Lectures 5,6 (Nov. 15 and 17, 2011)

From the point of view of general probability theory, we have so far in this course been engaged in the computation of expectation values of random variables. And not just any old random variables. The finite, combinatorial nature of the applications meant that our random variables X were "counting something". More precisely, they usually had the following properties :

(i) they were non-negative integer valued

(ii) they could be expressed as sums of identically distributed indicator variables.

From such computations we have been able to deduce interesting existence results, using essentially nothing more complicated than Proposition 10 and Corollary 11. From now on, the nature of our applications will be characterised by the following types of requirements :

(A) we will be interested in proving that certain events occur with high probability, not just non-zero probability

(B) it will not be enough to be able to compute $\mathbb{E}[X]$, we will also require information on how much X is "spread out" around its mean.

As indicated by Corollary 11, if you have a non-negative integer valued RV, then if you want to prove that $X = 0$ with high probability, it suffices to show that $\mathbb{E}[X]$ is "much smaller than 1". Actually, what one is using here is a very simple, but very useful general result, whose trivial proof we leave as an exercise :

Proposition 24 (Markov's Inequality) *Let X be a non-negative real valued RV, and $\alpha \geq 1$. Then*

$$\mathbb{P}(X \geq \alpha \mathbb{E}[X]) \leq \frac{1}{\alpha}. \tag{38}$$

This is the simplest example of a so-called *concentration inequality*. For applications to come it is by itself far too weak, though it is used all over the place in the proofs of much stronger results. The other problem is that it is "one-sided", i.e.: it only bounds the probability of X being too large. For the application to showing that $X = 0$ with high probability, when $\mathbb{E}[X] \ll 1$, that's fine. But suppose now instead you're interested in showing that $X > 0$ with high probability. The natural thing to do is to first show that $\mathbb{E}[X]$ is large. But this is, by itself, not enough.

Example : Suppose $X = 0$ with probability $0,99$ and $X = 10,000,000$ with probability $0,01$. Then $\mathbb{E}[X] = 100,000$ is still very large, but the event $X > 0$ is highly unlikely. In the book of Alon and Spencer (the first edition of which was written when the Cold War still hadn't quite ended), X is the number of deaths from nuclear war in the next 12 months.

The problem with the X in the above example is obviously that it is too spread out. Our main theoretical task in the coming lectures will be to develop tools which allow us to determine that certain random variables of interest are not too spread out, and therefore attain values in certain ranges with high probability. Sometimes we'll have an application where it's enough to know that $X > 0$ with high probability given that $\mathbb{E}[X]$ is large. Other times, we'll want X to be located in a narrow range around its mean value with high probability.

In our analyses we will make full use of the simplifying properties (i) and (ii) of the kinds of RV:s we encounter in combinatorial applications. The main obstacle to obtaining stronger results will be that, in most cases, the indicator variables in question are not *independent* of one another. This is a crucial point. On the one hand, independence simplifies lots of probabilistic analysis immensely. On the other hand, even with the current state of knowledge (we're talking 2011 !), effective tools for dealing with dependent events are few and far between. The techniques that have been developed all basically rely on knowing that either the amount of interdependence is "small" in some precisely quantifiable manner, or that the dependencies are "correlated" (we avoid defining this term precisely for the moment). Otherwise, you're probably screwed in terms of getting proofs : you might as well get out your computer and run simulations.

The first method we discuss is the simplest but most important one :

Second Moment Method

Basically this involves studying the *variance* of a RV as a measure of how far it is spread out. To simplify matters, unless otherwise stated, all RV:s are henceforth assumed to be non-negative integer valued, even if some of the things we prove hold more generally, and even with the same proofs (left to the reader to investigate these matters). At a later point we will specialise to the case of sums of indicator variables.

DEFINITION 14 : Let X be a RV. The *variance* of X, written as $\mathrm{Var}[X]$, is defined as

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

The square root of the variance is called the *standard deviation*.

Using linearity of expectation, it's easy to show that (exercise, if you have never done it before !)

$$\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2. \tag{39}$$

NOTATION : $\mathbb{E}[X] := \mu_X$, $\sqrt{\mathrm{Var}[X]} := \sigma_X$. We drop the subscripts when there can be no confusion about what RV is being considered.

