Handout 1: Mathematical Background

Boaz Barak

September 18, 2007

This is a brief review of some mathematical tools, especially probability theory that we will use. This material is mostly from discrete math (COS 340/341) but is also taught in many other courses.

Some good sources for this material are the lecture notes by Papadimitriou and Vazirani (see home page of Umesh Vaziarani), Lehman and Leighton (see home page of Eric Lehman, Chapters 18 to 24 are particularly relevant). The mathematical tool we use most often is discrete probability. The "Probabilistic Method" book by Alon and Spencer is a great resource in this area. Also, the books of Mitzenmacher and Upfal and Prabhakar and Raghavan cover probability from a more algorithmic perspective.

Although knowledge of algorithms is not strictly necessary, it would be quite useful. Good books are (1) Corment, Leiserson, Rivest and Smith, (2) Dasgupte, Papadimitriou and Vaziarni, (3) Kleinberg and Tardos. We do now require prior knowledge of computability but some basic familiar could be useful: Sipser (Into to theory of computation) is a great source. Victor Shoup's book (Computational Introduction to Number Theory and Algebra) is a great source for the number theory we'll need (and much more!).

1 Mathematical Proofs

Perhaps *the* mathematical prerequisite needed for this course is a certain level of comfort with mathematical proofs. While in everyday life we might use "proof" to describe a fairly convincing argument, in mathematics a proof is an argument that is convincing *beyond any shadow of a doubt.*[1] For example, consider the following mathematical statement:

Every even number greater than 2 *is equal to the sum of two primes.*

This statement, known as "Goldbach's Conjecture", was conjectured to be true by Christian Goldbach in 1742 (4 years before Princeton was founded). In the more than 250 years that have passed since, no one has ever found a counterexample to this statement. In fact, it has been verified to be true for all even numbers from 4 till $100,000,000,000,000,000$. Yet still it is not considered proven, since we have not ruled out the possibility that there is some (very large) even number that cannot be expressed as the sum of two primes.

The fact that a mathematical proof has to be absolutely convincing does not mean that it has to be overly formal and tedious. It just has to be clearly written, and contain no logical gaps. When you write proofs try to be clear and concise, rather than using too much formal notation. When

[1] In a famous joke, as a mathematician and an engineer drive in Scotland they see a white sheep on their left side. The engineer says "you see: all the sheep in Scotland are white". The mathematician replies "All I see is that there exists a sheep in Scotland whose right side is white".

you read proofs, try to ask yourself at every statement "am I really convinced that this statement is true?".

Of course, to be absolutely convinced that some statement is true, we need to be certain of what that statement means. This why there is a special emphasis in mathematics on very precise *definitions*. Whenever you read a definition, try to make sure you completely understand it, perhaps by working through some simple examples. Oftentimes, understanding the meaning of a mathematical statement is more than half the work to prove that it is true.

Example: Here is an example for a classical mathematical proof, written by Euclid around 300 B.C. Recall that a *prime number* is an integer $p > 1$ whose only divisors are p and 1, and that every number n is a product of prime numbers. Euclid's Theorem is the following:

Theorem 1. *There exist infinitely many primes.*

Before proving it, let's see that we understand what this statement means. It simply means that for every natural number k, there are more than k primes, and hence the number of primes is not finite.

At first, one might think it's obvious that there are infinitely many primes because there are infinitely many natural numbers, and each natural number is a product of primes. However, this is faulty reasoning: for example, the set of numbers of the form 3^n is infinite, even though their only factor is the single prime 3.

To prove Theorem 1, we use the technique of *proof by contradiction*. That is, we assume it is false and try to derive a contradiction from that assumption. Indeed, assume that all the primes can be enumerated as $p_1, p_2, \ldots, p_k$ for some number k. Define the number $n = p_1 p_2 \cdots p_k + 1$. Since we assume that the numbers $p_1, \ldots, p_k$ are *all* the primes, all of n's prime factors must come from this set, and in particular there is some i between 1 and k such that p_i divides n. That is, $n = p_i m$ for some number m. Thus,

$$p_i m = p_1 p_2 \cdots p_k + 1$$

or equivalently,

$$p_i m - p_1 p_2 \cdots p_k = 1 \quad .$$

But dividing both sides of this equation by p_i, we will get a whole number on the left hand side (as p_i is a factor of $p_1 p_2 \cdots p_k$) and the fraction $1/p_i$ on the right hand side, deriving a contradiction. This allows us to rightfully place the QED symbol $\square$ and consider Theorem 1 as proven.

2 Preliminaries

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 $|A| = |B|$.

I also assume familiarity with *big-Oh notation*: If f, g are two functions from $\mathbb{N}$ to $\mathbb{N}$, then **(1)** $f = O(g)$ if there exists a constant c such that $f(n) \leq c \cdot g(n)$ for every sufficiently large n, **(2)** $f = \Omega(g)$ if $g = O(f)$, **(3)** $f = \Theta(g)$ is $f = O(g)$ and $g = O(f)$, **(4)** $f = o(g)$ if for every $\epsilon > 0$, $f(n) \leq \epsilon \cdot g(n)$ for every sufficiently large n, and **(5)** $f = \omega(g)$ if $g = o(f)$.

To emphasize the input parameter, I often write $f(n) = O(g(n))$ instead of $f = O(g)$, and use similar notation for $o, \Omega, \omega, \Theta$.

3 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 $x \leftarrow_{\mathrm{R}} \{0, 1\}^n$.

4 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

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

Proof: Let $S = \{x : \#\text{ of 1's in } x \leq 50\}$. Let f be the function that flips all the bits of x from 1 to 0 and vice versa. Then f is a one-to-one and onto function from S to $\overline{S} = \{0, 1\}^{101} \setminus S$, meaning that $|S| = |\overline{S}| = \frac{2^{101}}{2}$. □

5 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 2 (Union bound)**.** *If Ω is a sample space and $E, E' \subseteq \Omega$ are two events over Ω. Then,*

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

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 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 $N \leq 50$ and the formula above can be phrased as

$$\Pr_{x \leftarrow_R \{0,1\}^{101}}[N(x) \leq 50] = \frac{1}{2}$$

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

7 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:

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

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 (this follows from the linearity of expectation, see below).

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 5, 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 $\mathbb{E}[cN] = c\mathbb{E}[N]$.

For example, the random variable N above is equal to $X_1 + \cdots + X_{101}$ with X_i equalling the i^{th} bit of the chosen string. Since $\mathbb{E}[X_i] = (1/2) \cdot 0 + (1/2) \cdot 1 = 1/2$, $\mathbb{E}[N] = 101 \cdot (1/2) = 50.5$.

8 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 3 (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. □

9 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 $\mathbb{E}[X] = \mu$,

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

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 $Var[X]$.

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

Theorem 4 (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

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

□

10 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.

11 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 5 (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,*

$$\Pr\left[\left|\frac{\sum X_i}{n} - \mu\right| > \epsilon\right] < 2^{-\epsilon^2 n/4}$$