COS 522: Assumed Knowledge

Boaz Barak

This document includes some mathematical notions I will assume. Most of these are taught in a discrete math class, and in any case are not complicated and easy to pick up. You can read it at your own pace — it will take a few weeks until we'll get to use all of this stuff. You do not have to do the exercises, and in any case don't need to submit them. If you feel you'd like more clarifications or pointers to reading material on any of these, please do not hesitate to contact me.

1 Some recommended reading

Some sources for this material are:

- Lecture notes for MIT course 6.042 "Mathematics for Computer Science" by Lehman and Leighton on `http://theory.lcs.mit.edu/~e_lehman/mathcs.pdf` (See also `http://theory.lcs.mit.edu/classes/6.042/spring05/`)
- 'Probability and Computing" by Upfal and Mitzenmacher.
- "The Probabilistic Method" by Alon and Spencer.

2 Mathematical Proofs

You will be expected to be able to read and write mathematical proofs, and be familiar with notions such as proof by contradiction and proof by induction.

3 Operations on Sets

I assume familiarity with basic notions of sets and operations on sets such as union (denoted $\cup$), intersection (denoted $\cap$), and set substraction (denoted $\setminus$). We denote by $|A|$ the size of the set A. I also assume familiarity with functions, and notions such one-to-one (injective) functions and onto (surjective) functions. If f is a function from a set A to a set B, we denote this by $f : A \to B$. If f is one-to-one then this implies that $|A| \leq |B|$. If f is onto then $|A| \geq |B|$. If f is a permutation/bijection (e.g., one-to-one *and* onto) then this implies that $|A| = |B|$.

If $f : A \to \{0,1\}$ is a function with binary (one bit) output, then it can also be considered as a subset of A (i.e., the subset $\{x \in A \mid f(x) = 1\}$). We'll interchange freely these two descriptions.

If A and B are finite sets then there are at most $|B|^{|A|}$ possible functions from A to B. Each such function can be described by its table of values (often called "truth table") which will contain $|A|$ elements from $|B|$, and so can be encoded using $|A| \log |B|$ bits. In particular a function from $\{0,1\}^n$ to $\{0,1\}$ can be described using 2^n bits, and the number of subsets of a set A (i.e., the number of functions from A to $\{0,1\}$) is $2^{|A|}$. The number of subsets of A of

A *relation* R on a set A is a subset of orderer pairs of A (i.e., $R \subseteq A \times A$). If $f : A \to A$ is a function, we can also view it as the relation consisting of all pairs of the form $\langle x, f(x) \rangle$.

4 Some notations we'll use:

- $\mathbb{N}$ set of natural numbers $\{0, 1, 2, \ldots\}$, $\mathbb{Z}$ set of whole numbers $\{0, \pm 1, \pm 2, \ldots\}$, $\mathbb{R}$ set of real numbers, $\mathbb{C}$ set of complex numbers. We denote by X^+ the intersection of the set X with the positive real numbers (i.e., the interval $(0, \infty)$).
- If X is a set then $x \leftarrow_{\mathrm{R}} X$ denotes a random variable uniformly distributed on the set X.
- $\forall$ for "for every" and $\exists$ for "there exists".
- Logical operations: AND ($\wedge$), OR ($\vee$), NOT ($\neg$), and XOR ($\oplus$). Simple equivalences such as De-Morgan laws.[1] Note that every Boolean function on n variables can be expressed as a formula consisting of $\neg$ and $\wedge$ only. (Can you give an upper bound on the size of this formula?)
- Limits: suppose that $f(\cdot)$ is a function from $\mathbb{N}$ to $\mathbb{R}$ (denoted $f : \mathbb{N} \to \mathbb{R}$). Let $c \in \mathbb{R}$. We say that $\lim_{n\to\infty} f(n) = c$ if for every $\epsilon > 0$ for every sufficiently large n, $|f(n) - c| < \epsilon$. In logical notation this condition is expressed as follows:

$$\forall \epsilon > 0 \; \exists N \in \mathbb{N} \text{ s.t. } \forall n > N \;\; |f(n) - c| < \epsilon$$

4.1 O notations

Let $f, g : \mathbb{N} \to \mathbb{R}^+$ be two functions. We say that:

- $f = O(g)$ there exists a constant $C > 0$ such that for every n, $f(n) \leq C \cdot g(n)$. Note that the definition would be the same if we said "for every sufficiently large n".
- $f = \Omega(g)$ if $g = O(f)$.
- $f = \Theta(g)$ if $f = O(g)$ and $g = O(f)$.
- $f = o(g)$ if $\lim_{n\to\infty} f(n)/g(n) = 0$.
- $f = \omega(g)$ if $g = o(f)$.

