prove that for all $a \neq 0$ and b, we have that $aX + b \sim N(a\mu + b, a^2\sigma^2)$.*

The next question one should ask is: how can we actually compute probabilities? (in general, we need the distribution function for that, but here this function is non-elementary so calculating integrals are very hard (or even impossible) for us). The first step will be the following simple corollary from (4.11).

Corollary 4.12. *Let $X \sim N(\mu, \sigma^2)$. Then, the random variable $Z = \frac{X-\mu}{\sigma}$ is normally distributed with parameters $\mu_Y = 0$ and $\sigma_Y^2 = 1$. That is, $Z \sim N(0, 1)$.*

A r.v $Z \sim N(0,1)$ is called *standard normal r.v* or *standard gaussian*. This corollary is helpful because now we only care about calculating the cumulative function of a standard gaussian. The standard notation for the distribution function of a standard gaussian is $\phi(x)$. That is, for $Z \sim N(0,1)$, we have that

$$\Phi(x) := \Pr[Z \leq x] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt.$$

The (approximate) values of $\Phi(x)$ for positive values of x are given in a table so we don't need (and actually, cannot really..) calculate them. For negative values of x we can use the same table plus the observation that $\Phi(x) = 1 - \Phi(-x)$, as the density function of a standard gaussian is symmetric around the origin.

Example 4.13. *Let $Z \sim N(0,1)$.*

- *Calculate $\phi(1.23)$, $\Phi(-1)$, $\Phi(-2.07)$.*
- *Calculate $\Pr[0 \leq Z \leq 1]$, $\Pr[-1 \leq Z \leq -0.5]$, $\Pr[-1 \leq Z \leq 0.5]$.*
- *Suppose that $X \sim N(\mu, \sigma^2)$. Calculate $\Pr[X \geq \mu + \sigma]$, $\Pr[\mu - \sigma \leq X \leq \mu + 1.2\sigma]$.*

4.2.4 The normal approximation to the Binomial distribution

A corner stone in probability, known as the DeMoivre-Laplace limit theorem, states that if n is large, then a binomial random variable with parameters n, p will have approximately the same distribution as a normal random variable with the same expectation and variance. This result is a special case (but historically is the first) of one of the most important results in probability theory, the Central Limit Theorem (that we will discuss later in this course).

For convenience, let us now state, without a proof, some version of the DeMoivre-Laplace theorem (we will prove the more general Central Limit Theorem in the near future):

Theorem 4.14 (DeMoivre-Laplace)**.** *Let $X \sim Bin(n,p)$, where p is fixed and n is large. Then, the random variable $Z = \frac{X-np}{\sqrt{np(1-p)}}$ is asymptotically standard gaussian. More precisely, for every $x \in \mathbb{R}$ we have that*

$$\Pr\left[\frac{X - np}{\sqrt{np(1-p)}} \leq x\right] \to \Phi(x),$$

as n tends to infinity.

Note that now we have two approximations for the binomial distribution:

1. When n is large and p is small, we saw in Math 130A that Poisson distribution is a good approximation for binomial (if you don't remember it, then please take a look at the book).
2. When n is large and p is large, we can use Theorem 4.14 (in general, this theorem gives a good approximation as long as $np(1-p) \geq 10$).
3. As a quantitative upper bound for the error between the binomail and the normal distributions in Theorem 4.14, one can prove that it converges to

$$\frac{0.5(p^2 + (1-p)^2)}{\sqrt{np(1-p)}},$$

which can be made arbitrarily small if n is large.

Example 4.15. *A roulette wheel has* 38 *slots:* 18 *red,* 18 *black, and* 2 *green. The ball ends up at one of these at random. You play many games and each game you place* 1*USD on red (if the outcome is red then you earn* 1*USD, otherwise you lose your money). After* n *games, what is the probability that you are not losing money? Answer it for* $n = 100$ *and then* $n = 1000$*. How would the answer change if it was a fair game? (meaning – no green slots).*

Example 4.16. *How many times do you need to toss a fair coin in order to get* 100 *heads with probability at least* 0.9*?*

Beat your intuition – a riddle We will end up this section with the following lovely (and famous) riddle. Suppose that a random real number x have been chosen (according to some distribution) and you have two sealed envelopes in front of you, one contain the number x and the other one contain the number $2x$.

- If you need to guess which of the envelopes contains the number x. What will be your success probability?
- Now, imagine that you are allowed to pick one envelope and to look at the number in it before you make a decision. Can you come up with a strategy that gives you a success rate better than in the previous bullet?

5 Practice problems for Midterm 1

1. A random variable X has density

$$f(x) = \begin{cases} c(x + \sqrt{x}) & \text{if } 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

   (*i*) Determine c (that is, for which value of c the function f is indeed a density function?).

   (*ii*) Compute $\mathbb{E}[1/X]$.

   (*iii*) Compute the density function of $Y = X^2$.

   (*iv*) Compute $Var(X)$.

   (*v*) Suppose that 10 independent trials are being performed, each of which is distributed as X^2 (so the output of each trial is some number). What is the probability that in exactly 3 of the trials we will obtain numbers smaller than 1/4?

2. Let $X \sim Poisson(1)$, $Y \sim Geo(2/3)$, and suppose that $N \sim N(\mu, \sigma^2)$, where

$$\mu = 9 \cdot \Pr[Y \geq 3 \mid Y \geq 1] \text{ and } \sigma^2 = 4\mathbb{E}[X].$$

   Calculate

$$\Pr[-1 \leq N \leq 2].$$

3. A random variable X has density

$$f(x) = \begin{cases} cx^3 & \text{if } 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

   (*i*) (10 points) Determine c (that is, for which value of c the function f is indeed a density function?).

(*ii*) (15 points) Compute the density function of X^2.

(*iii*) (15 points) Compute the density function of $1/X$

(*iv*) (15 points) Compute $Var(1/X)$.

(*v*) (15 points) A trial has a success probability $q = \Pr[1/4 \leq X \leq 1]$. Suppose that independent such trials are being performed and let Y be the first time that a success occurs. Calculate $\mathbb{E}[Y^2]$.

4. (30 points) Let $Y \sim U[0,2]$ and suppose that $X \sim N(\mu, \sigma^2)$, where

$$\mu = \mathbb{E}[Y^2], \text{ and } \sigma^2 = \frac{2}{9}\Pr[Y \geq 1.5 \mid Y \geq 1].$$

Compute $\Pr[1 \leq X \leq 2]$.

6 Joint distribution and independence

We are often interested in probability statements concerning two or more random variables, and the way that they vary together on the same domain Ω. In this section we develop tools to study such scenarios.

6.1 Joint distribution and joint density functions

Definition 6.1. *The* joint distribution *of the random variables X and Y is a function $F : \mathbb{R}^2 \to [0,1]$ such that*

$$F(x,y) = \Pr[X \leq x, Y \leq y].$$

If X, Y are both discrete random variables, then we can define their *joint probability mas function* as

$$p(x,y) := \Pr[X = x, Y = y],$$

and observe that

$$\Pr[X = x] = \Pr[X = x, Y < \infty] = \sum_y p(x,y)$$

is the the probability mass function of X, denoted by $p_X(x)$. Similarly, we obtain that

$$p_Y(y) = \sum_x p(x,y).$$

Observe that for discrete X, Y one can describe their joint probability mass function in a table where the rows correspond to the values of X, the columns correspond to the values of Y, and for every x, y, the (x,y) entry equals $p(x,y)$. The above computatuin show that if we sum the entries in the xth row we obtain $p_X(x)$ whereas the sum of the entries in the yth column gives us $p_Y(y)$. Since these values appear in the margins of the table, the probability mass functions p_X and p_Y are referred to as their *marginal probability mass functions*.

For continuous random variables X and Y, we cannot talk about the probability mass function (because it is identically 0), so we should define their *joint density*.

Definition 6.2. *The random variables X and Y are* jointly continuous *with a* joint density function *$f : \mathbb{R}^2 \to [0, \infty)$ if and only if for all $x, y \in \mathbb{R}$ we have*

$$F(x,y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f(u,v) du dv.$$

If F is differentiable at a point (x, y) then we can specify

$$f(x, y) = \frac{\partial^2}{\partial x \partial y} F(x, y).$$

In particular, if A is a sufficiently "nice" subset of $\mathbb{R}^2$, then

$$\Pr[(X, Y) \in A] = \int \int_A f(x, y) dx dy.$$

Similarly to the discrete case, if we think about an "uncountable table", the marginal distributions of X and Y are defined as

$$F_X(x) = \Pr[X \leq x] = \Pr[X \leq x, Y < \infty] = \int_{-\infty}^{x} \int_{\infty}^{\infty} f(u, v) dv du$$

and

$$F_Y(y) = \Pr[Y \leq y] = \Pr[X < \infty, Y \leq y] = \int_{-\infty}^{y} \int_{\infty}^{\infty} f(u, v) du dv$$

.

In particular it follows that the *marginal density of* X is

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy,$$

and the marginal density of Y is

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx.$$

Note that all probability statements about the pair (X, Y) can be described (at least in theory) by their distributions/ joint distribution functions. For example,

$$\Pr[X > x, Y \leq y] = \Pr[Y \leq y] - \Pr[X \leq x, Y \leq y] = F_Y(y) - F(x, y).$$

Exercise 6.3. *Write* $\Pr[a < X \leq b, c < Y \leq c]$ *in terms of the distribution functions.*

Example 6.4. *Let* $R = [a, b] \times [c, d]$ *be a rectangle, and suppose that the r.v* (X, Y) *is uniformly distributed inside* R*. That is,*

$$f(x, y) = \begin{cases} \frac{1}{(b-a)(d-c)} & \text{if } (x, y) \in R \\ 0 & \text{otherwise} \end{cases}.$$

Let us compute the marginal distribution of X *(the marginal of* Y *is left as a simple exercise). Observe that if* $x < a$ *then* $F_X(x) = 0$*, and if* $x > b$ *then* $F_X(x) = 1$ *(WHY?). Therefore, we choose* $x \geq a$ *and compute*

$$F_X(x) = \int_{-\infty}^{x} \int_{-\infty}^{\infty} f(u, v) dv du = \int_{a}^{x} \int_{c}^{d} \frac{1}{(b-a)(d-c)} dv du = \int_{a}^{x} \frac{1}{b-a} du,$$

which is the distribution function of $U[a, b]$*. In particular, the marginal density function of* X *is just*

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise.} \end{cases}$$

Exercise 6.5. *Consider a circle* C *of of radius* R *centred at the origin. Let* (X, Y) *be a point within the circle, chosen uniformly at random. That is,*

$$f(x, y) = \begin{cases} \frac{1}{\pi R^2} & \text{if } (x, y) \in C \\ 0 & \text{otherwise.} \end{cases}$$

Compute the marginal distributions of X *and* Y*.*

6.2 Expectation of a function of two random variables

Having defined joint distributions, we can ask for the expectation of a function of the two variables. That is, suppose that $g : \mathbb{R}^2 \to \mathbb{R}$ is a "nice" function. Then we define

$$\mathbb{E}[g(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f(x,y)dxdy.$$

6.3 Independent random variables

Recall that two discrete random variables X, Y are called *independent* if and only if $\Pr[X = x, Y = y] = \Pr[X = x] \cdot \Pr[Y = y]$ for all x, y. For continuous random variables we define independence based on the distribution functions:

Definition 6.6. *Two continuous random variables X and Y are* independent *if and only if for all $x, y \in \mathbb{R}$ we have*

$$F(x,y) = F_X(x) \cdot F_Y(y).$$

Note that this is equivalent to saying that $f(x,y) = f_X(x) \cdot f_Y(y)$ for all (x,y) for which F is differentiable at.

