On the use of the law of large numbers in practice

Adrien Hardy*

July 3, 2020

Consider a real valued random variable X defined on a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ such that $\mathbb{E}|X| < \infty$. The (strong) law of large numbers (LLN) states that, if $X_1, X_2, \dots$ are i.i.d copies of X, then the empirical mean $\overline{X}_n := \frac{1}{n}\sum_{i=1}^n X_i$ converges as $n \to \infty$ to the theoretical mean $\mathbb{E}(X)$ almost surely. At least, this is what the Wikipedia page on the LLN says on the 3rd of July 2020. For a precise statement, one needs to specify on which probability space does the infinite sequence $X_1, X_2, \dots$ live. This precision seems always missing in the textbooks, although the LLN is at the center of the probabilistic modeling that is used in every domains of science where statistics shows up. We present some thoughts on this problem in this note and put forward a subtlety that is still puzzling for its author.

Back to basics: Probabilistic modeling. Imagine that I obtain data by measuring n times some real quantity, so that I have at disposal a sample $x_1, \dots, x_n \in \mathbb{R}$, and that I want to know the typical value of this quantity. Say, the water level of a river at n different times measured at a specific place where I want to construct a bridge. Since the results of these measurements may vary in a way that I can't predict perfectly, I can model this quantity by a random variable[1] X, so that each observation x_i equals to a realization $X(\omega_i)$ of X for some $\omega_i \in \Omega$. For example, one can take $\omega \in \Omega := \mathbb{R}$ and $X(\omega)$ stands for the water level at time $\omega \in \mathbb{R}$ (fix a reference time to be zero and say the time unit is a day). A priori I don't know the distribution of the random variable X; I can't even compute the theoretical/true mean $\mathbb{E}(X)$. It is tempting to approximate $\mathbb{E}(X)$ by its empirical version $\overline{x}_n := \frac{1}{n}\sum_{i=1}^n x_i = \frac{1}{n}\sum_{i=1}^n X(\omega_i)$ to which I have access to. However, imagine that I have measured the water level at times ω_i when it was heavily raining just before the measurement. The number $\overline{x}_n$ I obtain in this case will certainly not be representative of the true mean $\mathbb{E}(X)$: I need the outcomes of the experiments to be representative of all possible behaviors of the random variable X in some sense.

Hi Heidi. One way to find representative ω_i's is to use the mathematical notion of independence and the LLN. To do so one proceeds as follow, although it is usually implicitly done. First, extend the probability space $(\Omega, \mathscr{F}, \mathbb{P})$ to the product space $(\Omega, \mathscr{F}, \mathbb{P})^{\otimes n} := (\Omega^n, \mathscr{F}^{\otimes n}, \mathbb{P}^{\otimes n})$ and define for every $1 \le i \le n$ the random variable X_i on $(\Omega, \mathscr{F}, \mathbb{P})^{\otimes n}$ by $X_i(\omega) := X(\omega_i)$ for every $\omega = (\omega_1, \dots, \omega_n) \in \Omega^n$. By definition of the product measure $\mathbb{P}^{\otimes n}$ it follows that the X_i's are *independent*[2] and have the *same distribution*[3] than X.

*Univ. Lille, CNRS, Inria, UMR 8524, Laboratoire Paul Painlevé, F-59000 Lille, France.

[1] Namely, a measurable function $X : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{T})$ for some σ-algebra $\mathscr{T}$.

[2] That is, $\mathbb{P}^{\otimes n}(\cap_{i=1}^n X_i^{-1}(B_i)) = \prod_{i=1}^n \mathbb{P}^{\otimes n}(X_i^{-1}(B_i))$ for every $B_1, \dots, B_n \in \mathscr{T}$.

[3] Namely, $\mathbb{P}^{\otimes n}(X_i^{-1}(B)) = \mathbb{P}(X^{-1}(B))$ for every $B \in \mathscr{T}$, where $\mathscr{T}$ is the σ-alegra we equipped $\mathbb{R}$ with.

