MASSACHUSETTS INSTITUTE OF TECHNOLOGY

6.436J/15.085J Fall 2008

Lecture 19 11/17/2008

LAWS OF LARGE NUMBERS – II

Contents

1 THE STRONG LAW OF LARGE NUMBERS

While the weak law of large numbers establishes convergence of the sample mean, in probability, the strong law establishes almost sure convergence.

Before we proceed, we point out two common methods for proving almost sure convergence.

Proposition 1: Let $\{X_n\}$ be a sequence of random variables, not necessarily independent.
(i) If $\sum_{n=1}^{\infty} \mathbb{E}[|X_n|^s] < \infty$, and $s > 0$, then $X_n \stackrel{\text{a.s.}}{\to} 0$.
(ii) If $\sum_{n=1}^{\infty} \mathbb{P}(|X_n| > \epsilon) < \infty$, for every $\epsilon > 0$, then $X_n \stackrel{\text{a.s.}}{\to} 0$.

Proof. (i) By the monotone convergence theorem, we obtain $\mathbb{E}[\sum_{n=1}^{\infty} |X_n|^s] < \infty$, which implies that the random variable $\sum_{n=1}^{\infty} |X_n|^s$ is finite, with probability 1. Therefore, $|X_n|^s \stackrel{\text{a.s.}}{\to} 0$, which also implies that $X_n \stackrel{\text{a.s.}}{\to} 0$.
(ii) Setting $\epsilon = 1/k$, for any positive integer k, the Borel-Cantelli Lemma shows that the event $\{|X_n| > 1/k\}$ occurs only a finite number of times, with probability 1. Thus, $\mathbb{P}(\limsup_{n\to\infty} X_n > 1/k) = 0$, for every positive integer k. Note that the sequence of events $\{\limsup_{n\to\infty} |X_n| > 1/k\}$ is monotone and converges to the event $\{\limsup_{n\to\infty} |X_n| > 0\}$. The continuity of probability measures implies that $\mathbb{P}(\limsup_{n\to\infty} |X_n| > 0) = 0$. This establishes that $X_n \stackrel{\text{a.s.}}{\to} 0$. $\square$

Theorem 1: Let $X, X_1, X_2, \dots$ be i.i.d. random variables, and assume that $\mathbb{E}[|X|] < \infty$. Let $S_n = X_1 + \cdots + X_n$. Then, S_n/n converges almost surely to $\mathbb{E}[X]$.

Proof, assuming finite fourth moments. Let us make the additional assumption that $\mathbb{E}[X^4] < \infty$. Note that this implies $\mathbb{E}[|X|] < \infty$. Indeed, using the inequality $|x| \leq 1 + x^4$, we have

$$\mathbb{E}\big[|X|\big] \leq 1 + \mathbb{E}[X^4] < \infty.$$

Let us assume first that $\mathbb{E}[X] = 0$. We will show that

$$\mathbb{E}\left[\sum_{n=1}^{\infty} \frac{(X_1 + \cdots + X_n)^4}{n^4}\right] < \infty.$$

We have

$$\mathbb{E}\left[\frac{(X_1 + \cdots + X_n)^4}{n^4}\right] = \frac{1}{n^4} \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \sum_{i_3=1}^{n} \sum_{i_4=1}^{n} \mathbb{E}[X_{i_1} X_{i_2} X_{i_3} X_{i_4}].$$

Let us consider the various terms in this sum. If one of the indices is different from all of the other indices, the corresponding term is equal to zero. For example, if i_1 is different from i_2, i_3, or i_4, the assumption $\mathbb{E}[X_i] = 0$ yields

$$\mathbb{E}[X_{i_1} X_{i_2} X_{i_3} X_{i_4}] = \mathbb{E}[X_{i_1}]\, \mathbb{E}[X_{i_2} X_{i_3} X_{i_4}] = 0.$$