All the notions discussed above can be naturally extended to more than two random variables. That is, given $X_1, \ldots, X_n$ random variables, their joint distribution is a function $F : \mathbb{R}^n \to [0,1]$ defined as follows:

$$F(x_1, \ldots, x_n) = \Pr[X_1 \leq x_1, \ldots, X_n \leq x_n].$$

Then, we can define the joint density as

$$\frac{\partial^n}{\partial x_1 \partial x_2 \ldots \partial x_n} F(x_1, \ldots, x_n),$$

in every point for which F is differentiable at.

Equivalently, the joint density is a function $f : \mathbb{R}^n \to [0, \infty)$ such that for every "nice" set $R \subseteq \mathbb{R}^n$ we have

$$\Pr[(X_1, \ldots, X_n) \in R] = \int \ldots \int_R f(x_1, \ldots, x_n)dx_1 \ldots dx_n.$$

Finally, the random variables $X_1, \ldots, X_n$ are *independent* if and only if for every non-empty subset $I \subseteq [n]$ the variables $(X_i)_{i \in I}$ are independent. In the continuous setting this is equivalent to say that

$$f(x_{i_1}, \ldots, x_{i_k}) = \prod_{j=1}^{k} f_{X_{i_j}}(x_{i_j}),$$

for all k and all $i_1 < \ldots < i_k$.

Example 6.7 (Buffon's needle)**.** *A plain is ruled by the lines $y = n$, $n \in \mathbb{Z}$, and a needle of length 1 is cast randomly on to the plane. What is the probability that the needle intersects some line?*

Proof. We let (X, Y) denote the coordinates of the centre of the needle in the plane. and let Θ be its angle, modulo π, with the x-axis. The distance of the needle's centre from the nearest line beneath is $Z = X - \lfloor X \rfloor$. We can assume that

- Z is uniformly distributed in $[0,1]$,
- Θ is uniformly distributed in $[0, \pi]$, and

- Z and Θ are independent (so in particular we have $f(z,\theta) = f_Z(z)f_\Theta(\theta)$.

By drawing a small diagram one can see that the needle intersects some line if and only if (Z,Θ) has the form $Z \leq \frac{1}{2}\sin\Theta$ or $1 - Z \leq \frac{1}{2}\sin\Theta$. Therefore

$$\Pr[\text{ intersection occurs}] = \int\int_B f(z,\theta)dzd\theta = \ldots = \frac{2}{\pi}.$$

Historically, Buffon designed this experiment in order to estimate the value of π. □

6.4 Sum of random variables

Suppose X, Y are random variables with F, F_X, F_Y being their joint, and marginal distributions, respectively. How does the random variable $X + Y$ distributed? Or more generally, what is the distribution of $X_1 + \ldots + X_n$?

Observe that

$$F_{X+Y}(t) = \Pr[X + Y \leq t] = \int_{-\infty}^{\infty}\int_{-\infty}^{t-x} f(x,y)dydx.$$

Now by a change of variable $y = z - x$, the above expressions equals

$$\int_{-\infty}^{\infty}\int_{-\infty}^{t} f(x,z-x)dzdx = \int_{-\infty}^{t}\int_{-\infty}^{\infty} f(x,z-x)dxdz.$$

By taking the derivative with respect to t, we obtain that

$$f_{X+Y}(t) = \int_{-\infty}^{\infty} f(x,t-x)dx.$$

In case that X and Y are independent, we have that

$$f_{X+Y}(t) = \int_{-\infty}^{\infty} f_X(x)f_Y(t-x)dx,$$

which is called the *convolution* of f_X and f_Y, and is denoted by $(f_X * f_Y)(t)$.

For more variables, we define the convolution inductively as

$$(f_{X_1} * \ldots f_{X_n})(t) = \left((f_{X_1} * \ldots f_{X_{n-1}}) * f_{X_n}\right)(t).$$

Another useful observation is that, for X, Y independent we have

$$F_{X+Y}(t) = \int_{-\infty}^{\infty}\int_{-\infty}^{t-y} f_X(x)f_Y(y)dxdy = \int_{-\infty}^{\infty} F_X(t-y)f_Y(y)dy. \tag{6}$$

Example 6.8. *Let $X_1, X_2, \ldots, X_n \sim U[0,1]$ independent random variables. Show that*

$$\Pr[X_1 + \ldots + X_n \leq 1] = \frac{1}{n!}.$$

Proof. We prove the following stronger statement by induction on n: for all $t \in [0,1]$ we have

$$\Pr[X_1 + \ldots + X_n \leq 1] = \frac{t^n}{n!}.$$

For $n = 1$ the statement is trivial. Suppose we know it holds for n and we want to prove it for $n+1$. Let $Y_n = X_1 + \ldots X_n$ and write

$$\Pr[X_1 + \ldots + X_n + X_{n+1} \leq t] = \Pr[Y_n + X_{n+1} \leq t],$$

which by (6) equals

$$\int_0^t F_{Y_n}(t-x) f_{X_{n+1}}(x) dx,$$

which by induction equals

$$\int_0^t \frac{(t-x)^n}{n!} dx = \frac{-(t-x)^{n+1}}{(n+1)!}|_0^t = \frac{t^{n+1}}{(n+1)!}.$$

□

Exercise 6.9. *Prove that if $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$ are indepedent, then*

$$X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2).$$

6.5 Conditional distribution

Let us first start with the discrete case. Suppose that X, Y are (discrete) random variables and suppose that $y \in \mathbb{R}$ is such that $\Pr[Y = y] > 0$. Then, we can easily compute the probability mass function of X *conditioned* on $Y = y$:

$$\Pr_{X|Y=y}[X = x \mid Y = y] = \frac{\Pr[X = x, Y = y]}{\Pr[Y = y]}.$$

Switching to the continuous case we see that the above definition is trickier as $\Pr[Y = y] = 0$ for all y. The way to go around it is as follows: suppose that we wish to compute $\Pr[X \leq x \mid Y = y]$, and that the density function of Y, f_Y is strictly positive at y (that is, $f_Y(y) > 0$). Then, by continuity we have

$$\Pr[X \leq x \mid Y = y] \approx \Pr[X \leq x \mid y \leq Y \leq y + dy] = \int_{-\infty}^{x} \frac{f(s, y)}{f_Y(y)} ds.$$

This leads to the following definition:

Definition 6.10. *For any y for which $f_Y(y) > 0$, the* conditional density function *of X given $Y = y$ is given by*

$$f_{X|Y}(x \mid y) = \frac{f(x, y)}{f_Y(y)}.$$

The conditional distribution function *of X given $Y = y$ is given by*

$$F_{X|Y}(x \mid y) = \int_{-\infty}^{x} f_{X|Y}(s \mid y) ds.$$

Observe that if X, Y are independent then

$$f_{X|Y}(x \mid y) = \frac{f(x, y)}{f_Y(y)} = \frac{f_X(x) f_Y(y)}{f_Y(y)} = f_X(x).$$

6.6 Joint distribution of functions of random variables

We have already seen example of the following form: suppose that X is a random variable and $Y = g(X)$ where g is a "sufficiently nice" function (we haven't defined what "sufficiently nice" means). Then,

$$\Pr[Y \leq y] = \Pr[g(X) \leq y] = \Pr[X \in g^{-1}(-\infty, y)].$$

In particular, if X is continuous with density f_X, then we have that

$$F_Y(y) = \int_{g^{-1}(-\infty,y)} f_X(x)dx,$$

which gives us a "recipe" for calculating the distribution of Y based on X. As a simple example, let's assume that $Y = aX + b$ and we wish to calculate f_Y. Then, observe that

$$F_Y(y) = \begin{cases} F_X((y-b)/a) & \text{if } a > 0 \\ 1 - F_X((y-b)/a) \text{ otherwise} \end{cases}.$$

In particular, by differentiating both sides we obtain that

$$f_Y(y) = |a|^{-1} f_X((y-b)/a).$$

This already gives us a hint how to generalize this argument to joint density, which is basically just to do change of variables in a multpile integral.

So, suppose that X_1 and X_2 are two random variables with joint density $f_{X_1,X_2}(x_1, x_2)$. Let $Y_1 = g_1(X_1, X_2)$ and $Y_2 = g_2(X_1, X_2)$ for some functions g_1 and g_2. How can we find the joint density of Y_1 and Y_2?

Assume that g_1 and g_2 satisfying:

- The transformation $(x_1, x_2) \to (g_1(x_1), g_2(x_2))$ is invertible on the domain of (X_1, X_2), and
- the functions g_1 and g_2 have continuous partial derivative at all points (x_1, x_2) and the *Jacobian* is non-zero on each such point. Recall that the Jacobian of (g_1, g_2) at the point (x_1, x_2) is defined by

$$J(x_1, x_2) = \begin{vmatrix} \frac{\partial g_1}{\partial x_1} & \frac{\partial g_1}{\partial x_2} \\ \frac{\partial g_2}{\partial x_1} & \frac{\partial g_2}{\partial x_2} \end{vmatrix}.$$

Then, one can easily prove (change of variables in a double integral) that

$$f_{Y_1,Y_2}(y_1, y_2) = f_{X_1,X_2}(x_1, x_2)|J(x_1, x_2)|^{-1}.$$

Example 6.11. *Suppose that $X_1, X_2 \sim Exp(\lambda)$ and are independent. Let $Y_1 = X_1 + X_2$ and $Y_2 = \frac{X_1}{X_2}$. We wish to find the joint density of Y_1, Y_2 and to show that they are also independent.*

Clearly, we have that

$$J(x_1, x_2) = \begin{vmatrix} 1 & 1 \\ \frac{1}{x_2} & -\frac{x_1}{x_2^2} \end{vmatrix} = -\frac{x_1}{x_2^2} - \frac{1}{x_2} = -\frac{x_1 + x_2}{x_2^2}.$$

Therefore, we have that

$$f_{Y_1,Y_2}(y_1, y_2) = \frac{x_2^2}{x_1 + x_2} f_{X_1,X_2}(x_1, x_2).$$

Note that $X_1 = Y_1Y_2/(1+Y_2)$, $X_2 = Y_1/(1+Y_2)$ and that $f_{X_1,X_2}(x_1, x_2) = f_{X_1}(x_1)f_{X_2}(x_2) = \lambda^2 e^{-\lambda(x_1+x_2)}$, we obtain that

$$f_{Y_1,Y_2}(y_1, y_2) = \lambda^2 y_1 e^{-\lambda y_1} \cdot \frac{1}{(1 + y_2)^2}.$$

Since the above expression factorizes as $g(y_1)h(y_2)$ we obtain that Y_1 and Y_2 are independent.