Since $X_i(\omega) = X(\omega_i) = x_i$ this construction means that we have *assumed* that the sample $x_1, \dots, x_n$ we have at hand is a realization of the *independent* random variables $X_1, \dots, X_n$. This should be justified in practice by designing a protocole before measurements such that the outputs x_i's of the experiments do not depend on each other, in some vague non-mathematical sense. Back to the water level example, it is likely that the height at two times close together should be correlated. One can also expect that, due to some chaotic dynamic resulting from the large number of factors on which the water level depends on, the heights may be considered independent if we wait long enough between two measurements.

Infinite sequence. Since by construction $\overline{x}_n = \frac{1}{n}\sum_{i=1}^n X_i(\omega) =: \overline{X}_n(\omega)$ for some $\omega \in \Omega^n$, we are now in position to use the LLN. Wait, almost, since we need to let $n \to \infty$ and our probability space depends on n. Thus, we need to construct a probability space $(\Omega^*, \mathscr{F}^*, \mathbb{P}^*)$ which "contains" $(\Omega^n, \mathscr{F}^{\otimes n})$ for every $n \geq 1$ and for which the restriction of the probability measure $\mathbb{P}^*$ to $(\Omega^n, \mathscr{F}^{\otimes n})$ is $\mathbb{P}^{\otimes n}$. More precisely, for every $n \geq 1$ there should exist a measurable map $\pi_n : (\Omega^*, \mathscr{F}^*) \to (\Omega^n, \mathscr{F}^{\otimes n})$ such that $\mathbb{P}^*(\pi_n^{-1}(A)) = \mathbb{P}^{\otimes n}(A)$ for all $A \in \mathscr{F}^{\otimes n}$. There are several ways to construct such a space $(\Omega^*, \mathscr{F}^*, \mathbb{P}^*)$ and we present two such constructions now.

a) The linear way: One construction, the one that I guess is implicitly done when saying "consider i.i.d copies $X_1, X_2, \dots$ of X", is to take the probability space

$$(\vec{\Omega}, \vec{\mathscr{F}}, \vec{\mathbb{P}}) := (\Omega, \mathscr{F}, \mathbb{P})^{\otimes \mathbb{N}}$$

with the mappings π_n defined by $\pi_n(\omega) := (\omega_1, \dots, \omega_n) \in \Omega^n$ when $\omega = (\omega_1, \omega_2, \dots) \in \Omega^{\mathbb{N}}$. We then define for every $i \geq 1$ the random variable $X_i : \vec{\Omega} \to \mathbb{R}$ by $X_i(\omega) := X(\omega_i)$, so that $\overline{X}_n := \frac{1}{n}\sum_{i=1}^n X_i$ is a random variable on the space $\vec{\Omega}$ *for every* $n \geq 1$. Now, the LLN precisely states that for $\vec{\mathbb{P}}$-almost every $\omega \in \vec{\Omega}$, we have

$$\lim_{n\to\infty} \overline{X}_n(\omega) = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n X(\omega_i) = \mathbb{E}(X).$$

If you want to illustrate this result on a computer, then sample $N \geq 1$ times independently your favorite random variable X satisfying $\mathbb{E}|X| < \infty$ so as to obtain $x_1, \dots, x_N \in \mathbb{R}$. Then compute $\overline{x}_n = \frac{1}{n}\sum_{i=1}^n x_i$ for every $n \in \{1, \dots, N\}$ and plot $\overline{x}_n$ as a function of $1 \leq n \leq N$; see the left hand side of Figure 1 below.

