NMAI059 – Probability and Statistics 1

Mykhaylo Tyomkyn

Lecture 9 - The laws of large numbers and the Central limit theorem.

Let $X_1, X_2, \dots$ be a sequence of independent, identically distributed L^1 random variables on some probability space. What can one say about the sequence $(S_n)_{n=1}^{\infty}$, where $S_n = (X_1 + \cdots + X_n)/n$? Since S_n measures the 'empirical average' of $X_1, \dots, X_n$, one might suspect that in some sense S_n converges to $\mu = \mathbb{E}(X_i)$. This is indeed the case, but how do you prove or even state it formally, i.e., what does it mean for a sequence of random variables to 'converge'? As it happens, there are multiple ways of formally defining convergence of random variables and we will introduce three of them. Here is the first.

Definition 1 (Convergence "in probability") *Let X and $X_1, X_2, \dots$ be random variables on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We say that the sequence $(X_n)_{n=1}^{\infty}$* converges in probability *to X, in notation $X_n \xrightarrow{P} X$, if for all $\epsilon > 0$ we have*[1]

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| \le \epsilon) = 1.$$

With this notion we can formalize our intuition about S_n. To simplify the proof we additionally assume that the X_i have finite variances.

Theorem 1 (Weak law of large numbers (WLLN), L^2-version) *Let $X_1, X_2, \dots$ be independent, identically distributed $L^2(\Omega, \mathcal{F}, \mathbb{P})$ random variables with $\mathbb{E}(X_i) = \mu$. Let $S_n = (X_1 + \cdots + X_n)/n$. Then*

$$S_n \xrightarrow{P} \mu.$$

In other words, the sequence $(S_n)_{n=1}^{\infty}$ converges in probability to the constant random variable, taking value μ always.

Proof Let σ^2 be the variance of each X_i. We have

$$\mathbb{E}(S_n) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(X_i) = \frac{n\mu}{n} = \mu,$$

and due to independence,

$$Var(S_n) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Fix some $\epsilon > 0$. Applying Chebyshev's inequality we obtain

$$\mathbb{P}(|S_n - \mu| \ge \epsilon) \le \frac{Var(S_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2},$$

[1] Equivalent versions: $\lim_{n\to\infty} \mathbb{P}(|X_n - X| < \epsilon) = 1$, $\lim_{n\to\infty} \mathbb{P}(|X_n - X| \ge \epsilon) = 0$, $\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \epsilon) = 0$.

which converges to 0 as $n \to \infty$ (since σ and ϵ are fixed). So, indeed $S_n \xrightarrow{P} \mu$. □

The above proof already works when the X_i are pairwise uncorrelated. In general, WLLN has many versions. For instance, it holds when the X_i are independent but merely in L^1. This, of course, would require a different proof.

Let us now introduce a second form of convergence of random variables.

Definition 2 (Convergence "almost surely") *Let X and $X_1, X_2, \dots$ be random variables on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We say that the sequence $(X_n)_{n=1}^{\infty}$* converges almost surely *to X, in notation $X_n \xrightarrow{a.s.} X$, if*

$$\mathbb{P}(\lim_{n\to\infty} X_n = X) = 1.$$

This notion of convergence is strictly stronger than $X_n \xrightarrow{P} X$.

Exercise 1 *(harder)*

- *Show that if $\mathcal{F}$ is a σ-algebra then $\{\lim_{n\to\infty} X_n = X\}$ is an event.*
- *Prove that $X_n \xrightarrow{a.s.} X$ implies $X_n \xrightarrow{P} X$.*

Example 1 (A sequence convergent in probability but not a.s.) *Let $\Omega = [0,1]$, $\mathcal{F} = \mathcal{L}$ and $\mathbb{P} = Unif[0,1]$. Let $X = 0$, and let*

$$\begin{aligned}
&X_1 = \mathbb{1}_{[0,1]}\\
&X_2 = \mathbb{1}_{[0,\frac{1}{2}]}, X_3 = \mathbb{1}_{[\frac{1}{2},1]},\\
&X_4 = \mathbb{1}_{[0,\frac{1}{4}]}, X_5 = \mathbb{1}_{[\frac{1}{4},\frac{2}{4}]}, X_6 = \mathbb{1}_{[\frac{2}{4},\frac{3}{4}]}, X_7 = \mathbb{1}_{[\frac{3}{4},1]},\\
&X_8 = \mathbb{1}_{[0,\frac{1}{8}]}, X_9 = \mathbb{1}_{[\frac{1}{8},\frac{2}{8}]}, X_{10} = \dots
\end{aligned}$$

Then $X_n \xrightarrow{P} X$, but not $X_n \xrightarrow{a.s.} X$ (in fact, $X_n(\omega)$ does not converge to $X(\omega)$ for any $\omega \in \Omega$).