Therefore, the nonzero terms in the above sum are either of the form $\mathbb{E}[X_i^4]$ (there are n such terms), or of the form $\mathbb{E}[X_i^2 X_j^2]$, with $i \neq j$. Let us count the number of terms of the second type. Such terms are obtained in three different ways: by setting $i_1 = i_2 \neq i_3 = i_4$, or by setting $i_1 = i_3 \neq i_2 = i_4$, or by setting $i_1 = i_4 \neq i_2 = i_3$. For each one of these three ways, we have n choices for the first pair of indices, and $n - 1$ choices for the second pair. We conclude that there are $3n(n-1)$ terms of this type. Thus,

$$\mathbb{E}\left[\frac{(X_1 + \cdots + X_n)^4}{n^4}\right] = \frac{n\mathbb{E}[X_1^4] + 3n(n-1)\mathbb{E}[X_1^2 X_2^2]}{n^4}.$$

Using the inequality $xy \leq (x^2 + y^2)/2$, we obtain $\mathbb{E}[X_1^2 X_2^2] \leq \mathbb{E}[X_1^4]$, and

$$\mathbb{E}\left[\frac{(X_1 + \cdots + X_n)^4}{n^4}\right] \leq \frac{\big(n + 3n(n-1)\big)\mathbb{E}[X_1^4]}{n^4} \leq \frac{3n^2\mathbb{E}[X_1^4]}{n^4} = \frac{3\mathbb{E}[X_1^4]}{n^2}.$$

It follows that

$$\mathbb{E}\left[\sum_{n=1}^{\infty}\frac{(X_1+\cdots+X_n)^4}{n^4}\right]=\sum_{n=1}^{\infty}\frac{1}{n^4}\mathbb{E}\big[(X_1+\cdots+X_n)^4\big]\leq\sum_{n=1}^{\infty}\frac{3}{n^2}\mathbb{E}[X_1^4]<\infty,$$

where the last step uses the well known property $\sum_{n=1}^{\infty} n^{-2}<\infty$. This implies that $(X_1+\cdots+X_n)^4/n^4$ converges to zero with probability 1, and therefore, $(X_1+\cdots+X_n)/n$ also converges to zero with probability 1, which is the strong law of large numbers.

For the more general case where the mean of the random variables X_i is nonzero, the preceding argument establishes that $\big(X_1+\cdots+X_n-n\mathbb{E}[X_1]\big)/n$ converges to zero, which is the same as $(X_1+\cdots+X_n)/n$ converging to $\mathbb{E}[X_1]$, with probability 1. $\square$

Proof, assuming finite second moments. We now consider the case where we only assume that $\mathbb{E}[X^2]<\infty$. We have

$$\mathbb{E}\left[\Big(\frac{S_n}{n}-\mu\Big)^2\right]=\frac{\mathrm{var}(X)}{n}.$$

If we only consider values of n that are perfect squares, we obtain

$$\sum_{i=1}^{\infty}\mathbb{E}\left[\Big(\frac{S_{i^2}}{i^2}-\mu\Big)^2\right]=\sum_{i=1}^{\infty}\frac{\mathrm{var}(X)}{i^2}<\infty,$$

which implies that $\big((S_{i^2}/i^2)-\mathbb{E}[X]\big)^2$ converges to zero, with probability 1. Therefore, S_{i^2}/i^2 converges to $\mathbb{E}[X]$, with probability 1.

Suppose that the random variables X_i are nonnegative. Consider some n such that $i^2\leq n<(i+1)^2$. We then have $S_{i^2}\leq S_n\leq S_{(i+1)^2}$. It follows that

$$\frac{S_{i^2}}{(i+1)^2}\leq\frac{S_n}{n}\leq\frac{S_{(i+1)^2}}{i^2},$$

or

$$\frac{i^2}{(i+1)^2}\cdot\frac{S_{i^2}}{i^2}\leq\frac{S_n}{n}\leq\frac{(i+1)^2}{i^2}\cdot\frac{S_{(i+1)^2}}{(i+1)^2}.$$

As $n\to\infty$, we also have $i\to\infty$. Since $i/(i+1)\to 1$, and since $S_{i^2}\cdot i^2$ converges to $\mathbb{E}[X]$, with probability 1, we see that for almost all sample points, S_n/n is sandwiched between two sequences that converge to $\mathbb{E}[X]$. This proves that $S_n/n\to\mathbb{E}[X]$, with probability 1.

