Lecture 9 (Nov. 29, 2011)

A central class of results in probability theory are so-called *Central Limit Theorems*. A weaker set of results, called *Laws of Large Numbers*, capture the layman's notion that things tend to average out over time. The Central Limit Theorems are more precise : they tell you that random variables which are long-term averages tend to have normal distributions. Recall that the normal distribution with mean μ and standard deviation σ is the real-valued random variable $X = N(\mu, \sigma)$ such that, for every $z \in \mathbb{R}_+$,

$$\mathbb{P}(|X - \mu| \geq z) = 2 \cdot \int_z^\infty \frac{1}{\sqrt{2\pi}} e^{-t^2/2\sigma} \, dt. \tag{60}$$

The normal distribution is thus well concentrated about its mean. For example, (60) implies that, for any $\lambda > 0$,

$$\mathbb{P}(|X - \mu| \geq \lambda\sigma) \leq e^{-\lambda^2/2}.$$

This should be compared with the totally general Chebyshev inequality.

Classically, "the" Central Limit Theorem is about sums of independent, identically distributed (i.i.d.) random variables. It is an old result which says, basically, that if $X_1, X_2, ...$ is a sequence of i.i.d. random variables, each with mean μ and variance σ, and $Y_n = (X_1 + \cdots + X_n)/n$ is the average of the first n of them, then Y_n approaches $N(\mu, \sigma)$. What do we mean here by "approaches" ? Well, there are different possibilities, but the simplest notion, which is also the weakest and thus the easiest to get results about, is that, for every positive real number z,

$$\lim_{n\to\infty} \mathbb{P}(|Y_n - \mu| \geq z) = 2 \cdot \int_z^\infty \frac{1}{\sqrt{2\pi}} e^{-t^2/2\sigma} \, dt. \tag{61}$$

While the CLT is a fundamental theoretical result, there are several problems associated with its application :

(I) it assumes identical distributions
(II) it assumes independence
(III) it is qualitiative, not a quantitative result. In other words, it doesn't say anything about the rate of convergence to the limit in (61).

Problem (I) is not serious : the CLT can be extended to sums of variables with different distributions. (II) and (III) are much more serious, though. There are CL Theorems that concern dependent variables, but results are limited. In a seminal paper, Chernoff (1952) dealt significantly with problem (III). His results concern sums of independent indicator variables. Chernoff was interested in statistics, and his results are of great importance in that field. We've already seen in this course that sums of indicator variables are also ubiquitous in combinatorial applications, so Chernoff's results deserve attention. On the other hand, the fact that they don't address the

issue of independence limits their applicability. Nevertheless, the methods employed by Chernoff lay the foundation for much subsequent work on addressing the independence issue and knowledge of his method (and of the various qualitative CL Theorems) is a prerequisite for appreciating these later developments. We thus present a detailed proof of the following result

Theorem 31 (Chernoff's bound) *Let X be a random variable which is a sum of independent indicator variables. Let $\mathbb{E}[X] := \mu$. Then for any $\epsilon > 0$ there exists a positive constant c_ϵ, depending only on ϵ, such that*

$$\mathbb{P}(|X - \mu| > \epsilon\mu) < 2e^{-c_\epsilon \mu}. \tag{62}$$

In fact one can take

$$c_\epsilon = \min\left\{\frac{\epsilon^2}{2}, (1+\epsilon)\ln(1+\epsilon) - \epsilon\right\}. \tag{63}$$

The crucial point here is that c_ϵ does not depend on X, i.e.: it doesn't depend on how many indicator variables X is the sum of, nor on the distributions of these.

We will deduce Theorem 31 from a *normalised* version of it. Let X_i be an indicator variable, say

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

The *normalisation* of X_i, which we denote $\hat{X}_i$, is the variable $X_i - p_i$, i.e.:

$$\hat{X}_i = \begin{cases} 1 - p_i, & \text{with probability } p_i, \\ -p_i, & \text{with probability } 1 - p_i. \end{cases} \tag{64}$$

Thus $\hat{X}_i$ has mean zero. It has the same variance as X_i, namely $p_i(1-p_i)$.

Now let $\hat{X}$ be a r.v. which is a sum of n normalised indicator variables, for some fixed n. Write $\hat{X} = \hat{X}_1 + \cdots + \hat{X}_n$, with the $\hat{X}_i$ as above, and define the number p by $np = p_1 + \cdots + p_n$. Finally, let a be any positive real number. We will prove the following two inequalities :

$$\mathbb{P}(\hat{X} > a) < \exp\left[a - pn\ln\left(1 + \frac{a}{pn}\right) - a\ln\left(1 + \frac{a}{pn}\right)\right], \tag{65}$$

$$\mathbb{P}(\hat{X} < -a) < \exp\left[-\frac{a^2}{2pn}\right]. \tag{66}$$

Note that the theorem follows from (65) and (66) upon setting $a = \epsilon pn$. We will prove (65) in detail. The proof of (66) is very similar and thus omitted, but the full proof can be found in [AS], Appendix A. First, though, a couple of remarks are in order :