We will now state (without proof) the strong law of large numbers.

Theorem 2 (Strong law of large numbers (SLLN)) *Let $X_1, X_2, \dots$ be independent, identically distributed $L^1(\Omega, \mathcal{F}, \mathbb{P})$ random variables with $\mathbb{E}(X_i) = \mu$. Let $S_n = (X_1 + \cdots + X_n)/n$. Then*

$$S_n \xrightarrow{a.s.} \mu.$$

The subtle difference between WLLN and SLLN can be better visualized, say when $X_i \sim Bern(1/2)$, using the infinite 'tree of outcomes' of the coin tosses. WLLN is a statement about the 'levels' of the tree, while SLLN is a statement about the 'branches'. Let us now look at two applications of the laws of large numbers.

Example 2 (Monte Carlo volume computation) *Suppose we are given a k-dimensional compact 'body' $D \subseteq [0,1]^k$, typically defined by a system of algebraic inequalities. We want to (approximately) compute the volume of D. We use the* Monte Carlo *method: create many points $x_1, x_2 \dots x_n$ in $[0,1]^k$ uniformly at random, independently, and for each i check whether $x_i \in D$ (checking is cheap: just plug* *the coordinates of x_i into the inequalities defining D). Let $X_i = 1$ if $x_i \in D$ and $X_i = 0$ otherwise. Then $X_1, \dots, X_n$ are iid*[2] *random variables with $X_i \sim Bern(vol(D))$. So, by the laws of large numbers, $S_n = (X_1 + \cdots + X_n)/n$ is a good estimate for $\mathbb{E}(X_i) = vol(D)$. How good exactly is a question of Statistics.*

Example 3 ("Normal" numbers) *A real number $x \in [0,1]$ is called* normal *if in its decimal representation $x = \sum_{i \geq 1} x_i 10^{-i}$ any sequence of digits $a = (a_1, \dots, a_k) \in \{0, \dots, 9\}^k$ appears in x consecutively with frequency 10^{-k}. In other words*

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{(x_i \dots, x_{i+k-1}) = a\}} = 10^{-k}.$$

Then SLLN implies that $X \sim Unif[0,1]$ is normal a.s. (to see this, apply SLLN to every possible finite sequence a, and take the countable intersection of probability 1 events). In other words 'almost all' numbers are normal. By the same argument (applied a countable number of times), $X \sim Unif[0,1]$ is a.s. absolutely normal, *that is normal in every expansion base $q \in \{2, 3, \dots\}$ ($q = 10$: normal). So, almost all numbers are absolutely normal, but in practice absolutely normal numbers are very hard to construct (e.g., rational numbers are not normal, as they are periodic). It is conjectured that $e, \pi, \sqrt{2}$ are absolutely normal, but this is open.*

Let us now introduce a third type of convergence of random variables, the weakest of the three but at the same time probably the most important one.

Definition 3 (Weak convergence) *A sequence $X_1, X_2, \dots$ of random variables (not necessarily defined on the same probability space) with respective cdf's $F_1, F_2, \dots$* converges in distribution / weakly /in law *to a random variable X with cdf F, if*

$$\lim_{n \to \infty} F_n(x) = F(x) \quad \textit{for every } x \in \mathbb{R} \textit{ at which } F \textit{ is continuous.}^{3}$$

We write $X_n \xrightarrow{\mathcal{L}} X$.

Exercise 2 *Prove that $X_n \xrightarrow{P} X$ implies $X_n \xrightarrow{\mathcal{L}} X$.*

The reverse implication, as a statement, would only make sense if all X_i and X were defined on the same probability space. But even then, the implication would not hold.