For a general random variable X, we write it in the form $X=X^+-X^-$, where X^+ and X^- are nonnegative. The strong law applied to X^- and X^- separately, implies the strong law for X as well. $\square$

The proof for the most general case (finite mean, but possibly infinite variance) is omitted. It involves truncating the distribution of X, so that its moments are all finite, and then verifying that the "errors" due to such truncation are not significant in the limit.

2 The Chernoff bound

Let again $X, X_1, \ldots$ be i.i.d., and $S_n = X_1 + \cdots + X_n$. Let us assume, for simplicity, that $\mathbb{E}[X] = 0$. According to the weak law of large numbers, we know that $\mathbb{P}(S_n \geq na) \to 0$, for every $a > 0$. We are interested in a more detailed estimate of $\mathbb{P}(S_n \geq na)$, involving the rate at which this probability converges to zero. It turns out that if the moment generating function of X is finite on some interval $[0, c]$ (where $c > 0$), then $\mathbb{P}(S_n \geq na)$ decays exponentially with n, and much is known about the precise rate of exponential decay.

2.1 Upper bound

Let $M(s) = \mathbb{E}[e^{sX}]$, and assume that $M(s) < \infty$, for $s \in [0, c]$, where $c > 0$. Recall that $M_{S_n}(s) = \mathbb{E}[e^{s(X_1+\cdots+X_n)}] = (M(s))^n$. For any $s > 0$, the Markov inequality yields

$$\mathbb{P}(S_n \geq na) = \mathbb{P}(e^{sS_n} \geq e^{nsa}) \leq e^{-nsa}\mathbb{E}[e^{sS_n}] = e^{-nsa}(M(s))^n.$$

Every nonnegative value of s, gives us a particular bound on $\mathbb{P}(S_n \geq a)$. To obtain the tightest possible bound, we minimize over s, and obtain the following result.

Theorem 2. (Chernoff upper bound) Suppose that $\mathbb{E}[e^{sX}] < \infty$ for some $s > 0$, and that $a > 0$. Then,

$$\mathbb{P}(S_n \geq na) \leq e^{-n\phi(a)},$$

where

$$\phi(a) = \sup_{s \geq 0} \big(sa - \log M(s)\big).$$

For $s = 0$, we have

$$sa - \log M(s) = 0 - \log 1 = 0,$$

where we have used the generic property $M(0) = 1$. Furthermore,

$$\frac{d}{ds}\big(sa - \log M(s)\big)\Big|_{s=0} = a - \frac{1}{M(s)} \cdot \frac{d}{ds}M(s)\Big|_{s=0} = a - 1 \cdot \mathbb{E}[X] > 0.$$

Since the function $sa - \log M(s)$ is zero and has a positive derivative at $s = 0$, it must be positive when s is positive and small. It follows that the supremum $\phi(a)$ of the function $sa - \log M(s)$ over all $s \geq 0$ is also positive. In particular, for any fixed $a > 0$, the probability $\mathbb{P}(S_n \geq na)$ decays at least exponentially fast with n.

Example: For a standard normal random variable X, we have $M(s) = e^{s^2/2}$. Therefore, $sa - \log M(s) = sa - s^2/2$. To maximize this expression over all $s \geq 0$, we form the derivative, which is $a - s$, and set it to zero, resulting in $s = a$. Thus, $\phi(a) = a^2/2$, which leads to the bound

$$\mathbb{P}(X \geq a) \leq e^{-a^2/2}.$$

2.2 Lower bound

Remarkably, it turns out that the estimate $\phi(a)$ of the decay rate is tight, under minimal assumptions. To keep the argument simple, we introduce some simplifying assumptions.

Assumption 1.
(i) $M(s) = \mathbb{E}[e^{sX}] < \infty$, for all $s \in \mathbb{R}$.
(ii) The random variable X is continuous, with PDF f_X.
(iii) The random variable X does not admit finite upper and lower bounds. (Formally, $0 < F_X(x) < 1$, for all $x \in \mathbb{R}$.)

We then have the following lower bound.

Theorem 2. (Chernoff lower bound) Under Assumption 1, we have

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(S_n \geq na) = -\phi(a), \qquad (1)$$

for every $a > 0$.

We note two consequences of our assumptions, whose proof is left as an exercise:
(a) $\lim_{s\to\infty} \dfrac{\log M(s)}{s} = \infty$;
(b) $M(s)$ is differentiable at every s.