We say that $f : \mathbb{N} \to \mathbb{R}^+$ is *polynomially bounded* if $f(n) = n^{O(1)}$ (i.e., there exists a constant C such that $f(n) \leq n^C$ for every sufficiently n). We say that $f, g : \mathbb{N} \to \mathbb{R}$ are *polynomially related* if $f(n) = g(n)^{O(1)}$ and $f(n) = g(n)^{\Omega(1)}$. We say that f is *super-polynomial* if $f(n) = n^{\omega(1)}$.

Exercise 1. Which of these functions is super-polynomial?

- $n \mapsto n^2$
- $n \mapsto 2^{\log^2 n}$. ($\log^c n$ denotes $(\log n)^c$)
- $n \mapsto \frac{2^{\sqrt{n}}}{n^3}$
- $n \mapsto 2^{\log^{1/2} n}$
- $n \mapsto n^{\log n}$
- $n \mapsto n \log n$.

[1] $\neg(x \wedge y)$ is equivalent to $\neg x \vee \neg y$. Since $\forall$ is the same as many ANDs and $\exists$ is the same as many ORs, another formulation is that $\neg \forall x \phi(x)$ is equivalent to $\exists x \neg \phi(x)$.

5 Graphs

An *undirected graph* over a set V is a collection E of unordered pairs of V. Each element $v \in V$ is called a *vertex* while each element $e = \{u, v\} \in E$ is called an *edge*. If a graph has n vertices we'll often assume that $V = [n] = \{1, \ldots, n\}$. A *directed* graph over a set V is a collection E of *ordered* pairs. I assume you are familiar with basic notions such as paths in graphs and connected components.

6 Discrete Probability

I assume familiarity with basic probability theory on finite sample spaces, as described below.

6.1 Sample Spaces

For every probabilistic experiment (for example, tossing a coin or throwing 3 dice) the set of all possible results of the experiment is called a *sample space*. For example, if the experiment is to toss a coin and see if the result is "heads" or "tails" then the sample space is the set $\{H, T\}$, or equivalently (if we denote heads by 1 and tails by 0) the set $\{0, 1\}$. With every element x of the sample space we associate the probability p_x that the result of the experiment will be x. The number p_x is between 0 and 1 and the sum of all the p_x's is equal to 1. We sometimes denote the sample space by Ω, but many times it will be clear from the context. **Hint:** Whenever a statement about probability is made, it is a good habit to ask yourself what is the sample space that this statement refers.

As another example, consider the experiment of tossing three coins. In this case there are 8 possible results and hence the sample space is $\{000, 001, 010, 011, 100, 101, 110, 111\}$. Each element in the sample space gets chosen with probability $\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{2^3} = \frac{1}{8}$. An equivalent way to state the experiment of tossing n coins is to say that we choose a random n-long binary string. We call this the *uniform* distribution over $\{0, 1\}^n$ (because every string gets chosen with the same probability). If we want to say that we let x record the result of this experiment then we use the notation $x \leftarrow_{\mathrm{R}} \{0, 1\}^n$.

6.2 Events

An *event* is a subset of the sample space. The probability that an event happens is the probability that the result of the experiment will fall inside that subset. For example, if we consider the sample space of tossing 101 coins, then we can denote by E the event that most of the coins came up tails — at most 50 of the coins come up "heads". In other words, E is the set of all length-101 strings with at most 50 ones. We denote the probability that an event E occurs by $\Pr[E]$. For example, in this case we can write

$$\Pr_{x \leftarrow_{\mathrm{R}} \{0,1\}^{101}} [\# \text{ of 1's in } x \leq 50] = \frac{1}{2}$$

Exercise 2 (15 points)**.** Prove that indeed this probability is equal to $\frac{1}{2}$.[2]

[2] Hint: This will follow if you'll show that $|E| = |\{0, 1\}^{101} \setminus E|$. You can do this by exhibiting a bijection between these two sets.

6.3 Union Bound

If E and E' are events over the same sample space then another way to look at the probability that *either* E or E' occurs is to say that this is the probability that the event $E \cup E'$ (the union of E and E') occurs. A very simple but useful bound is that this probability is *at most* the sum of the probability of E and the probability of E'. This is called the *union bound.*

Theorem 1 (Union bound)**.** *Ω be some sample space and let $E, E' \subseteq \Omega$ be two events over Ω. Then,*

$$\Pr_{\Omega}[E \cup E'] \le \Pr_{\Omega}[E] + \Pr_{\Omega}[E']$$

Exercise 3 (15 points)**.** Prove Theorem 1.