Remark At this point it is worth clarifying the terminology *second moment method*. Let X be a RV. The *exponential generating function* of X is the RV e^X. Thus

$$e^X = \sum_{k=0}^{\infty} \frac{X^k}{k!}.$$

Under suitable convergence conditions, linearity of expectation yields that

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

The quantity $\mathbb{E}[X^k]/k!$ in this expression is called the *k:th moment* of the r.v. X. From (39) we see that the variance of X involves its second moment, hence the name.

A rough analogy to studying the 2nd moment of a r.v. is to study the second derivative of a smooth function in calculus. And just as it is pretty hard to find a real-life situation where one is interested in the third derivative of a smooth function, so in probability theory it is pretty rare to study the third moment of a r.v. Basically, if you can't get a handle on the second moment, then you're probably in a whole lot of trouble !

Finally, it should now not come as a great shock that the term *first moment method* is applied when one just studies the expectation of a r.v. itself. So this is the method we've been using in the course up to now.

The basic concentration estimate involving variance is

Proposition 25 (Chebyshev's Inequality) *Let X be a r.v. with mean μ and standard deviation σ. Let $\lambda \geq 1$. Then*

$$\mathbb{P}(|X - \mu| \geq \lambda\sigma) \leq \frac{1}{\lambda^2}. \tag{40}$$

PROOF : Define a new r.v. Y by $Y := |X - \mu|^2$. Then the left-hand side of (40) is just, by definition of variance, $\mathbb{P}(Y \geq \lambda^2 \mathbb{E}[Y])$. Markov's inequality (38) now gives the result immediately.

Corollary 26 *Let X be a r.v., $\epsilon > 0$. Then*

$$\mathbb{P}(|X - \mu| \geq \epsilon\mu) \leq \frac{\sigma^2}{\epsilon^2\mu^2}. \tag{41}$$

In particular,

$$\mathbb{P}(X = 0) \leq \frac{\sigma^2}{\mu^2}. \tag{42}$$

PROOF : For the first part, take $\lambda = \epsilon\mu/\sigma$ in (40). For the second part, set $\epsilon = 1$.

According to this corollary, we get good concentration of X around its mean provided that $\mathrm{Var}[X]$ is small compared to $\mathbb{E}[X]^2$. We now specialise to the case where

$$X = X_1 + \cdots + X_n$$

is a sum of indicator RV:s. We do not assume the X_i to be identically distributed though. Indeed let us denote by A_i the event indicated by X_i and $p_i := \mathbb{P}(A_i)$. Thus

$$X_i = \begin{cases} 1, & \text{with probability } p_i, \\ 0, & \text{with probability } 1 - p_i. \end{cases}$$

Also denote $\mu_i := \mathbb{E}[X_i]$, $\sigma_i^2 := \mathrm{Var}[X_i]$. Clearly, $\mu_i = p_i$. Also, by (39) and the fact that $X_i^2 = X_i$ since X_i only takes on the values 0 and 1, we have

$$\sigma_i^2 = p_i - p_i^2 = p_i(1 - p_i). \tag{43}$$

We thus have the inequality

$$\sigma_i^2 \leq \mu_i. \tag{44}$$

Since in applications the individual probabilities p_i are usually very small (even if the number of events A_i is usually very large), we are not losing much information in using (44).

We want an expression for the variance of X. Using (39) and several applications of linearity of expectation (LOE from now on), we obtain that

$$\sigma^2 = \sum_{i=1}^{n} \sigma_i^2 + \sum_{i \neq j} \mathrm{Cov}(X_i, X_j), \tag{45}$$

where the *covariance* of X_i and X_j is defined by

$$\mathrm{Cov}(X_i, X_j) = \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\mathbb{E}[X_j].$$

By (44) and LOE, the first sum on the right of (45) is at most μ. This is good, since we are interested in having σ^2 much smaller than μ^2 in situations where μ is large. So we can focus in on the sum of covariances. Since the X_i are indicator variables, we have

$$\mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\mathbb{E}[X_j] = \mathbb{P}(A_i \cap A_j) - \mathbb{P}(A_i)\mathbb{P}(A_j).$$

Hence $\mathrm{Cov}(X_i, X_j) = 0$ if the events A_i and A_j are independent[1]. So independent pairs don't contribute anything to the sum. Let $i \sim j$ denote that