The first property guarantees that for any $a > 0$ we have $\lim_{s\to\infty}(\log M(s) - sa) = \infty$. Since $M(s) > 0$ for all s, and since $M(s)$ is differentiable, it follows that $\log M(s)$ is also differentiable and that there exists some $s^* \geq 0$ at which

$\log M(s) - sa$ is minimized over all $s \geq 0$. Taking derivatives, we see that such a s^* satisfies $a = M'(s^*)/M(s^*)$, where M' stands for the derivative of M. In particular,

$$\phi(a) = s^* a - \log M(s^*). \tag{2}$$

Let us introduce a new PDF

$$f_Y(x) = \frac{e^{s^* x}}{M(s^*)} f_X(x).$$

This is a legitimate PDF because

$$\int f_Y(x)\, dx = \frac{1}{M(s^*)} \int e^{s^* x} f_X(x)\, dx = \frac{1}{M(s^*)} \cdot M(s^*) = 1.$$

The moment generating function associated with the new PDF is

$$M_Y(s) = \frac{1}{M(s^*)} \int e^{sx} e^{s^* x} f_X(x)\, dx = \frac{M(s+s^*)}{M(s^*)}.$$

Thus,

$$\mathbb{E}[Y] = \frac{1}{M(s^*)} \cdot \frac{d}{ds} M(s+s^*)\Big|_{s=0} = \frac{M'(s^*)}{M(s^*)} = a,$$

where the last equality follows from our definition of s^*. The distribution of Y is called a "tilted" version of the distribution of X.

Let $Y_1, \ldots, Y_n$ be i.i.d. random variables with PDF f_Y. Because of the close relation between f_X and f_Y, approximate probabilities of events involving $Y_1, \ldots, Y_n$ can be used to obtain approximate probabilities of events involving $X_1, \ldots, X_n$.

We keep assuming that $a > 0$, and fix some $\delta > 0$. Let

$$B = \Big\{ (x_1, \ldots, x_n) \;\Big|\; a - \delta \leq \frac{1}{n} \sum_{i=1}^{n} x_i \leq a + \delta \Big\} \subset \mathbb{R}^n.$$

Let $S_n = X_1 + \ldots + X_n$ and $T_n = Y_1 + \ldots + Y_n$. We have

$$\begin{aligned}
\mathbb{P}\big(S_n \geq n(a-\delta)\big) &\geq \mathbb{P}\big(n(a-\delta) \leq S_n \leq n(a+\delta)\big) \\
&= \int_{(x_1,\ldots,x_n)\in B} f_X(x_1)\cdots f_X(x_n)\, dx_1 \cdots dx_n \\
&= \int_{(x_1,\ldots,x_n)\in B} (M(s^*))^n e^{-s^* x_1} f_Y(x_1) \cdots e^{-s^* x_n} f_Y(x_n)\, dx_1 \cdots dx_n \\
&\geq (M(s^*))^n e^{-ns^*(a+\delta)} \int_{(x_1,\ldots,x_n)\in B} f_Y(x_1)\cdots f_Y(x_n)\, dx_1 \cdots dx_n \\
&= (M(s^*))^n e^{-ns^*(a+\delta)}\, \mathbb{P}(T_n \in B). \qquad (3)
\end{aligned}$$

The second inequality above was obtained because for every $(x_1, \ldots, x_n)] \in B$, we have $x_1 + \cdots + x_n \leq n(a + \delta)$, so that $e^{-s^* x_1} \cdots e^{-s^* x_n} \geq e^{-ns^*(a+\delta)}$.

By the weak law of large numbers, we have

$$\mathbb{P}(T_n \in B) = \mathbb{P}\Big(\frac{Y_1 + \cdots + Y_n}{n} \in [na - n\delta,\, na + n\delta]\Big) \to 1,$$

as $n \to \infty$. Taking logarithms, dividing by n, and then taking the limit of the two sides of Eq. (3), and finally using Eq. (2), we obtain

$$\liminf_{n \to \infty} \frac{1}{n} \log \mathbb{P}(S_n > na) \geq \log M(s^*) - s^* a - \delta = -\phi(a) - \delta.$$

This inequality is true for every $\delta > 0$, which establishes the lower bound in Eq. (1).