(i) there is an obvious asymmetry in the estimates (65) and (66), depending

on whether $\hat{X}$ is positive or negative. Unfortunately, this is a feature of Chernoff's method.
(ii) the connection to the normal distribution is clear in (66), as the variance of $\hat{X}$ is about np if the individual p_i are small, as is usually the case in applications. With (65), the connection is not so obvious. However, if a is small compared to pn and we use the fact that $\ln(1+u) \geq u - u^2/2$ when $0 < u < 1$, then we can deduce from (65) that

$$\mathbb{P}(\hat{X} > a) < \exp\left[-\frac{a^2}{2pn} + \frac{a^3}{2(pn)^2}\right]. \tag{67}$$

Note that (67) gives no information when a is large compared to np as then the cubic term dominates. Again, this is a feature of Chernoff's method, but is not important, since we're only interested in having concentration close to the mean anyway.

Proof of (65) : The proof uses the *exponential generating function* of $\hat{X}$, namely : Let $\lambda > 0$. Then we will consider the r.v.

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

Now $\hat{X} > a$ if and only if $e^{\lambda \hat{X}} > e^{\lambda a}$. The simple Markov inequality gives a bound

$$P(e^{\lambda \hat{X}} > e^{\lambda a}) < \frac{\mathbb{E}[e^{\lambda \hat{X}}]}{e^{\lambda a}}. \tag{68}$$

We will estimate the expectation and then the clever part of the proof is that λ, which at this point is still some arbitrary positive real number, will be chosen so as to minimise the right-hand side of (68). The estimate of the expectation will use the concavity of the logarithm. Let us begin by formally defining what this means :

Definition 21 : A function f on the positive reals is said to be *concave* if, for any n, any positive reals $x_1 \leq x_2 \leq \cdots \leq x_n$ and any positive reals $a_1, ..., a_n$ satisfying $\sum a_i = 1$, it holds that

$$f\left(\sum_{i=1}^{n} a_i x_i\right) \geq \sum_{i=1}^{n} a_i f(x_i).$$

Concavity has a simple geometric interpretation, namely that the graph of f lies on or above the straight line drawn between any two points on it.

Lemma 32 *Let $C > 0$. Then the function $f(x) = \ln(Cx + 1)$ is concave.*

Proof of Lemma : Exercise.

Now let us return to the proof of (65). Since $\hat{X} = \sum \hat{X}_i$, one easily sees that

$$e^{\lambda \hat{X}} = \prod_{i=1}^{n} e^{\lambda \hat{X}_i}.$$

Now we use the independence of the $\hat{X}_i$. Recall that if A, B are independent random variables, then $\mathbb{E}[AB] = \mathbb{E}[A]\mathbb{E}[B]$. Thus, by induction,

$$\mathbb{E}[e^{\lambda \hat{X}}] = \prod_{i=1}^{n} \mathbb{E}[e^{\lambda \hat{X}_i}]. \tag{69}$$

But from (64), the definition of e.g.f. and linearity of expectation (convergence is not a problem), one easily computes that

$$\mathbb{E}[e^{\lambda \hat{X}_i}] = p_i e^{\lambda(1-p_i)} + (1-p_i)e^{-\lambda p_i} = e^{-\lambda p_i}\left[p_i(e^{\lambda}-1)+1\right].$$

Substituting into (69) and recalling the definition of p, we thus have

$$\mathbb{E}[e^{\lambda \hat{X}}] = e^{-\lambda pn} \prod_{i=1}^{n} [p_i(e^{\lambda}-1)+1]. \tag{70}$$

But

$$\prod_{i=1}^{n} [p_i(e^{\lambda}-1)+1] \leq [p(e^{\lambda}-1)+1]^n. \tag{71}$$

Indeed this follows from taking logarithms and using Lemma 32. So substituting (71) back into (70) and in turn back into (68), we have the estimate

$$\mathbb{P}(\hat{X} > a) < e^{-\lambda pn}[pe^{\lambda} + (1-p)]^n e^{-\lambda a}. \tag{72}$$

It is now a horrid calculus exercise to compute the precise value of λ which minimises the right hand side of (72)[1]. However, a good approximation when $a \ll np$ is to take $\lambda = \ln(1 + a/pn)$. Substituting this into (72) we get the desired relation (65) upon noticing that, with this choice of λ,

$$[pe^{\lambda} + (1-p)]^n = (1 + a/n)^n \leq e^a.$$

This completes the proof of Theorem 31. Applications will follow in the next lecture(s).

Lecture 10 (Dec. 1, 2011)

Devdatt presents an application of Chernoff bounds to *network routing*. See Section 4.5.1 of [MU].

[1]The right value turns out to be

$$\lambda = \ln\left[\left(\frac{1-p}{p}\right)\left(\frac{a+np}{n-(a+np)}\right)\right].$$