7 Properties of expectation

In this chapter we will explore more properties and results coming from the expectation of random variables.

7.1 Expectation of sums of random variables

We've already seen in Theorem 3.3 that the expectation of discrete random variables is *linear*. That is,

$$\mathbb{E}[a_0 + \sum_{i=1}^{n} a_i X_i] = a_0 + \sum_{i=1}^{n} a_i \mathbb{E}[X_i].$$

It is not hard to prove that the same holds also for continuous random variables. Moreover, instead of taking the sum of random variables, one can consider any other function $g(X_1, \ldots, X_n)$, and for this it is not hard to prove that

$$\mathbb{E}[g(X_1, \ldots, X_n)] = \int \ldots \int g(x_1, \ldots, x_n) f(x_1, \ldots, x_n) dx_1 \ldots dx_n,$$

where f is the joint density function of $(X_1, \ldots, X_n)$.

Another useful (and yet, simple) property of expectation is *monotonicity*. That is, suppose that X is a random variable for which, with probability 1, we have that $a \leq X \leq b$, then $a \leq \mathbb{E}[X] \leq b$. More generally, suppose that X and Y are two random variables for which $X \geq Y$ on each point of the sampling space. Then, $\mathbb{E}[X] \geq \mathbb{E}[Y]$ (prove these properties!).

In this section, along with the homework, we will see few neat applications for these two simple properties (linearity, and monotonicity).

7.2 Application – random walk on the plane

This is example 2ℓ from Chapter 7 in the book.

Consider a particle initially located at a given point in the plane, and suppose that it moves into a new position after each step by moving one unit of distance from the previous position and at an angle of orientation from the previous position that is uniformly distributed over $[0, 2\pi)$. Compute the expected square of the distance from the origin after n steps.

Proof. Let (X_i, Y_i) denote the change of the position at the ith step. That is, if at the end of step $i-1$st the particle was at some position (x, y), then after the ith step it moves to $(x + X_i, y + Y_i)$. In particular, we have that $X_i = \cos(\theta_i)$ and $Y_i = \sin(\theta_i)$ where $\theta_i \sim U([0, 2\pi))$ is the angle chosen at step i.

Let $X = \sum_{i=1}^{n} X_i$ and $Y = \sum_{i=1}^{n} Y_i$ be the position of the particle after the nth step. Then, letting D be the distance of the particle from the origina, we have that

$$D^2 = X^2 + Y^2 = \sum_{i=1}^{n} (\cos(\theta_i)^2 + \sin(\theta_i)^2) + \sum_{i \neq j} (2 \cos\theta_i \cos\theta_j + 2 \sin\theta_i \sin\theta_j).$$

Therefore, we have that

$$\mathbb{E}[D^2] = n.$$

□

7.3 Applications – Probabilistic method

7.3.1 Van der Waerden's theorem

As a first example, we consider an application about coloring integers. Recall that an *arithmetic progression* is a subset of "equally spaced" numbers. More precisely, a *k-term arithmetic progression* (or a k-AP for short) is a subset of the form

$$\{a, a+b, \ldots, a+(k-1)b\},$$

where a, b are some fixed integers.

A *coloring* of a subset $S \subseteq \mathbb{N}$ is a function $c : S \to \{1, \ldots, C\}$, where the range is considered as the set of *colors*. A 2-coloring of S is a coloring with $C = 2$, where the colors 1 and 2 can be considered as "red" and "blue".

The following theorem is known as "Van der Waerden's Theorem". It is an example of a result in the so-called area nowadays known as *Ramsey Theory*.

Theorem 7.1. *For every $k \in \mathbb{N}$, there exists an $n \in$ such that for every 2-coloring of $\{1, 2, \ldots, n\}$, there exists some k-AP which is monochromatic (that is, a k-AP for which all its elements have the same color).*

We won't prove this theorem in these notes as the proof has nothing to do with probability, but feel free to google it and read its proof.

Now, let $W(k)$ be the smallest possible choice of n in the theorem, for a particular value of k. That is, $W(k)$ is the smallest value of n for which any 2-coloring of the numbers $\{1, \ldots, n\}$. Alternativey, $W(k) - 1$ is the largest value of n for which **there exists** a 2-coloring of $\{1, 2, \ldots, n\}$ with **no** monochromatic k-AP.

For example, it is not hard to show (do it!) that $W(2) = 3$. It also turns out that $W(3) = 9$ which is not as easy to prove.

These $W(k)$s are known as *van der Waerden numbers* and have received a lot of attention. Unfortunately, not too much is known about them, and in particular we only have exact values for $W(k)$ for fairly small values of k.

Here we are going to prove a *lower bound* for $W(k)$:

Theorem 7.2. *For any $k \in \mathbb{N}$,*

$$W(k) \geq \sqrt{k-1} \cdot 2^{(k-1)/2}.$$

Proof. Let n be the largest integer strictly smaller than $\sqrt{k-1} \cdot 2^{(k-1)/2}$. Our goal is to show that there exists a 2-coloring of $\{1, 2, \ldots, n\}$ with *no* monochromatic k-APs. The main point is that we will not explicitly construct such a coloring (and in fact, such a construction is unknown!). Instead, we will prove that a *randomly chosen* coloring has a positive probability of being good. In particular, this implies that there exists such a coloring. Indeed, if there was no such coloring, the chance of getting such a coloring would have to be zero! This approach is known as "the probabilistic method".

So let us take a random coloring of $\{1, 2, \ldots, n\}$ by assigning, for each $1 \leq i \leq n$, a color $C(i) \in \{1, 2\}$ uniformly at random, where all the $C(i)$s are independent random variables. Now define the random variable

$$X = \#\text{monochromatic } k\text{-APs in the set } \{1, \ldots, n\}.$$

Our goal is to show that $\Pr[X = 0] > 0$. The strategy is quite simple. We are going to show that $(X) < 1$, and then, by the following lemma we will obtain the desired.

Lemma 7.3. *Let Y be any random variable taking on only nonnegative integer values, and satisfying $\mathbb{E}[Y] < 1$. Then $\Pr(Y = 0) > 0$.*

Proof. We have

$$\mathbb{E}[Y] = \sum_{i=0}^{\infty} i \cdot \Pr[Y = i] \geq \sum_{i=1}^{\infty} \Pr[Y = i] = 1 - \Pr[Y = 0].$$

Rearranging,

$$\Pr[Y = 0] \geq 1 - \mathbb{E}[Y] > 0.$$

□

In order to compute $\mathbb{E}[X]$ we will use linearity of expectation. Define, for any $a, b \in \mathbb{N}$ for which $A_{ab} := \{a, a+b, \ldots, a+(k-1)b\} \subseteq [n]$, an indicator random variable X_{ab} for which $X_{ab} = 1$ if and only if the corresponding k-AP is monochromatic.

Then

$$X = \sum_{(a,b):A_{ab} \subseteq \{1,\ldots,n\}} X_{ab}.$$

Thus, using linearity of expectation,

$$\begin{aligned}\mathbb{E}[X] &= \sum_{(a,b):A_{ab} \subseteq \{1,\ldots,n\}} \mathbb{E}[X_{ab}] \\ &= \sum_{(a,b):A_{ab} \subseteq \{1,\ldots,n\}} \Pr[A_{ab} \text{ is monochromatic}].\end{aligned}$$

Clearly, for any a, b we have

$$\Pr[A_{ab} \text{ is monochromatic}] = \frac{1}{2^k} \cdot 2 = \frac{1}{2^{k-1}}.$$

The number of k-APs is easily upper bounded by $n \cdot \frac{n}{k-1} = \frac{n^2}{k-1}$ (n choices for a and at most $n/(k-1)$ choices for b such that $a + (k-1)b \leq n$).

All in all, by linearity of expectation we have

$$\mathbb{E}[X] \leq \frac{n^2}{k-1} \cdot \frac{1}{2^{k-1}},$$

which, by the way we defined n, is smaller than 1. This completes the proof. □

Note that this proof does not tell us *how* to find such a coloring (except perhaps by trying random colorings until we find one that works), and doesn't tell us what these colorings look like. There are many examples where we can prove existence of certain structures, using the probabilistic method, but have no explicit deterministic construction! The statements proved with the method have (typically) nothing to do with probability; probability is introduced for the purpose of the proof.

7.3.2 Ramsey numbers

In the 1950s, the Hungarian sociologist S. Szalai studied friendship relationships between children. He noticed that among 20 children, he was always able to find four children each pair of which were friends, or four children such that no two of them were friends. Before deducing any sociological conclusions, Szalai asked three distinguished mathematicians from Hungary: Erdős, Turán, and Sós who noticed that indeed this is a mathematical phenomenon and not a sociological one. To see this, form the friendship graph between children, any two children that are friends form an edge. In any graph on 20 vertices there are four vertices which form a clique (a subset of vertices, each pair of which form an edge) or an independent set (a subset of vertices, no pair of which form an edge). Let's look at a smaller case:

Proposition 7.4. *Among any six people, there are three any two of whom are friends, or there are three such that no two of them are friends.*

This is not a sociological claim, but a very simple graph-theoretic statement. In other words, in any graph on 6 vertices, there is a triangle or three vertices with no edges between them.

Proof. Let $G = (V, E)$ be a graph and $|V| = 6$, i.e. G has six vertices. Fix a vertex $v \in V$. We consider two cases.

Case 1: The degree of v is at least 3. In this case, consider three neighbors of v, call them x, y, z. If any two among $\{x, y, z\}$ are friends, we are done because they form a triangle together with v. If not, no two of $\{x, y, z\}$ are friends and we are done as well.

Case 2: The degree of v is at most 2, then there are at least three other vertices which are not neighbors of v, call them x, y, z. In this case, the argument is complementary to the previous one. If $\{x, y, z\}$ are mutual friends, then we are done. Otherwise, there are two among $\{x, y, z\}$ who are not friends, for example x and y, and then no two of $\{v, x, y\}$ are friends.

□

The number 6 in the above proposition cannot be replaced by 5. This is because the 5-cycle C_5 has neither a triangle nor three vertices with no edges between them.

The above proposition is a special case of *Ramsey's theorem* proved in 1930, which is a foundational result in *Ramsey theory*. This consists of a large body of deep results in mathematics that which roughly say, according to Motzkin, that "complete disorder is impossible." In other words, any very large system contains a large well-organized subsystem.

Definition 7.5. *The Ramsey number $R(s, t)$ is the minimum n such that ever graph with n vertices contains a clique of order s or an independent set of order t.*

So Proposition 7.4 can be stated as $R(3, 3) = 6$.

By replacing a graph by its complement, we can deduce that $R(s, t) = R(t, s)$. Further, we have $R(2, t) = t$ as any graph on t vertices either contains an edge or is an independent set of order t, and any smaller empty graph has neither an edge nor an independent set of order t.

Theorem 7.6. *For any positive integers s and t, the Ramsey number $R(s, t)$ exists. Further, it satisfies*

$$R(s, t) \leq \binom{s + t - 2}{s - 1}.$$

The bound given here is due to Erdős and Szekeres, and is considerably better then the bound in Ramsey's proof.