[1]More generally, for any two random variables X and Y, if X and Y are *independent* then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$, though the converse need not hold (find an example !). What does it mean for two random variables to be independent in general ? It means simply what one would expect, namely that knowledge of the value of one variable does not give any information on the value of the other. There are several equivalent ways to express this formally. In the finite setting the following definition suffices : we say that real-valued RV:s X and Y are *independent* if, for all real numbers r, s, $\mathbb{P}(X = r | Y = s) = \mathbb{P}(X = r)$ and $\mathbb{P}(Y = r | X = s) = \mathbb{P}(Y = r)$. In words, the probability that X (resp. Y) attains the value r given that Y (resp. X) is known to have the value s, is the same as it was before the value of Y (resp. X) was known.

events A_i and A_j are not independent. We have at the very least the bound

$$\sum_{i \neq j} \mathrm{Cov}(X_i, X_j) \leq \sum_{i \sim j} \mathbb{P}(A_i \cap A_j).$$

Since $\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i) \cdot \mathbb{P}(A_j | A_i)$, we can rewrite the last sum as a double-sum, namely

$$\sum_{i \sim j} \mathbb{P}(A_i \cap A_j) = \sum_i \mathbb{P}(A_i) \sum_{j \sim i} \mathbb{P}(A_j | A_i).$$

Let us now make one further simplifying assumption, namely that the inner sum above is independent of i. This is a kind of "symmetry" requirement which holds for most applications. Following Alon-Spencer, we now denote the inner sum Δ^*. Thus we have

$$\sum_{i \neq j} \mathrm{Cov}(X_i, X_j) \leq \Delta^* \cdot \sum_i \mathbb{P}(A_i) = \Delta^* \cdot \sum_i \mu_i = \Delta^* \cdot \mu.$$

So let's summarise where we stand : assuming that our r.v. X is a sum of indicator variables, and that a certain symmetry condition is fulfilled, we have that

$$\mathrm{Var}[X] \leq (1 + \Delta^*) \mathbb{E}[X].$$

Hence, to show that $\mathrm{Var}[X]$ is much smaller than $\mathbb{E}[X]^2$, it suffices to show that Δ^* is much smaller than $\mathbb{E}[X]$. This is the crux of the second moment method.

Application : Distinct subset sums

We now describe an application of the second moment method to a problem in number theory. It is a relatively simple application from a theoretical viewpoint, in that it only uses Chebyshev's inequality, and (45) in the special case where the indicator variables are independent, and hence all the covariances are zero.

Definition 15 : Let $A = \{a_1, ..., a_k\}$ be a finite set of integers. A is said to have *distinct subset sums* if, for every two distinct subsets I, J of $\{1, ..., k\}$, the sums $\sum_{i \in I} a_i$ and $\sum_{j \in J} a_j$ have different values[2].

Let $f(n)$ be the maximum possible size of a subset of $\{1, ..., n\}$ which has distinct subset sums.

Lower Bounds :

Take $n = 2^k$ and $A = \{2^i : 0 \leq i \leq k\}$. This example shows that

[2]If I is the empty set, the sum is assigned the value zero. The definition extends to infinite sets, but the notation will just become a bit more complicated.

$f(n) \geq 1 + \lfloor \log_2 n \rfloor$. Erdős offered 500 dollars for a proof that there exists a universal constant C such that $f(n) \leq \log_2 n + C$. Note that he's not asking here for a computation of the optimal C or even a decent estimate of it, just a proof that some such constant exists, in other words that $f(n) = \log_2 n + O(1)$. The base-2 example shows that $C \geq 1$. If we confine ourselves to integer C then a number of authors, starting with John Conway and Richard Guy in 1969, have produced examples showing that $C \geq 2$. The point here is that the powers-of-2 example is not optimal. Note that, in order to get a better lower bound on C, it suffices to do so for a single n, because of the following trick : if $A = \{a_1, ..., a_k\}$ is a subset of $\{1, ..., n\}$ with distinct subset sums, and u is any odd number s.t. $1 \leq u \leq 2n$, then $A' = \{2a_1, ..., 2a_k, u\}$ is a subset of $\{1, ..., 2n\}$ with distinct subset sums and one additional element. This means that if $f(n) > \log_2 n + C$ then $f(N) > \log_2 N + C$ for every N of the form $N = 2^t n$.

