Lecture 34

Andrei Antonenko

April 22, 2005

1 Weak Law of Large Numbers

The weak law of large numbers asserts that the sample mean of a large number of independent identically distributed random variables is very close to the true mean, with high probability.

As in the introduction to this chapter, we consider a sequence $X_1, X_2, \dots$ of independent identically distributed random variables with mean μ and variance σ^2, and define the sample mean by

$$M_n = \frac{X_1 + \cdots + X_n}{n}. \tag{1}$$

We have

$$\mathbf{E}[M_n] = \frac{\mathbf{E}[X_1] + \cdots + \mathbf{E}[X_n]}{n} = \frac{n\mu}{n} = \mu, \tag{2}$$

and, using independence,

$$\mathbf{var}(M_n) = \frac{\mathbf{var}(X_1 + \cdots + X_n)}{n^2} = \frac{\mathbf{var}(X_1) + \cdots + \mathbf{var}(X_n)}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}. \tag{3}$$

We apply Chebyshev's inequality and obtain

$$P(|M_n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}, \tag{4}$$

for any $\epsilon > 0$. We observe that for any fixed $\epsilon > 0$, the right-hand side of this inequality goes to zero as n increases. As a consequence, we obtain the weak law of large numbers, which is stated below. It turns out that this law remains true even if the X_i have infinite variance, but a much more elaborate argument is needed, which we omit. The only assumption needed is that $\mathbf{E}[X_i]$ is well-defined and finite.

Theorem 1.1 (Weak Law of Large Numbers)**.** *Let $X_1, X_2, \dots$ be independent identically distributed random variables with mean μ. For every $\epsilon > 0$, we have*

$$P(|M_n - \mu| \geq \epsilon) = P\left(\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| \geq \epsilon\right) \to 0, \quad \text{as } n \to \infty. \tag{5}$$

The WLLN states that for large n, the "bulk" of the distribution of M_n is concentrated near μ. That is, if we consider a positive length interval $[\mu - \epsilon, \mu + \epsilon]$ around μ, then there is high probability that M_n will fall in that interval; as $n \to \infty$, this probability converges to 1. Of course, if ϵ is very small, we may have to wait longer (i.e., need a larger value of n) before we can assert that M_n is highly likely to fall in that interval.

Example 1.2 (Probabilities and Frequencies.)**.** Consider an event A defined in the context of some probabilistic experiment. Let $p = P(A)$ be the probability of that event. We consider n independent repetitions of the experiment, and let M_n be the fraction of time that event A occurred; in this context, M_n is often called the empirical frequency of A. Note that

$$M_n = \frac{X_1 + \cdots + X_n}{n},$$

where X_i is 1 whenever A occurs, and 0 otherwise; in particular, $\mathbf{E}[X_i] = p$. The weak law applies and shows that when n is large, the empirical frequency is most likely to be within ϵ of p. Loosely speaking, this allows us to say that empirical frequencies are faithful estimates of p. Alternatively, this is a step towards interpreting the probability p as the frequency of occurrence of A.

Example 1.3 (Polling.)**.** Let p be the fraction of voters who support a particular candidate for office. We interview n "randomly selected" voters and record the fraction M_n of them that support the candidate. We view M_n as our estimate of p and would like to investigate its properties. We interpret "randomly selected" to mean that the n voters are chosen independently and uniformly from the given population. Thus, the reply of each person interviewed can be viewed as an independent Bernoulli trial X_i with success probability p and variance $\sigma^2 = p(1-p)$. The Chebyshev inequality yields

$$P(|M_n - p| \geq \epsilon) \leq \frac{p(1-p)}{n\epsilon^2}.$$

The true value of the parameter p is assumed to be unknown. On the other hand, it is easily verified that

$$p(1-p) \leq 1/4,$$

which yields

$$P(|M_n - p| \geq \epsilon) \leq \frac{1}{4n\epsilon^2}.$$

For example, if $\epsilon = 0.1$ and $n = 100$, we obtain

$$P(|M_100 - p| \geq 0.1) \leq \frac{1}{4 \cdot 100 \cdot (0.1)^2} = 0.25.$$

In words, with a sample size of $n = 100$, the probability that our estimate is wrong by more than 0.1 is no larger than 0.25.

Suppose now that we impose some tight specifications on our poll. We would like to have high confidence (probability at least 95%) that our estimate will be very accurate (within .01 of p). How many voters should be sampled? The only guarantee that we have at this point is the inequality

$$P(|M_n - p| \geq 0.01) \leq \frac{1}{4n(0.01)^2}.$$

We will be sure to satisfy the above specifications if we choose n large enough so that

$$\frac{1}{4n(0.01)^2} \leq 1 - 0.95 = 0.05,$$

which yields $n \geq 50,000$. This choice of n has the specified properties but is actually fairly conservative, because it is based on the rather loose Chebyshev inequality.