Proof. We claim that

$$R(s, t) \leq R(s - 1, t) + R(s, t - 1), \tag{7}$$

and then deduce the desired theorem. To show this, let $n = R(s - 1, t) + R(s, t - 1)$, and consider any graph G on n vertices. Fix any vertex v. We consider two cases:

Case 1: The degree of v is at least $R(s - 1, t)$. Then, by the definition of $R(s, t - 1)$, the set of neighbors of v either contains a clique of order $s - 1$, or an independent set of order t. In the second case, we are done as this is an independent set in G. In the first case, we can extend the clique by adding v, and hence G contains a clique of order s, completing this case.

Case 2: The degree of v is at most $R(s - 1, t) - 1$. In this case, v has at least $n - 1 - (R(s - 1, t) - 1) = R(s, t - 1)$ nonneighbors. Then, by the definition of $R(s, t - 1)$, the set of nonneighbors of v either contains a clique of order s, or an independent set of order $t - 1$. In the first case, we are done as this is a clique in G. In the second case, we can extend the independent set by adding v, and hence G contains an independent set of order t, completing the proof of the claim.

Given (7), it follows by induction that these Ramsey numbers are finite. Moreover, we get an explicit bound. First $R(s, t) \leq \binom{s+t-2}{s-1}$ holds in the base case $s = 1$ or $t = 1$ since every graph contains a clique of order 1 and an independent set of order 1. The inductive step is:

$$R(s,t) \le R(s-1,t) + R(s,t-1) \le \binom{s+t-3}{s-2} + \binom{s+t-3}{s-1} = \binom{s+t-2}{s-1},$$

where the equality is Pascal's identity for binomial coefficients. □

How good is the above bound for the diagonal case $s = t$? We get the upper bound

$$R(s,s) \le \binom{2s-2}{s-1} \le \frac{4^s}{\sqrt{s}}.$$

This upper bound has not been significantly improved in roughly 70 years! All we know currently is that the exponential growth is the right order of magnitude, but the base of the exponential is not known. The following is an old lower bound of Erdős. Note that to get a lower bound, we need to show that there is a large graph without cliques and independent sets of a certain order. This is quite difficult to achieve by an explicit construction. (The early lower bounds on $R(s,s)$ were only polynomial in s.)

The amazing thing about Erdős' proof below is that he never presents a specific graph. He simply shows that one exists by considering a *random* graph almost always works. This was one of the first occurences of the *probabilistic method* in combinatorics. The probabilistic method has been used in discrete mathematics ever since with phenomenal success.

Theorem 7.7. *For $s \ge 3$,*

$$R(s,s) \ge 2^{s/2}.$$

Proof. Let n be the largest integer less than $2^{s/2}$. Consider a random graph G on n vertices, where each pair is an edge with probability $1/2$ chosen independently from the other edges. For any particular set S of s vertices, the probability that S forms a clique is $2^{-\binom{s}{2}}$, and the probability that S forms an independent set is $2^{-\binom{s}{2}}$. Since these are disjoint events, the probability that S forms a clique or independent set is $2^{1-\binom{s}{2}}$. The number of such sets S on s vertices is $\binom{n}{s}$. By linearity of expectation, the expected number of cliques or independent sets in the random graph G is

$$\binom{n}{s} 2^{1-\binom{s}{2}} = \frac{n!}{s!(n-s)!} \frac{2}{2^{s(s-1)/2}} < \frac{2n^s}{s! 2^{s(s-1)/2}} \le \frac{2^{1+s/2}}{s!} < 1,$$

the inequality $\le$ comes from the constraint on n and s. Since the number of cliques or independent sets of order s is a nonnegative integer, and the expected (i.e. average) number is less than one, the probability that that there is no clique or independent set of order s must be positive. Thus there must be a graph on n vertices (a point in our sample space) with no clique or independent set of order n. We conclude that $R(s,s) \ge 2^{s/2}$. □

Determining Ramsey numbers exactly, even for rather small values of s, is a notoriously difficult problem. It is known that $R(4,4) = 18$, but even $R(5,5)$ is not known (it is known to be between 43 and 49), and determining $R(6,6)$ seems hopeless (it is between 102 and 165).

7.3.3 Independent sets in graphs

Recall that a graph $G = (V,E)$ consists of a set of *vertices* V, and a set of *edges* E, where each edge is an unordered pair of vertices. For every vertex $v \in V$, we let $d(v)$ denote its *degree*; that is, $d(v)$ is the number of edges $e \in E$ for which $v \in e$. The following simple exercise is useful for us:

Exercise 7.8. *Prove that for every graph $G = (V,E)$ we have that $\sum_{v \in V} d(v) = 2|E|$.*

Given a graph $G = (V,E)$, a subset $X \subseteq V$ is called an *independent set* if and only if it contains no edges of G. We also let $\alpha(G)$ be the size of the largest independent set in G.

In this section we prove the following (deterministic) theorem using the probabilistic method:

Theorem 7.9. *Let $G = (V, E)$ be a graph on n vertices and with $|E| = nd/2$, where $d \geq 1$. Then, $\alpha(G) \geq \frac{n}{2d}$.*

Proof. Let $S \subseteq V$ be a random subset defined by

$$\Pr[v \in S] = p,$$

independently at random, for all $v \in V$ (p will be determined shortly).

Consider the random variable $|S|$ (that is, the size of the random subset S), and $Y=$ the number of edges within S. For each edge $e = xy \in E$, we let Y_e be an indicator random variable for the event '$e \subseteq S$'. That is, $\Pr[Y_e = 1] = p^2$ for all $e \in E$. Observe that

$$Y = \sum_{e \in E} Y_e,$$

and therefore, by linearly of expectation we have that

$$\mathbb{E}[Y] = |E|p^2.$$

Moreover, as we clearly have

$$\mathbb{E}[|S|] = np,$$

(WHY?) it follows that

$$\mathbb{E}[|S| - Y] = np - |E|p^2 = np - \frac{ndp^2}{2}.$$

We can now determine $p = \frac{1}{d}$ (which maximizes the right hand side of the above equation) and obtain

$$\mathbb{E}[|S| - Y] = \frac{n}{2d}.$$

In particular, it means that there exists a specfic set $S \subseteq V$ for which $|S| - Y \geq \frac{n}{2d}$. In words, it means that there exists a subset S for which the number of its vertices minus the number of the edges that it contain is at least $n/2d$. Clearly, by deleting one vertex from each edge we "kill" all the edges and are left with an independent set. Therefore, there exists an independent set of size at least $n/2d$ as desired. □

7.4 Expectation and independence

An important property of expectation is *multiplicativity* for independent factors. It is not hard to find random variables X, Y for which $\mathbb{E}[XY] \neq \mathbb{E}[X]\mathbb{E}[Y]$. Apparently, for independent random variables n even stronger property holds:

Theorem 7.10. *Two random variables X, Y are independent if and only if for every two 'nice' function f, g we have that*

$$\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)]\mathbb{E}[g(Y)].$$

Proof. Let's start with the (perhaps) easier direction: suppose that for every two 'nice' function f, g we have that

$$\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)]\mathbb{E}[g(Y)],$$

and we wish to prove that X and Y are indepdent. Recall that X, Y are independent if and only if

$$F(x, y) = F_X(x)F_Y(y)$$

for all x, y.

Fix x, y and let I_x and I_y be the indicator random variables for the events

$$`X \leq x' \text{ and } `Y \leq y', \text{ respectively.}$$

By the assumption we have

$$F(x,y) = \Pr[X \leq x, Y \leq y] = \mathbb{E}[I_x I_y] = \mathbb{E}[I_x]\mathbb{E}[I_y] = F_X(x)F_Y(y)$$

as desired.

For the other direction, observe that

$$\begin{aligned}
\mathbb{E}[f(X)g(Y)] &= \int\int f(x)g(y)f_{X,Y}(x,y)dxdy \\
&= \int\int g(y)f_X(x)f_Y(y)dxdy \\
&= \left(\int f(x)f_X(x)dx\right)\left(\int g(y)f_Y(y)dy\right) \\
&= \mathbb{E}[f(X)]\mathbb{E}[g(Y)].
\end{aligned}$$

This completes the proof. □

The following exercise will be very important for us later in this course.

Exercise 7.11. *Suppose that $X_1, \ldots, X_n \sim Ber(1/2)$ and are independent. Calculate*

$$\mathbb{E}[e^{t\sum_{i=1}^n X_i}].$$

7.5 Variance of sum of random variables

Observe that if X, Y are independent, then by Theorem 7.10 we have that

$$\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0.$$

Observe that this condition is **not** equivalent to independence, as for example, if we take $\Pr[X = 1] = \Pr[X = 0] = \Pr[X = -1] = \frac{1}{3}$, and $Y = X^2$, it is clear that X and Y are not independent (prove it!) but

$$\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \mathbb{E}[X^3] - \mathbb{E}[X]\mathbb{E}[X^2] = 0.$$

In general, even though $\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0$ is not equivalent to independence, it does give us some correlation between the variables.

Definition 7.12. *Given random variables X, Y, we define their* covariance *by*

$$COV(X,Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])].$$

As an easy exercise, convince yourselves that

$$COV(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

We say that X, Y are *uncorrelated* if and only if $COV(X,Y) = 0$ (in particular, if X and Y are independent, then they are also uncorrelated).

Note that if X and Y are indicators random variables for events A and B, respectively, then

$$COV(X,Y) = \Pr[A \cap B] - \Pr[A]\Pr[B] = \Pr[A]\left(\Pr[B \mid A] - \Pr[B]\right).$$

In particular, if $COV(X,Y) > 0$ then the events A and B are *positively correlated* (in the sense that knowing A increases the probability that B occurs) and if $COV(X,Y) < 0$ then they are *negatively correlated*.

For general random variables X and Y, if $COV(X,Y) > 0$ we say that X and Y are *positively correlated* in the sense that "on average", increasing Y will result in a larger X, and if $COV(X,Y) < 0$ we say that X and Y are *negatively correlated* in the sense that increasing Y, on average, will result in a smaller X.

We will get back to these definitions later in this course, and we now try to show the connection between the covariance and the the variance of a sum of random variables.

First, let us summarize some basic properties of covariance, and the proofs are left as easy exercises:

Proposition 7.13. *The following properties hold:*

1. $COV(X,Y) = COV(Y,X)$
2. $COV(X,X) = Var(X)$
3. $COV(aX,Y) = aCOV(X,Y)$ *for all* $a \in \mathbb{R}$
4. $COV(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j) = \sum_{i=1}^{n} \sum_{j=1}^{m} COV(X_i, Y_j)$.