Note that that there are examples of E and E' such that $\Pr[E \cup E']$ is strictly less than $\Pr[E] + \Pr[E']$. For example, this can be the case if E and E' are the same set (and hence $E \cup E' = E$). If E and E' are *disjoint* (i.e., mutually exclusive) then $\Pr[E \cup E'] = \Pr[E] + \Pr[E']$.

6.4 Random Variables

A random variable is a function that maps elements of the sample space to another set (often, but not always, to the set $\mathbb{R}$ of real numbers). For example, in the case of the uniform distribution over $\{0,1\}^{101}$, we can define the random variable N to denote the number of ones in the string chosen. That is, for every $x \in \{0,1\}^{101}$, $N(x)$ is equal to the number of ones in x. Thus, the event E we considered before can be phrased as the event that $N \le 50$ and the formula above can be phrased as

$$\Pr_{x \leftarrow_{\mathrm{R}} \{0,1\}^{101}}[N(x) \le 50] = \frac{1}{2}$$

For the remainder of this handout, we will only consider *real* random variables (that is random variables whose output is a *real number*).

6.5 Expectation

The *expectation* of a random variable is its weighted average. That is, it is the average value it takes, when the average is weighted according to the probability measure on the sample space. Formally, if N is a random variable on a sample space Ω (where for every $x \in \Omega$, the probability that x is obtained is given by p_x) then the expectation of N, denoted by $\mathbb{E}[N]$ is defined as follows:

$$\mathbb{E}[N] \overset{def}{=} \sum_{x \in \Omega} N(x) \cdot p_x$$

For example, if the experiment was to choose a random U.S. citizen (and hence the sample space is the set of all U.S. citizens) and we defined the random variable H to be the height of the person chosen, then the expectation of H (denoted by $\mathbb{E}[H]$) is simply the average height of a U.S. citizen.

There can be two different random variables with the same expectation. For example, consider the sample space $\{0,1\}^{101}$ with the uniform distribution, and the following two random variables:

- N is the random variable defined above: $N(x)$ is the number of ones in x.
- M is defined as follows: if x is the all ones string (that is $x = 1^{101}$) then $M(x) = 50.5 \cdot 2^{101}$. Otherwise (if $x \neq 1^{101}$) then $M(x) = 0$.

The expectation of N equals 50.5 (you'll prove this below in Exercise 4).

The expectation of M is also 50.5: with probability 2^{-101} it will be $2^{101} \cdot 50$ and with probability $1 - 2^{-101}$ it will be 0.

Note that even though the average of M is 50.5, the probability that for a random x, $M(x)$ will be close to 50.5 or even bigger than zero is very very small. This is similar to the fact that if Bill Gates is in a room with 99 poor people (e.g. theoretical computer scientists), then the average worth of a random person in this room is more than \$100M even though with probability 0.99 a random person in the room will be worth much less than that amount. Hence the name "expectation" is somewhat misleading.

In contrast, it will follow from Theorem 4, that for a random string x, even though it will never have $N(x)$ equal to exactly 50.5 (after all, $N(x)$ is always a whole number), with high probability $N(x)$ will be close to 50.5.

The fact that two different variables can have the same expectation means that if we know the expectation it does not give us *all* the information about the random variable but only *partial* information.

Linearity of expectation. The expectation has a very useful property which is that it is a *linear function.* That is, if N and M are random variables over the same sample space Ω, then we can define the random variable $N+M$ in the natural way: for every $x \in \Omega$, $(N+M)(x) = N(x)+M(x)$. It turns out that $\mathbb{E}[N + M] = \mathbb{E}[N] + \mathbb{E}[M]$. For every fixed number c and random variable N we define the random variable cN in the natural way: $(cN)(x) = c \cdot N(x)$ It turns out that $\mathbb{E}[cN] = c\mathbb{E}[N]$.

Exercise 4 (30 points)**.** 1. Prove that the expectation is linear.

2. Prove that the expectation of the random variable N defined above over the sample space $\{0, 1\}^{101}$ is equal to 50.5.