Example 4 *Let $X \sim Bern(1/2)$ and $Y = 1 - X$. Then Y is also $Bern(1/2)$-distributed, so $F_X = F_Y$. Consider now the sequence $X, Y, X, Y, X, Y, \dots$. It converges to X in distribution, since all the distributions are identical, but not in probability since $|Y - X| = 1$ always.*

The following theorem is arguably the most important theorem in Probability theory and Statistics, both in theoretical and practical terms. It underlines the importance of the normal distribution. We do not have time or mean to prove it, but I will try to give a motivation.

[2] Independent, identically distributed

[3] In most practical applications F will be continuous everywhere.

Theorem 3 (Central limit theorem (CLT)) *Let $X_1, X_2, \ldots$ be a sequence of iid*[4] *$L^2(\Omega, \mathcal{F}, \mathbb{P})$ random variables with $\mathbb{E}(X_i) = \mu$ and $\operatorname{Var}(X_i) = \sigma^2 > 0$. Then for the sequence $(S_n^*)_{n=1}^{\infty}$, where*

$$S_n^* = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma}$$

we have

$$S_n^* \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

The crucial term in CLT is the $\sqrt{n}$ in the denominator. Suppose for simplicity that $\mu = 0$ and $\sigma = 1$. Then by (the weakest form of) LLN

$$\frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow{\mathcal{L}} 0,$$

while by CLT

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \xrightarrow{\mathcal{L}} \mathcal{N}(0,1),$$

which is not a contradiction. In the second case we simply use a larger 'magnification ratio'.

What is the intuition behind $\sqrt{n}$? Let us suppose $X_i \sim Bern(1/2)$, so $X_1 + \cdots + X_n = X \sim Bin(n, 1/2)$. The pmf $p_X(k) = 2^{-n}\binom{n}{k}$ is maximized at $k = n/2$ (assuming for simplicity that n is even), and by Stirling's formula[5] we have

$$2^{-n/2}\binom{n}{n/2} \approx \frac{1}{c\sqrt{n}},$$

for some (unimportant) constant $c > 0$. Using this fact one can show that the sum of $o(\sqrt{n})$ values of p_X centred at $k = n/2$ is $o(1)$, while the sum of $\omega(\sqrt{n})$ such values is $1 - o(1)$. With more care for the constants (notice the $\sqrt{\pi}$ term in both Stirling and $\mathcal{N}(0,1)$), this is how CLT can be proved for $X_i \sim Bern(1/2)$. But the astonishing fact is that CLT holds for *any distribution* of $X_i \in L^2$.

Example 5 *Let us revisit the example from Lecture 8, where we tossed a fair coin 100 times and asked for the probability to get at least 60 'heads'. Chebyshev's inequality gave an upper bound of 12.5%. Applying CLT ("the normal approximation") with $X_i \sim Bern(1/2)$, $\mu = \mathbb{E}(X_i) = 1/2$ and $\sigma = \sigma(X_i) = 1/2$ gives*

$$\mathbb{P}\left(\sum_{i=1}^{100} X_i \geq 60\right) = \mathbb{P}\left(\sum_{i=1}^{100} (X_i - \mu) \geq 10\right) = \mathbb{P}\left(\frac{1}{\sqrt{100}} \sum_{i=1}^{100} \frac{X_i - \mu}{\sigma} \geq 2\right) \approx 1 - \Phi(2) \approx 2.5\%.$$

Here $\Phi(x)$ stands (and is a common abbreviation) for the cdf of the $\mathcal{N}(0,1)$ distribution. It does not have a closed form using elementary functions and avoiding integrals, but is well-understood numerically and can, for example, be looked up in so-called normal tables.

A good rule of thumb to remember: the probability for $B(n, 1/2)$ to be within 2 standard deviations from its mean is ca. 95%. Note that the scaling with n is not linear: if we toss the coin 10000 times instead then with probability ca. 95% we will get between 4900 and 5100 heads. It is remarkable how tightly the binomial distribution (and more generally any distribution arising as a sum of iid L^2-variables) is concentrated around its mean.

[4] Independent, identically distributed

[5] $n! = (1 + o(1))\sqrt{2\pi n}(\frac{n}{e})^n$