Now we are ready to prove the following theorem:

Theorem 7.14. *Let $X_1, \ldots, X_n$ be random variables. Then,*

$$Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} Var(X_i) + 2 \sum_{1 \le i < j \le n} COV(X_i, X_j).$$

Proof. Recall that by the definition of the variance, linearity of expectation, and by Proposition 7.13 we have that

$$\begin{aligned}
Var(\sum_i X_i) &= \mathbb{E}[(\sum_i X_i - \sum_i \mathbb{E}[X_i])^2] \\
&= \mathbb{E}[(\sum_i (X_i - \mathbb{E}[X_i]))^2] \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])] \\
&= \sum_i Var(X_i) + \sum_{i \ne j} COV(X_i, X_j) \\
&= \sum_{i=1}^{n} Var(X_i) + 2 \sum_{1 \le i < j \le n} COV(X_i, X_j).
\end{aligned}$$

This completes the proofs. □

7.6 Conditional expectation

Recall that if X and Y are jointly discrete random variables, then the conditional probability mass function of X given that $Y = y$, is defined, for all y such that $\Pr[Y = y] > 0$, by

$$p_{X|Y}(x|y) := \Pr[X = x \mid Y = y] = \frac{p(x,y)}{p_Y(y)}.$$

Moreover, observe that $X \mid Y = y$ is a random variable and that $p_{X|Y}(x|y)$ is its probability mass function. Therefore, its expectation is

$$\mathbb{E}[X \mid Y = y] = \sum_x x p_{X|Y}(x|y).$$

For continuous random variables X and Y we defined the conditional density of X given $Y = y$ by

$$f_{X|Y}(x \mid y) = \frac{f(x,y)}{f_Y(y)}.$$

In particular, we obtain that

$$\mathbb{E}[X \mid Y = y] = \int_{-\infty}^{\infty} x f_{X|Y}(x|y) dx,$$

provided that $f_Y(y) > 0$ (otherwise it is undefined). Now, given X and Y, we can define the random variable $\mathbb{E}[X \mid Y]$ which is a function of Y and its value at $Y = y$ is given by $\mathbb{E}[X \mid Y = y]$. Having defined this random variable, we are ready to state (and prove) the following theorem which is one of the most important properties of conditional expectation.

Theorem 7.15. *For every two random variables X and Y we have that*

$$\mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}[X].$$

Note that in the discrete case the theorem asserts that

$$\mathbb{E}[X] = \sum_y \mathbb{E}[X \mid Y = y] \Pr[Y = y].$$

We now prove the theorem for the continuous setting while the discrete case is being left as an exercise.

Proof. Note that

$$\begin{aligned}
\mathbb{E}[\mathbb{E}[X \mid Y]] &= \int_{-\infty}^{\infty} \mathbb{E}[X \mid Y = y] f_Y(y) dy \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{X|Y}(x|y) f_Y(y) dx dy \\
&= \int_{-\infty}^{\infty} x \left(\int_{-\infty}^{\infty} f(x,y) dy \right) dx \\
&= \int_{-\infty}^{\infty} x f_X(x) dx \\
&= \mathbb{E}[X].
\end{aligned}$$

□

Example 7.16. *A miner is trapped in a mine containing 3 doors. The first door leads to a tunnel that will take him to safety after 3 hours of travel. The second door leads to a tunnel that will return him to the mine after 5 hours of travel. The third door leads to a tunnel that will return him to the mine after 7 hours. If we assume that the miner is at all times equally likely to choose any one of the doors, what is the expected length of time until he reaches safety?*

Proof. Let X denote the amount of time (in hours) until the miner reaches safety, and let Y denote the door he initially chooses.

Now,

$$\begin{aligned}\mathbb{E}[X] &= \mathbb{E}[X \mid Y=1]\Pr[Y=1] + \mathbb{E}[X \mid Y=2]\Pr[Y=2] + \mathbb{E}[X \mid Y=3]\Pr[Y=3] \\ &= 3\Pr[Y=1] + (5+\mathbb{E}[X])\Pr[Y=2] + (7+\mathbb{E}[X])\Pr[Y=3].\end{aligned}$$

To complete the proof, just isolate $\mathbb{E}[X]$ in the above expression and recall that $\Pr[Y=1] = \Pr[Y=2] = \Pr[Y=3] = \frac{1}{3}$. □

Example 7.17 (Expectation of a sum of a random number of random variables)**.** *Suppose that the number of people entering a store on a given day is a random variable with expectation* 50*. Suppose further that the amounts of money spent by these customers are independent random variables having a common expectation of* 8 *USD. Finally, suppose that the amount of money spent by a customer is independent of the total number of customers who enter the store. What is the expected amount of money spent in the store at a given day?*

Proof. Let N be the number of customers entering the store on a given day, and let X_i be the amount of money spent by Customer i. Then, the total amount of money spent by the customers is $\sum_{i=1}^{N} X_i$. Now,

$$\mathbb{E}[\sum_{i=1}^{N} X_i] = \mathbb{E}[\mathbb{E}[\sum_{i=1}^{N} X_i \mid N]],$$

and observe that

$$\mathbb{E}[\sum_{i=1}^{N} X_i \mid N=n] = \mathbb{E}[\sum_{i=1}^{n} X_i] = n\mathbb{E}[X_1] = 8n.$$

In particular, this can be written as

$$\mathbb{E}[\sum_{i=1}^{N} X_i \mid N] = N\mathbb{E}[X_i].$$

Therefore, we conclude that

$$\mathbb{E}[\sum_{i=1}^{N} X_i] = \mathbb{E}[N\mathbb{E}[X_1]] = \mathbb{E}[N]\mathbb{E}[X_1] = 50 \cdot 8 = 400.$$

This completes the proof. □

7.7 Moment generating function

The kth *moment* of a random variable X is defined as $\mathbb{E}[X^k]$. Moments of random variables encode important information about their distribution. For example, the first moment ($k=1$) is just the expectation, and the second moment is closely related to the variance.

In this chapter we will learn a new quantity, called the *moment generating function* of X which encodes information of all of these moments in some way. Our starting point is the following fact from Calculus about the Taylor's expansion of the exponential function:

$$e^{tx} = \sum_{k=0}^{\infty} \frac{(tx)^k}{k!}.$$

Now, suppose that $t \neq 0$ is some fixed real and that X is some random variable, then we obtain

$$\mathbb{E}[e^{tX}] = \sum_{k=0}^{\infty} \frac{t^k}{k!} \cdot \mathbb{E}[X^k].$$

Let us now define the *moment generating function of* X as

$$M(t) := \mathbb{E}[e^{tX}].$$

To justify the name "moment generating function", observe that by successively differentiating $M(t)$ and then evaluating at $t = 0$ we can obtain all the moments of X. Indeed,

$$M^{(k)}(t) = \frac{d\mathbb{E}[e^{tX}]}{dt^k} = \mathbb{E}\left[\frac{de^{tX}}{dt^k}\right] = \mathbb{E}[X^k e^{tX}],$$

and therefore,

$$M^{(k)}(0) = \mathbb{E}[X^k].$$

The main application that we will see for moment generating function is in proving Chernoff's bounds on the next chapter. First, we will give some example for calculating it for general distributions.

Bernoulli distribution Let $X \sim Ber(p)$. Then,

$$M_X(t) = \mathbb{E}[e^{tX}] = (1-p)e^0 + pe^t = 1 + p(e^t - 1).$$

Binomial distribution Let $X \sim Bin(n, p)$. Recall that one can write

$$X = \sum_{i=1}^{n} X_i,$$

where the X_is are iid random variables with distribution $Ber(p)$. Therefore, we have that

$$M_X(t) = \mathbb{E}[e^{tX}] = \mathbb{E}[\prod_{i=1}^{n} e^{tX_i}].$$

Now, since the X_is are independent and the fact that we've already computed the generating function for Bernoulli r.v, we have

$$M_X(t) = \prod_{i=1}^{n} \mathbb{E}[e^{tX_i}] = \prod_{i=1}^{n} M_{X_i}(t) = (1 + p(e^t - 1))^n$$

.

Exercise 7.18. *Calculate the variance of $Bin(n, p)$ from its generating function.*
Hint: differentiate once to obtain $\mathbb{E}[X]$, and then differentiate again to obtain $\mathbb{E}[X^2]$.

Normal distribution Suppose that $Z \sim N(0, 1)$. Then,

$$\begin{aligned} M_Z(t) = \mathbb{E}[e^{tZ}] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx} e^{-x^2/2} dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\frac{-(x-t)^2}{2}} e^{t^2/2} dx \\ &= e^{t^2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-s^2/2} ds \\ &= e^{t^2/2}. \end{aligned}$$

(in the third line we had a change of variable $x - t \to s$).

Exercises 7.19. *Calculate $M_X(t)$ where*

1. $X \sim N(\mu, \sigma^2)$.
2. $X \sim Exp(\lambda)$.
3. $X \sim Geo(p)$.
4. $X \sim U[a, b]$ *(continuous uniform distribution).*

A very important result is that the moment generating function of a random variable X uniquely determines its distribution. That is, if $M_X(t)$ exists and is finite in some region around $t = 0$, then the distribution of X is uniquely determined. For example, if some given random variable X has a moment generating function $e^{t^2/2}$, then we know that X has a standard normal distribution. This result will be extremely useful in the chapter about limit theorems.

Exercises 7.20. 1. *Suppose that $M_X(t) = e^{7t}$. What is $\mathbb{E}[X^2]$?*

2. *Suppose that $M_X(t) = e^{t^2/2}$, what is* $\Pr[-1 \leq X \leq 1]$?

8 Practice problems for Midterm 2

1. Every day, Alice comes to the bus stop exactly at 7am. She takes the first bus that arrives. The arrival of the first bus is an exponential random variable with expectation 20 minutes. Also, every working day, and independently, Bob comes to the same bus stop at a random time, uniformly distributed between 7 and 7:30am.

   (a) (10 points) What is the probability that tomorrow Alice will wait for more than 30 minutes?

   (b) (15 points) Assume day-to-day independence. Consider Bob late if he comes after 7:20. What is the probability that Bob will be late on 2 or more working days among the next 10 working days?

   (c) (15 points) What is the probability that Alice and Bob will meet at the station tomorrow?

2. Let X, Y, and Z be independent random variables, each of which is uniformly distributed on the interval $[0, 1]$.

   (a) (15 points) Find the joint density function of XY and Z^2, and compute $\Pr[XY < Z^2]$.

   (b) (15 points) Compute $Var(XY + Z)$.

3. (30 points) Let K_n be the *complete graph* on n vertices; that is, its edge-set consists of all $\binom{n}{2}$ possible unordered pairs of vertices. Suppose that some coloring of the edge-set of K_n is given. A triangle is called *rainbow* if it has at most one edge from each color. Show that there exists a coloring of the edges of K_n using three colors with at least $\binom{n}{3}\frac{2}{9}$ rainbow triangles.

4. Let X, Y be independent exponential random variables with parameter $\lambda = 1$. Find the joint density function of $U = X + Y$ and $V = X/(X + Y)$, and prove that V is uniformly distributed on the interval $[0, 1]$.

5. The joint density of X, Y is given by

$$f(x, y) = \begin{cases} 3x & \text{if } 0 \leq y \leq x \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

(a) Find the marginal densities f_X and f_Y.

(b) Compute $COV(X, Y)$.

(c) Compute $Var(X + Y)$.

6. Show that there exists some constant $c > 0$ such that for all n sufficiently large, there exists a subset $S \subseteq \{1, \ldots, n\}$ of size at least $|S| \geq cn^{1/3}$ such that S does not contain non-trivial solutions to $a + b = c + d$ (where $a, b, c, d \in S$, $a \neq b$, $c \neq d$, and $\{a, b\} \neq \{c, d\}$). That is, show that no two pairs of elements of S have the same sum.