6.6 Markov Inequality

As we saw above, sometimes we want to know not just the expectation of a random variable but also the probability that the variable is close to (or at least not too far from) its expectation. Bounds on this probability are often called "tail bounds". The simplest one of them is *Markov* inequality, which is a one-sided inequality. It says that with high probability a non-negative random variable is never much larger than its expectation. (Note that the random variable M defined above was an example of a non-negative random variable that with high probability is much *smaller* than its expectation.) That is, it is the following theorem:

Theorem 2 (Markov Inequality)**.** *Let X be a random variable over a sample space Ω such that for all $x \in \Omega$, $X(x) \geq 0$. Let $k \geq 1$. Then,*

$$\Pr[X \geq k\mathbb{E}[X]] \leq \frac{1}{k}$$

Proof. Denote $\mu = \mathbb{E}[X]$. Suppose for the sake of contradiction that $\Pr[X \geq k\mu] > 1/k$. Let $S = \{x \in \Omega \mid X(x) \geq k\mu\}$ and $\overline{S} = \Omega \setminus S$. By the definition of expectation

$$\mathbb{E}[X] = \sum_{x \in \Omega} X(x)p_x = \sum_{x \in S} X(x)p_x + \sum_{x \in \overline{S}} X(x)p_x$$

However, we know that for each $x \in S$, $X(x) \geq k\mu$ and hence

$$\sum_{x \in S} X(x)p_x \geq \sum_{x \in S} k\mu p_x = k\mu \sum_{x \in S} p_x$$

Yet $\sum_{x \in S} p_x = \Pr[S] > \frac{1}{k}$ under our assumption and hence $\sum_{x \in S} X(x)p_x > k\mu\frac{1}{k} = \mu$.

Since $X(x) \geq 0$ for all $x \in \Omega$ we get that $\sum_{x \in \overline{S}} X(x)p_x \geq 0$ and hence $\mathbb{E}[x] > \mu$, yielding a contradiction. □

6.7 Variance and Chebychev inequality

We already noted that the distance from the expectation is an interesting parameter. Thus, for a random variable X with expectation μ we can define a new random variable $\tilde{X}$ which to be the distance of X from its expectation. That is, for every $x \in \Omega$, we define $\tilde{X}(x) = |X - \mu|$. (Recall that $|\cdot|$ denotes the absolute value.) It turns out that it is hard to work with $\tilde{X}$ and so we look at the variable $\tilde{X}^2$, which is equal to $(X - \mu)^2$. We define the *variance* of a random variable X to be equal to the expectation of $\tilde{X}^2$. That is, for X with $\mathbb{E}[X] = \mu$,

$$Var[X] \overset{def}{=} \mathbb{E}[\tilde{X}^2] = \mathbb{E}[(X - \mu)^2]$$

In other words $Var[X]$ is defined to be $\mathbb{E}[(X - \mathbb{E}[X])^2]$.

We define the *standard deviation* of X to be the square root of $Var[X]$.

Exercise 5 (30 points)**.**

1. Prove that $Var[X]$ is always non-negative.
2. Prove that $Var[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$.
3. Give an example for a random variable X such that $\mathbb{E}[X^2] \neq \mathbb{E}[X]^2$.
4. Give an example for a random variable X such that its standard deviation is *not equal* to $\mathbb{E}[|X - \mathbb{E}[X]|]$.
5. Give an example for two random variables X, Y over the same sample space such that $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.
6. Give an example for two random variables X, Y over the same sample space such that $\mathbb{E}[XY] \neq \mathbb{E}[X]\mathbb{E}[Y]$.

If we have a bound on the variance then we can have a better tail bound on the variables:

Theorem 3 (Chebyshev's inequality)**.** *Let X be a random variable over Ω with expectation μ and standard deviation σ. Let $k \geq 1$. Then,*

$$\Pr[|X - \mu| \geq k\sigma] \leq 1/k^2$$

Proof. The variable $Y = (X - \mu)^2$ is non-negative and has expectation $Var(X) = \sigma^2$. Therefore, by Markov inequality, $\Pr[Y \geq k^2\sigma^2] \leq 1/k^2$.

However, whenever $|X - \mu| \geq k\sigma$ it holds that $|X - \mu|^2$ (which is equal to Y) is at least $k^2\sigma^2$. This means that

$$\Pr[|X - \mu| \geq k\sigma] \leq \Pr[Y^2 \leq k^2\sigma^2] \leq 1/k^2$$

□

6.8 Conditional probabilities and independence

Let A be some event over a sample space Ω (with $\Pr[A] > 0$). By a probability *conditioned on* A we mean the probability of some event, assuming that we already know that A happened. For example if Ω is our usual sample space of uniform choices over $\{0,1\}^{101}$ and A is the event that the first coin turned out head, then the conditional space is the space of all length-101 strings whose first bit is 1.