One can then use a computer to help find individual examples ... For up-to-date information on lower bounds see, for example,

http://garden.imacs.sfu.ca/?q=op/sets_with_distinct_subset_sums

Upper Bounds :

If A has size k and is contained in $\{1, ..., n\}$ then there are 2^k distinct subset sums and each is among $\left\{0, ..., nk - \frac{k(k-1)}{2}\right\}$. Thus

$$2^{f(n)} \leq 1 + nf(n) - \frac{f(n)(f(n)-1)}{2}. \tag{46}$$

Taking base-2 logs, we have

$$f(n) \leq \log_2 n + \log_2 f(n) + O(1), \tag{47}$$

which leads to a bound of the form

$$f(n) \leq \log_2 n + \log_2 \log_2 n + O(1). \tag{48}$$

Erdős improved this to the following

Theroem 27

$$f(n) \leq \log_2 n + \frac{1}{2} \log_2 \log_2 n + O(1). \tag{49}$$

Proof : The idea is to refine the basic counting argument which leads to (48) by using the fact that the 2^k subset sums for a set $A = \{a_1, ..., a_k\}$ are not "uniformly distributed" in the interval $\left[0, nk - \frac{k(k-1)}{2}\right]$, but that there is a higher concentration of sums close to the mean. To make this precise requires a second moment analysis, which we now perform in detail.

Let $A = \{a_1, ..., a_k\}$ be a subset of $\{1, ..., n\}$ with distinct subset sums.

For each $i = 1, ..., k$, let X_i be the r.v. given by

$$X_i = \begin{cases} a_i, & \text{with probability } 1/2, \\ 0, & \text{with probability } 1/2. \end{cases} \tag{50}$$

The X_i:s are assumed to be independent, and we let $X := \sum_{i=1}^k X_i$. In words, X is the value of a subset sum of A, where the subset is chosen uniformly at random from all 2^k subsets of A. Though it is of no interest for the proof, note that, by LOE,

$$\mu = \mathbb{E}[X] = \frac{1}{2}\left(\sum_{i=1}^k a_i\right). \tag{51}$$

What we are interested in is the variance. By (43), (45) and independence, we have

$$\sigma^2 = \mathrm{Var}(X) = \frac{1}{4}\left(\sum_{i=1}^k a_i^2\right) \leq \frac{kn^2}{4}, \tag{52}$$

hence $\sigma \leq n\sqrt{k}/2$. Now let $\lambda \geq 1$. By Chebyshev's inequality,

$$\mathbb{P}\left(|X - \mu| \geq \frac{\lambda n\sqrt{k}}{2}\right) \leq \frac{1}{\lambda^2}. \tag{53}$$

This is equivalent to saying that

$$\mathbb{P}\left(|X - \mu| < \frac{\lambda n\sqrt{k}}{2}\right) \geq 1 - \frac{1}{\lambda^2}. \tag{54}$$

Now, on the one hand, X is integer-valued, and the number of integers satisfying $|X - \mu| < \frac{\lambda n\sqrt{k}}{2}$ is less than $1 + \lambda n\sqrt{k}$. On the other hand, (54) says that the probability that a uniformly randomly chosen subset sum satisfies this inequality is at least $1 - 1/\lambda^2$. Since there are 2^k subset sums, and they are assumed to be all distinct, it follows that there must be at least $\left(1 - \frac{1}{\lambda^2}\right) 2^k$ integers satisfying the inequality. We conclude that

$$\left(1 - \frac{1}{\lambda^2}\right) 2^{f(n)} < 1 + \lambda n\sqrt{f(n)}. \tag{55}$$

Taking base-2 logs, we have

$$f(n) \leq \log_2 n + \frac{1}{2}\log_2 f(n) + O(1), \tag{56}$$

where the $O(1)$-term depends on λ. From this one easily deduces (49).

Application : Average number of prime divisors

I did not have time to go through this example, but I handed out the text from Alon-Spencer. Note that this application is slightly more complicated, since there are non-zero covariances.

Interlude : Randomized Algorithms

At this point we took a break from the applications of the second moment method, and Devdatt gave an introduction to the algorithmic perspective on the general probabilistic method. Lecture notes will not be written up (by me at any rate) for this part.