7. Let $G = (V, E)$ be a simple graph. Recall that an independent set in G is a set of vertices, no two of which have an edge between them. The *independence number* $\alpha(G)$ is the size of the largest independent set in G. Here we'll prove that

$$\alpha(G) \geq \sum_{v \in V(G)} \frac{1}{d(v) + 1}. \tag{8}$$

(a) Choose an ordering $v_1, \ldots, v_n$ of $V(G)$ uniformly at random. Let S be the set of all vetices that appear before all of their neighbors in the ordering. Show that S is an independent set.

(b) For each vertex v, let X_v be the indicator random variable which has value 1 if $v \in S$ and 0 otherwise. Compute $E[X_v]$. Use this to compute $E[|S|]$ and deduce (8).

9 Concentration inequlities

In this section we will learn basic techniques to prove that certain random variables (in particular, random variables that can be represented as a sum of "many" random variables with "not too many" dependencies) typically have values which are "close" to their expectation. Such results are referred to as *concentration inequalities* because they measure how much a random variable is concentrated around its mean.

9.1 Markov's inequality

Let us start with the most basic inequality, due to Markov, which is the key ingredient in all the extensions that we will see below.

Theorem 9.1 (Markov's inequality)**.** *Let X be a random variable for which $\Pr[X \geq 0] = 1$. Then, for every $k > 0$ we have that*

$$\Pr[X \geq k] \leq \frac{\mathbb{E}[X]}{k}.$$

Observe that the above bound is non-trivial only in case that $k > \mathbb{E}[X]$.

Proof. We prove it for the case that X has a continuous distribution. The discrete case is left as an exercise. First, observe that

$$\mathbb{E}[X] = \int_0^\infty x f_X(x) dx \geq \int_k^\infty k f_X(x) dx.$$

Indeed, since $\Pr[X \geq 0] = 1$ we can start integrating from 0 (and not from $-\infty$), and since $x f_X(x)$ is positive, by monotonicity of the integral (meaning, we calculate a smaller area) we clearly have that

$$\int_0^\infty x f_X(x) dx \geq \int_k^\infty x f_X(x) dx.$$

Then, the above inequality is obtained by observing that for $x \geq k$ we have that

$$x f_X(x) \geq k f_X(x).$$

Now, recall that, by definition, we have

$$\int_k^\infty f_X(x)dx = \Pr[X \geq k],$$

and therefore, we obtain that

$$\mathbb{E}[X] \geq k \Pr[X \geq k].$$

The desired result is obtained by dividing both sides by k. This completes the proof. □

Example 9.2. *Suppose that $X \sim Bin(n, \frac{1}{4})$. Then we have that $\mathbb{E}[X] = \frac{n}{4}$, and X is non-negative by definition. Therefore, by Markov's inequality we obtain that*

$$\Pr[X \geq \frac{n}{2}] \leq \frac{\mathbb{E}[X]}{n/2} = \frac{1}{2}.$$

9.2 Chebyshev's inequality

Now, based on Markov's inequality, we can prove way stronger bounds (in particular for binomial random variables). The first inequality we want to consider is the following classical inequality due to Chebyshev.

Theorem 9.3 (Chebyshev's inequality)**.** *Let X be a random variable and $t > 0$ be any positive real. Then,*

$$\Pr[|X - \mathbb{E}[X]| \geq t] \leq \frac{Var(X)}{t^2}.$$

Proof. First, observe that

$$\Pr[|X - \mathbb{E}[X]| \geq t] = \Pr[(X - \mathbb{E}[X])^2 \geq t^2].$$

Now, considering the random variable $Y = (X - \mathbb{E}[X])^2$, we clearly have that $\Pr[Y \geq 0] = 1$, and therefore, by Markov's inequality we obtain

$$\Pr[Y \geq t^2] \leq \frac{\mathbb{E}[Y]}{t^2} = \frac{\mathbb{E}[(X - \mathbb{E}[X])^2]}{t^2} = \frac{Var(X)}{t^2}.$$

This completes the proof. □

Let us illustrates its strength by the following example:

Example 9.4. *Suppose that $X \sim Bin(n, \frac{1}{4})$. Then we have that $\mathbb{E}[X] = \frac{n}{4}$ $Var(X) = \frac{3n}{16}$. Now,*

$$\Pr[X \geq \frac{n}{2}] \leq \Pr[|X - \mathbb{E}[X]| \geq \frac{n}{4}],$$

and therefore, by Chebyshev's inequality we obtain that

$$\Pr[X \geq \frac{n}{2}] \leq \frac{Var(X)}{n^2/16} \leq \frac{C}{n},$$

for some absolute constant C that doesn't depend on n. Compare this bound with the one from Example 9.2.

Exercise 9.5. *Try to find the smallest k for which, using Chebyshev's inequality, we obtain*

$$\Pr[|X - \frac{n}{4}| \geq k] \leq \frac{1}{1000},$$

where, as before, $X \sim Bin(n, \frac{1}{4})$.

It is not that hard to come up with examples to show that both Markov's inequality and Chebyshev's inequality are tight (try to do it!). In particular, it means that they are best possible in general. This leads to the question of whether we can do better in a more restricted, yet interesting, class of random variables. This question brings us to consider the case of a random variable that is the sum of a number of independent random variables. Such random variables play a central role in probability, randomized algorithms and other areas.

So, can Markov's and Chebyshev's Inequalities be improved for this particular kind of random variable?

First, let us check what Chebyshev's Inequality (the stronger of the two) gives us for a sum of independent random variables.

Theorem 9.6. *Let $X_1, X_2, \ldots, X_n$ be independent random variables with $\mathbb{E}[X_i] = \mu_i$ and $\mathrm{Var}(X_i) = \sigma_i^2$. Then, for any $a > 0$:*

$$\Pr[|\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mu_i| \geq a] \leq \frac{\sum_{i=1}^{n} \sigma_i^2}{a^2}$$

Proof. This follows from Chebyshev's Inequality applied to $\sum_{i=1}^{n} X_i$ and the fact that $\mathrm{Var}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} \mathrm{Var}(X_i)$ for independent variables. □

Note that in Example 9.4 we actually used this bound without saying, since a binomial random variable is a sum of indicators.

Since often we will use the above inequality for i.i.d (independent and identically distributed) random variables, let us state the following immediate coroallary for this scenario:

Suppose that $X_1, \ldots, X_n$ are i.i.d random variables with expectation μ and variance σ^2, then

$$\Pr\left[\left|\frac{\sum_{i=1}^{n} X_i}{n} - \mu\right| \geq t\right] \leq \frac{\sigma^2}{nt^2}$$

for any $t > 0$. Please pause reading and try to prove this inequality to make sure that we are all on the same page.

You can now wonder whether this result be improved or is it tight? Note that we used the independence of the variables $\{X_i\}$ to get $\mathrm{Var}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} \mathrm{Var}(X_i)$. However, we *only used pairwise independence* of the variables $\{X_i\}$. Indeed, it is possible to show that Theorem 9.6 is tight when all the variables $\{X_i\}$ are just guaranteed to be pairwise independent.

We are now ready to tackle the case of a sum of independent random variables, and present the strongest general concentration inequality known for this setting.

9.3 Chernoff's inequality

There are many different forms of Chernoff's bounds, each tuned to slightly different assumptions. We will start with the statement of the bound for the simple case of a sum of independent Bernoulli random variables.

Theorem 9.7 (Chernoff Bounds)**.** *Let $X = \sum_{i=1}^{n} X_i$, where $X_i = 1$ with probability p_i and $X_i = 0$ with probability $1 - p_i$, and all X_i are independent. Let $\mu = \mathbb{E}[X] = \sum_{i=1}^{n} p_i$. Then*

(i) **Upper Tail:** $\Pr[X \geq (1+\delta)\mu] \leq e^{-\frac{\delta^2}{2+\delta}\mu}$ *for all $\delta > 0$;*

(ii) **Lower Tail:** $\Pr[X \leq (1-\delta)\mu] \leq e^{-\mu\delta^2/2}$ *for all* $0 < \delta < 1$*;*

Note that, curiously, the lower and upper tail don't look exactly the same.

For $\delta \in (0,1)$, we can combine the lower and upper tails in Theorem 9.7 to obtain the following simple and useful bound:

Corollary 9.8. *Let* $X = \sum_{i=1}^{n} X_i$*, where* $X_i = 1$ *with probability* p_i *and* $X_i = 0$ *with probability* $1 - p_i$*, and all* X_i *are independent. Let* $\mu = \mathbb{E}[X] = \sum_{i=1}^{n} p_i$*. Then,*

$$\Pr[|X - \mu| \geq \delta\mu] \leq 2e^{-\mu\delta^2/3} \quad \text{for all } 0 < \delta < 1.$$

Before proving Theorem 9.7 we will give some applications that you might find helpful in your projects.

Example 9.9. *Let* $X \sim Bin(n, \frac{1}{4})$*. Then, in particular we can write* $X = \sum_{i=1}^{n} X_i$*, where the* X_i*s are i.i.d Bernoulli random variables with* $\Pr[X_1 = 1] = \frac{1}{4}$ *and* $\Pr[X_1 = 0] = \frac{3}{4}$*. Therefore, for evry* $\varepsilon > 0$*, by chernoff's inequality we obtain that*

$$\Pr[|X - \frac{n}{4}| \geq \varepsilon n] \leq 2e^{-\frac{n\varepsilon^2}{12}}.$$

Compare this bound to the one obtained in Example 9.4.

Example 9.10. *Let* $G_{n,p}$ *be a graph on* n *the vertex set* $\{1, \ldots, n\}$*, for which every pair* $\{i, j\}$ *is being added to* $E(G_{n,p})$ *with probability* p*, independently. Show that for every* $\varepsilon > 0$ *and* $p = \omega\left(\frac{\log n}{n}\right)$*, with probability* $1 - o(1)$ *we have that for every vertex* $v \in [n]$ *we have* $(1-\varepsilon)(n-1)p \leq d(v) \leq (1+\varepsilon)(n-1)p$*. Check what would be the smallest* p *(if any) for which you can prove such a statement using Chebyshev's.*

Now we can prove Chernoff's bounds.

Proof. First, let us describe the general strategy: Let X be any random variable, and $a \in \mathbb{R}$. We will make use of the same idea which we used to prove Chebyshev's inequality from Markov's inequality, where instead of squaring the expression we will exponentiate it. Specifically, for any $s > 0$,

$$\begin{aligned} \Pr[X \geq a] &= \Pr[e^{sX} \geq e^{sa}] \\ &\leq \frac{\mathbb{E}[e^{sX}]}{e^{sa}} \end{aligned} \tag{9}$$

So we have some upper bound on $\Pr[X \geq a]$ in terms of $\mathbb{E}[e^{sX}]$. Similarly, for any $s > 0$, we have

$$\begin{aligned} \Pr[X \leq a] &= \Pr[e^{-sX} \geq e^{-sa}] \\ &\leq \frac{\mathbb{E}[e^{-sX}]}{e^{-sa}} \end{aligned}$$

Next, since $X = \sum X_i$ and the X_is are independent, we know that the Moment generating function of X can be written as a product of the moment generating functions of the X_is (WHY?). Therefore, to prove the theorem it will be enough to bound the moment generating function of each X_i individually.