Formally this is defined in the natural way: we consider A as a sample space by inheriting the probabilities from Ω (and normalizing so the probabilities will sum up to one). That is, for every $x \in A$ we define $p_{x|A}$ (the probability that x is chosen conditioned on A) to be $\frac{p_x}{\Pr[A]}$. For an event B we define $\Pr[B|A]$ (the probability that B happens conditioned on A) to be $\sum_{x \in A \cap B} p_{x|A} = \frac{\Pr[A \cap B]}{\Pr[A]}$.

Independent events. We say that B is independent from A if $\Pr[B|A] = \Pr[B]$. That is, knowing that A happened does not give us any new information on the probability that B will happen. By plugging the formula for $\Pr[B|A]$ we see that B is independent from A if and only if

$$\Pr[B \cap A] = \Pr[A]\Pr[B]$$

This means that B is independent from A iff A is independent from B and hence we simply say that A and B are independent events.

For example, if, as above, A is the event that the first coin toss is heads and B is the event that the second coin toss is heads then these are independent events. In contrast if C is the event that the number of heads is at most 50 then C and A are *not* independent (since knowing that A happened increases somewhat the chances for C).

If we have more than two events then it's a bit more messy: we say that the events $E_1, \ldots, E_n$ are mutually independent if not only $\Pr[E_1 \cap E_2 \cap \cdots \cap E_n] = \Pr[E_1] \cdots \Pr[E_n]$ but also this holds for every subset of $E_1, \ldots, E_n$. That is, for every subset I of the numbers $\{1, \ldots, n\}$,

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

Independent random variables. We say that U and V are *independent random variables* if for every possible values u and v, the events $U = u$ and $V = v$ are independent events or in other words $\Pr[U = u \text{ and } V = v] = \Pr[U = u]\Pr[V = v]$. We say that $U_1, \ldots, U_n$ are a collection of independent random variables if for all values $u_1, \ldots, u_n$, the events $U_1 = u_1, \ldots, U_n = u_n$ are mutually independent.

Exercise 6 (20 points).

1. Prove that if U and V are independent then $\mathbb{E}[UV] = \mathbb{E}[U]\mathbb{E}[V]$.
2. Prove that if U and V are independent then $Var[U + V] = Var[U] + Var[V]$.

6.9 The Chernoff Bound

Suppose that 60% of a country's citizens prefer the color blue over red. A poll is the process of choosing a random citizen and finding his or her favorite color. Suppose that we do this n times and we define the random variable X_i to be 0 if the color of the i^{th} person chosen is red and 1 if it is blue. Then, for each i the expectation $\mathbb{E}[X_i]$ is 0.6, and by linearity of expectation $\mathbb{E}[\sum_{i=1}^{n} X_i] = 0.6n$. The estimate we get out of this poll for the fraction of blue-preferrers is $\frac{\sum X_i}{n}$ and we would like to know how close this is to the real fraction of the population (i.e., 0.6). In other words, for any

$\epsilon > 0$, we would like to know what is the probability that our estimate will be ϵ off from the real value, i.e., that $|\frac{\sum X_i}{n} - 0.6| > \epsilon$.

It turns out that in this case we have a very good bound on the deviation of $\sum X_i$ from its expectation, and this is because all of the X_i's are independent random variables (since in each experiment we draw a new random person independently of the results of previous experiments). This is the Chernoff bound, which we state here in a simplified form:

Theorem 4 (Chernoff bound)**.** *Let* $X_1, \ldots, X_n$ *be independent random variables with* $0 \leq X_i \leq 1$ *and* $\mathbb{E}[X_i = \mu]$*. Then,*

$$\Pr\left[\left|\frac{\sum X_i}{n} - \mu\right| > \epsilon\right] < 2^{-\epsilon^2 n}$$

Proof. See Section 24.3 (page 320) in the Lehman and Leighton notes (`http://theory.lcs.mit.edu/~e_lehman/mathcs.pdf`). □

6.10 The Probabilistic Method

The probabilistic method relies on the following fact: if A is an event over a sample space that has non-zero probability then there exists a point in the sample space satisfying A. This trivial observation has many surprising and powerful implications. It is often combined with other simple facts such as:

- If X is a random variable and μ a number and $\mathbb{E}[X] = \mu$ then $\Pr[X \leq \mu] > 0$.
- If $B_1, \ldots, B_N$ are events with $\Pr[B_i] < 1/N$ for every i, then $\Pr[\neg B_1 \wedge \cdots \wedge \neg B_N] > 0$.

Exercise 7. If G is a graph then let $\omega(G)$ denote the size of the largest clique in G, and let $\alpha(G)$ denote the size of the largest independent set in G. Prove that there exists a graph G on n vertices with $\omega(G), \alpha(G) \leq 3 \log n$.