Note that since the generating function of $X_i \sim Ber(p_i)$ is $1 + p_i(e^s - 1)$, and since $1 + x \leq e^x$ holds for all $x \in \mathbb{R}$, we obtain that

$$M_{X_i}(s) = 1 + p_i(e^s - 1) \leq e^{p_i(e^s - 1)}.$$

Therefore, we obtain that

$$M_X(s) = \prod_{i=1}^{n} M_{X_i}(s) \leq \prod_{i=1}^{n} e^{p_i(e^s-1)} = e^{(e^s-1)\sum_{i=1}^{n} p_i} \leq e^{(e^s-1)\mu}, \tag{10}$$

where we used the fact that $\sum_{i=1}^{n} p_i = \mathbb{E}[X] = \mu$.

For the proof of the upper tail, we can now apply the strategy described in Equation 9, with $a = (1+\delta)\mu$ and choosing $s = \ln(1+\delta)$ (where δ is given in the statement of the theorem):

$$\begin{aligned}\Pr[X \geq (1+\delta)\mu] &\leq e^{-s(1+\delta)\mu} e^{(e^s-1)\mu)} \\ &= \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^\mu.\end{aligned}$$

(Our choice of s is motivated as follows: since we try to minimize the upper bound on the tail probability, we try to minimize our expression for the upper bound as a function of s. Taking the derivative of the exponent shows that this minimum is achieved exactly at $s = \log(1+\delta)$. Prove it!)

Taking the natural logarithm of the right-hand side yields

$$\mu(\delta - (1+\delta)\ln(1+\delta)).$$

Using the following inequality for $x > 0$ (left as an exercise):

$$\ln(1+x) \geq \frac{x}{1+x/2},$$

we obtain

$$\mu(\delta - (1+\delta)\ln(1+\delta)) \leq -\frac{\delta^2}{2+\delta}\mu.$$

Hence, we have the desired bound for the upper tail:

$$\Pr(X \geq (1+\delta)\mu) \leq \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^\mu \leq e^{-\frac{\delta^2}{2+\delta}\mu}.$$

The proof of the lower tail is entirely analogous. It proceeds by taking $s = \ln(1-\delta)$ and applies the following inequality for the logarithm of $(1-\delta)$ in the range $0 < \delta < 1$:

$$\ln(1-\delta) \geq -\delta + \frac{\delta^2}{2}.$$

Details are left as an exercise. □

10 Central Limit Theorem

The Central Limit Theorem (CLT for short), is one of the most important results in probability theory, with tons of applications. Roughly speaking, it asserts that the sum of "many" independent random variables is approximately distributed like a normal distribution (with the corresponding mean and variance). Here we will state and prove the following simple form of CLT:

Theorem 10.1 (CLT)**.** *Let $X_1, X_2, \ldots$ be a sequence of i.i.d random variables, each of which has mean μ and variance σ^2. Then, as n tends to infinity, we have that*

$$\Pr\left[\frac{X_1 + \ldots + X_n - n\mu}{\sigma \cdot \sqrt{n}} \leq a\right] \to \phi(a),$$

where $\phi(a) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-x^2/2} dx$, is the distribution function of a standard gaussian.

11 Derandomization

The probabilistic method supplies, in many cases, effective randomized algorithms for various algorithmic problems. In some cases, these algorithms can be derandomized and converted into deterministic ones. In this chapter we discuss some examples.

In this section we will show some examples of how to convert probabilistic algorithms into deterministic ones (that is, to *derandomize* the algorithms). To explain the basic idea, let us start with a simple example. The starting point is the following, relatively trivial, theorem:

Proposition 11.1. *There exists a 2-coloring of the edges of K_n (the complete graph on n vertices) such that the number of monochromatic copies of K_4 is at most $\binom{n}{4}\frac{1}{2^5}$.*

The proposition is trivial as the above quantity is just the expected number of monochromatic K_4s in case that we take a random coloring. This naturally leads to the question:

Can we actually find deterministically such a coloring in time which is polynomial in n?

Let us describe a procedure that does it, and is a special case of a general technique called the *method of conditional probabilities*.

We want to define, for every partial coloring of $E(K_n)$, a *weight function*. Given a coloring of some of the edges of K_n by red and blue, we define, for each copy K of K_4 in Kn, a weight $\omega(K)$ as follows: If at least one edge of K is colored red and at least one edge is colored blue, then $\omega(K) = 0$. If no edge of K is colored, then $\omega(K) = 2^5$, and if $r \geq 1$ edges of K are colored, all with the same color, then $\omega(K) = 2^{r-6}$. Also define the total weight W of the partially colored K_n as the sum

$$W = \sum_K \omega(K).$$

Observe that the weight of each copy of K_4 is precisely the probability that it will be monochromatic, if all the presently uncolored edges of K_n will be assigned randomly and independently.

Hence, by Linearity of Expectation, the total weight W is simply the expected number of monochromatic copies of K_4 in such a random extension of the partial coloring of K_n to a full coloring.

We can now describe the procedure for finding a coloring as in the Proposition.

Let $e_1, \ldots, e_N$, where $N = \binom{n}{2}$, be an arbitrary ordering of the edges of K_n. Construct the desired two-coloring by consecutively coloring each edge either red or blue, where in each turn i we color the edge e_i. Suppose $e_1, \ldots, e_{i-1}$ have already been colored, and we wish to color e_i. Let W be the weight of K_n, as defined above, with respect to the given partial coloring of $e_1, \ldots, e_{i-1}$. Similarly, let W_{red} be the weight of K_n with respect to the partial coloring obtained from the current coloring by assigning e_i red, and let W_{blue} be the weight of K_n with respect to the partial coloring obtained from the current coloring by assigning e_i blue. By the definition of W (and as follows from its interpretation as an expected value), we have

$$W = \frac{W_{red} + W_{blue}}{2}.$$

The color of e_i is now chosen to minimize the resulting weight. That is, if $W_{red} \leq W_{blue}$, then we color e_i red, otherwise we color it blue. By the above inequality, the weight function never increases during the algorithm. Since at the beginning its value is exactly $\binom{n}{4} \cdot 2^{-5}$, its value at the end is at most this quantity. However, at the end all edges are colored, and the weight is precisely the number of monochromatic copies of K_4. Thus the procedure above produces, deterministically and in polynomial time, a 2-edge-coloring of K_n, satisfying the conclusion of the Proposition. This completes the proof.

The second, and more interesting, example that we want to consider is related to the area of *positional games*. In a positional game there are two players, I and II, alternating turns in claiming previously unclaimed elements of some *board* V, with I going first. There is also a predetermined family of subsets

of V, denoted by $\mathcal{F}$ which is considered as the collection of *winning sets* (the pair $(V, \mathcal{F})$ is considered as the *hypergraph* of the game). The winner of the game is the **first** player to fully occupy all the elements of one of the winning sets in $\mathcal{F}$. If there is no winner by the time that all the elements in V have been previously claimed, the game is declared as a draw.

A complicated definition? not so... let us illustrate the definition by a simple example, namely the traditional child-game Tic-Tac-Toe. This game is played on a 3×3 board, where the players alternate turns in marking $0/X$ in previously unmarked spots. The winner is the first player to fully mark some *combinatorial line* (that is, horizontal, vertical, or diagonal line). This game can be easily described as a positional game! (do it).

A pioneering result in combinatorics is the following theorem by Erdős and Selfridge from 1973. The impact of their result completely changed the subject and initiated the so-called "derandomization" technique in this area.

Theorem 11.2 (Erdős-Selfridge)**.** *Let $\mathcal{F}$ be an n-uniform hypergraph, and assume that*

$$|\mathcal{F}| + MaxDed(\mathcal{F}) < 2^n.$$

Then, playing on $\mathcal{F}$ the second player can force a strong draw.

Remark 11.3. *Beck has managed to extend this theorem to the* biased *case. That is, when player I is allowed to take a elements per turn and Player II claims b elements per turn.*

Remark 11.4. *Note that if the second player can force a draw then so does the first.*

Proof. Let $\mathcal{F} = \{A_1, \ldots, A_M\}$. Assume that at some round of the game the first player has already occupied the elements $x_1, \ldots, x_i$ and the second player has occupied $y_1, \ldots, y_{i-1}$. Note that by choosing an element y_i, the second player "turns off" all the hyperedges containing y_i and refer to those sets as "dead sets". The winning sets which are not "dead" are referred to as "survivors", and they all have a chance to be occupied by the first player. We wish to define a potential function measuring the "danger" of each "survivor" and to measure the "danger" of the whole position. To do so we define

$$D_i = \sum_{s \in S_i} 2^{-u_s}$$

where u_s is the number of unoccupied elements of s and S_i is the set of all survivors at the current round.

A natural choice for second will thus be an element y_i which minimizes the danger function D_{i+1}. How to do so? suppose y_i and x_{i+1} are the next two steps of the players. Let us measure the effect on the danger function. Clearly,

$$D_{i+1} \leq D_i - \sum_{s \in S_i} 2^{-u_s} 1_{y_i \in s} + \sum_{s \in S_i} 2^{-u_s} 1_{x_{i+1} \in s}.$$

All in all, a natural choice for y_i is the unoccupied z for which $\sum_{s \in S_i} 2^{-u_s}$ attains its maximum. Then, we obtain

$$D_{i+1} \leq D_i.$$

Therefore, the second player can force

$$D_1 \geq D_2 \geq \ldots \geq D_{last}.$$

Note that if first wins then there exists an $s \in S_{last}$ with $u_s = 0$ and therefore $D_{last} \geq 1$. On the other hand, by assumption we have

$$D_1 = \sum_{s \in S_1} 2^{-n+1} 1_{x_1 \in s} + \sum_{s \in S_i} 2^{-n} 1_{x_1 \notin s} \leq (|\mathcal{F}| + MaxDeg(\mathcal{F}))2^{-n} < 1.$$

This completes the proof. □

Remark 11.5. *Note that if the two players play completely at random, then the expected number of monochromatic edges is*

$$2^{-n+1}|\mathcal{F}|.$$

Now, if the expectation if smaller than 1 *then there exists a drawing position (it doesn't mean that either of the players can actually force such a position!). This explains the beauty of the argument – the proof is a "derandomization" of the above argument. It "upgrades" the existing drawing position to a Drawing strategy! Moreover, note that this condition is tight (a binary tree with* n *levels has* 2^{n-1} *edges and first player clearly can occupy a full branch).*

Remark 11.6. *A non-uniform version can be stated as follows: If*

$$\sum_{A\in\mathcal{F}} 2^{-|A|} < 1/2$$

then the second player can force a strong draw. (Exercise!)

12 Practice problems for the Final

1. Let $X, Y \sim U[0,1]$, be independent and let $Z = \max\{X, Y\}$.

   (a) (10 points) Calculate $\Pr[Z \leq a]$.

   (b) (10 points) Calculate the density function of Z.

   (c) (5 points) Calculate $Var(Z)$.

2. Let X be uniformly distributed on the interval $[0, \frac{\pi}{2}]$ and let Y be exponentially distributed with mean $1/3$. Assume X and Y are independent.

   (a) Compute the joint density function of U and V, where U and V are given by

   $$U = e^Y \cos(X), \quad V = e^Y \sin(X).$$

   (b) Compute $E[U+V]$ and $[U+V]$.

3. A factory produces bolts of a certain width. The bolts don't come out perfectly each time – their widths deviate from the specified width by a random amount X, measured in micrometers. Suppose X has density function $f(x) = \frac{C}{1+(x/10)^2}$, where $-20 \leq x \leq 20$.

   (a) Find the value of C such that f is indeed a probability density function.

   (b) Calculate the mean and variance of X.

   (c) Suppose the factory produces 10,000 bolts a day. Assuming that each bolt is produced independently, estimate the number of bolts whose widths deviate from specification by more than 10 micrometers.

4. Video projector light bulbs are known to have a mean lifetime of $\mu = 100$ hours and standard deviation $\sigma = 75$ hours. The university uses the projectors for 9000 hours per semester.

   (a) (15 points) Use the central limit theorem to estimate the probability that 100 light bulbs will last the whole semester?

   (b) (10 points) Explain how to estimate the number of light bulbs necessary to have a 1% chance of running out of light bulbs before the semester ends. Don't actually do the whole computation.

5. You and one million other people have bought tickets for this week's lottery. Each person has a one in one million chance of selecting winning numbers. If there is more than one person with winning numbers, one winner is randomly chosen from them to win the prize.

   Suppose your ticket has winning numbers. Let X be the number of other matching tickets belonging to the other one million players. Since there are one million other tickets, each of which has a one in one million chance of winning, we can assume that X is approximately a Poisson random variable with parameter 1.

   (a) What is the probability that you have the only winning ticket?

   (b) The prize winner is chosen at random from all the winning ticket holders. What is the probability that you win, given that there are x other winners?

   (c) What is the probability that you win, given that you have a matching ticket? *Hint: conditional expectation.*

6. Suppose that X is random variable with a moment generating function $M_X(t) = \left(\frac{1+e^{100t}}{2}\right)^n$.

   (a) (10 points) Find $\mathbb{E}[X]$.

   (b) (7 points) Find $\mathbb{E}[X^2]$ and then compute $Var(X)$.

   (c) (5 points) What is the probability that $X \geq 10\mathbb{E}[X]$? Use Chebyshev's inequality to upper bound this quantity.

   (d) (3 points) Do you see how to obtain a better bound than the one you obtained with Chebyshev's? if yes, then explain in words without computing.

7. Suppose that X is uniformly distributed on the set of n even numbers $\{2, 4, 6, 8, \ldots, 2n\}$.

   (a) Calculate $M_X(t)$.

   (b) Calculate $\mathbb{E}[X^3]$.

   (c) Let $X_1, \ldots, X_k$ be independent copies of X and let $S_k = X_1 + \ldots + X_k$. Calculate $M_{S_k}(t)$.

8. Let X be any set, and let $\mathcal{F}$ be a collection of subsets of X, each of size exactly n (you can assume that n is sufficiently large if needed).

   (a) (15 points) Prove that if $|\mathcal{F}| \leq 2^{n-1} - 1$ then there exists a partition $X = X_1 \cup X_2$ such that for all $F \in \mathcal{F}$ we have that $F \cap X_1 \neq \emptyset$ and $F \cap X_2 \neq \emptyset$.

   (b) (10 points) Prove, using Chernoff's bounds, that if $|\mathcal{F}| \leq n^{100}$, then there exists a partition $X = X_1 \cup X_2$ such that for all $F \in \mathcal{F}$ we have $\left||F \cap X_1| - \frac{n}{2}\right| \leq C\sqrt{n \log n}$, for some fixed constant $C > 0$ that doesn't depend on n.

9. You and your friend are playing a game. You start by selecting a number X uniformly at random from $[0, 1]$. Your friend picks numbers $Y_1, Y_2, \ldots$ uniformly in $[0, 1]$ until they pick a number larger than $X/2$.

   (a) (10 points) Find the expected number of times, N, your friend needs to pick a number. *Hint: Condition on $X = x$.*

   (b) (10 points) Find the expected sum of your friend's numbers given that they had to pick N numbers.

   (c) (5 points) Find the expected sum of your friend's numbers.

10. (a) Let X be a random variable. Use Chebyshev's inequality to show that

$$\Pr[X = 0] \leq \frac{[X]}{E[X]^2}.$$

Deduce that if X_n is a sequence of random variables and $[X_n] = o(E[X_n]^2)$, then $X_n = 0$ with high probability.

(b) Show that if $p \geq Cn^{-2/3}$ for some large constant C, then the random graph $G(n,p)$ contains a clique of size 4 with high probability.

13 Asymptotics estimates and 'big O' notation

In most of the undergrad courses we deal with *exact* formulas. Among many more examples, in calculus we learnt Taylor's expansion and saw that $e^x = \sum_{k=1}^{\infty} \frac{x^k}{k!}$, $\cos x = \sum_{k=0}^{\infty} (-1)^k x^{2k} 2k!$, etc. In Math 130A we saw many identities of the form $2^n = \sum_{k=0}^{n} \binom{n}{k}$, $\binom{n+1}{m+1} = \sum_{k=m}^{n} \binom{k}{m}$, and more.

In the real world, however, in most scenarios it will be very hard and even impossible to obtain a closed formula. There are many natural and simple-to-state problems in combinatorics, probability, number theory and more, which an exact formula is still unknown, or even worse – it is known that an exact formula does not even exist.

What we do in such cases is to derive "asymptotic estimates" for these quantities. That is, we derive approximations of the quantities by "simple" functions, which show how these quantities behave "asymptotically". For example, in Section 1.2 we showed that the number of triangles in every graph on n vertices with at least $(1/4 + \epsilon)n^2$ edges is "of order" cn^3, without finding the exact constant c (Which clearly is not unique for all graphs and depends on the graph structure).

Let us start by introducing a convenient (and nowadays standard) notation.

Definition 13.1. *Given two functions $f, g : \mathbb{R} \to \mathbb{R}$, we say that*

1. *$f = O(g)$ if and only if there exist constants $C_1 > 0$ and $C_2 > 0$ such that*

$$\text{for all } x \geq C_1 \text{ we have } f(x) \leq C_2 g(x).$$

2. *$f = \Omega(g)$ if and only if $g = O(f)$. Or equivalently, $f = \Omega(g)$ if and only if there exist constants $C_1 > 0$ and $C_2 > 0$ such that*

$$\text{for all } x \geq C_1 \text{ we have } f(x) \geq C_2 g(x).$$

3. *$f = \Theta(g)$ if and only if both $f = O(g)$ and $f = \Omega(g)$.*

4. *$f = o(g)$ if and only if $\lim_{x\to\infty} \frac{f}{g} = 0$.*

5. *$f = \omega(g)$ if and only if $g = o(f)$.*

As simple exercises try to write the full definition of items $3 - 5$.

Let us try to get a feeling for this notation that we've just introduces. For example, try to convince yourselves formally that the following hold:

1. $x^2 + 150x + 2^{100} = O(x^2)$,

2. $x^{1000} = o(2^x)$,

3. $x^3 + 10^{10}x^2 + 100 = \Omega(10000x^3)$,

4. $\frac{1}{x} = o(1)$,
5. $(1+\frac{1}{x})^x = \Theta(1)$,
6. $(1+\frac{1}{x})^{x^2} = \omega(1)$.

We now describe (Without proofs) some asymptotic estimates which will be very useful in our course.

Stirling's approximation Perhaps the most famous estimate is Stirling's approximation for factorial. It gives:

$$n! = \sqrt{2\pi n} \cdot \left(\frac{n}{e}\right)^n \cdot \left(1 + O\left(\frac{1}{n}\right)\right).$$

Harmonic numbers The nth *harmonic number* H_n is defined as follows

$$H_n = \sum_{k=1}^{n} \frac{1}{k}.$$

It is known that

$$H_n = \ln n + \gamma + o(1),$$

where $\gamma = 0.5772\ldots$ is the Euler constant.

Binomial coefficients One of the most useful inequalities in discrete mathematics is the following:

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

The proof of the lower bound is more or less trivial (prove it!). For the upper bound, we will prove it as follows:

Recall from the binomial formula that

$$(1+x)^n = \sum_{i=0}^{n} \binom{n}{i} x^i.$$

Therefore, if we choose $x \geq 0$ we obtain that

$$(1+x)^n \geq \sum_{i=0}^{k} \binom{n}{i} x^i,$$

which implies that

$$\frac{(1+x)^n}{x^k} \geq \sum_{i=0}^{k} \binom{n}{i} x^{i-k}.$$

Now, suppose that $0 < x < 1$ and observe that for all $i \leq k$ we have $x^{i-k} \geq 1$. This implies that

$$\frac{(1+x)^n}{x^k} \geq \sum_{i=0}^{k} \binom{n}{i}.$$

Finally, choose $x = \frac{k}{n}$ (and observe that $0 < x < 1$ and therefore this is a legal choice), and obtain that

$$\binom{n}{k} \leq \sum_{i=0}^{k} \binom{n}{i} \leq \frac{(1+x)^n}{x^k} = \frac{(1+\frac{k}{n})^n}{\left(\frac{k}{n}\right)^k} \leq \left(\frac{en}{k}\right)^k.$$

This completes the proof of the upper bound.

General estimates We will also make use of the following estimates:

1. $e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$ for all x,
2. $\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \ldots$, for all $-1 < x < 1$,
3. $1 + x \leq e^x$ for all x,
4. $1 - x \geq e^{-x-x^2/2}$ for all $0 < x < 1$.

The following exercises are highly recommended:

Exercises 13.2. *1. Let f be a function satisfying $\lim_{x\to\infty} f(x) = \infty$. Show that $\ln \Theta(f) = \Theta(\ln f)$.*

2. Show that if $f = O(g)$ then we also have $f + g = O(g)$.

3. Show that $\sum_{k=1}^{\infty} \frac{k \log k}{2^k} = O(1)$.

4. Arrange the following functions according to the relation $f < g$ if and only if $f = o(g)$ (this is not a total ordering): $\sqrt{x}, x!, 2^x, x^x, x^2, \log x, (\log\log x)^{\log\log x}, 100$.

5. Show that $\binom{n}{\alpha n} = \frac{1+o(1)}{\sqrt{2\pi\alpha(1-\alpha)n}} \cdot 2^{H(\alpha)n}$, where $H(\alpha) = -\alpha \log_2 \alpha - (1-\alpha)\log_2(1-\alpha)$ is the binary entropy *function.*

References

[1] Gravner, Janko. Online lecture notes, https://www.math.ucdavis.edu/ gravner/MAT135A/resources/lecturen

[2] Mitzenmacher, Michael, and Eli Upfal. Probability and computing: Randomization and probabilistic techniques in algorithms and data analysis. Cambridge university press, 2017.

[3] Ross, Sheldon M. A first course in probability. Vol. 7. Upper Saddle River, NJ: Pearson Prentice Hall, 2006.