b) The triangular way: Next, consider instead the probability space

$$(\Omega^{\triangle}, \mathscr{F}^{\triangle}, \mathbb{P}^{\triangle}) := \bigotimes_{n=1}^{\infty} (\Omega, \mathscr{F}, \mathbb{P})^{\otimes n},$$

so that $\Omega^{\triangle} = \Omega \times (\Omega \times \Omega) \times (\Omega \times \Omega \times \Omega) \times \cdots$, and the mappings $\pi_n(\omega) := (\omega_1^{(n)}, \dots, \omega_n^{(n)})$ where $\omega = (\omega_i^{(n)} : \ 1 \leq i \leq n, \ n \geq 1) \in \Omega^{\triangle}$. In this construction, we identified in our mind the probability space $(\Omega, \mathscr{F}, \mathbb{P})^{\otimes n}$ with the n-th layer of $(\Omega^{\triangle}, \mathscr{F}^{\triangle}, \mathbb{P}^{\triangle})$. We then define the random variables $X_i^{(n)} : \Omega^{\triangle} \to \mathbb{R}$ by $X_i^{(n)}(\omega) := X(\omega_i^{(n)})$ and the empirical mean by $\overline{X}_n := \frac{1}{n}\sum_{i=1}^n X_i^{(n)}$. At the simulation level, the construction differs from the linear

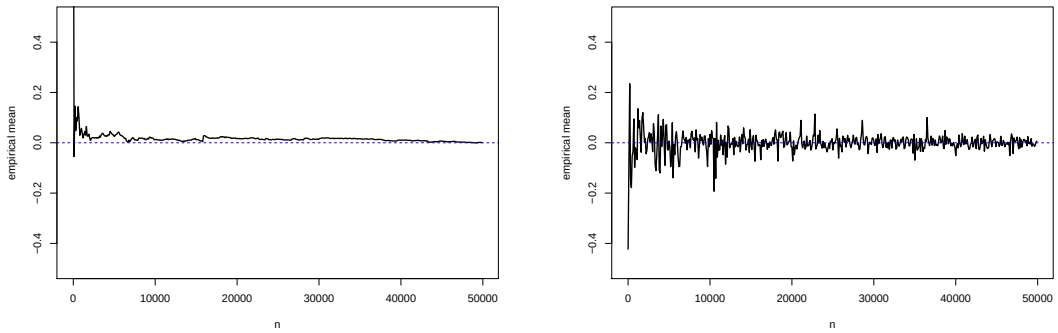

Figure 1: Realizations of the empirical mean $\overline{x}_n$ as a function of n for Student random variables of parameter 2 and $N = 50000$. The left figure is made the linear way while the right one is made the triangular way.

way since you now sample independently $x_1, \ldots, x_n$, store $\overline{x}_n$, and redo this operation *independently* for n ranging from 1 to N; see the right hand side of Figure 1.

The main difference with the linear way is that the variables $(\overline{X}_n)_{n\geq 1}$ are now independent. The two Borel-Cantelli lemmas combined together provide that $\overline{X}_n \to \mathbb{E}(X)$ as $n \to \infty$ $\mathbb{P}^{\triangle}$-almost surely if and only if for every $\varepsilon > 0$ we have

$$\sum_{n=1}^{\infty} \mathbb{P}^{\otimes n}\Big(|\overline{X}_n - \mathbb{E}(X)| > \varepsilon\Big) < \infty. \tag{C}$$

With a slight abuse of notations, we have used that $\mathbb{P}^{\triangle}(A_{n,\varepsilon}) = \mathbb{P}^{\otimes n}(A_{n,\varepsilon})$ since the event $A_{n,\varepsilon} := \{|\overline{X}_n - \mathbb{E}(X)| > \varepsilon\}$ satisfies $\pi_n(A_{n,\varepsilon}) = A_{n,\varepsilon}$. Now the condition (C) is known to be equivalent to the condition $\mathbb{E}(X^2) < \infty$; this is a result of Hsu & Robbins [1947] for one implication and Erdös [1949] for the converse implication.

Thus, when $\mathbb{E}(X) < \infty$ but $\mathbb{E}(X^2) = \infty$, for instance when X follows a Student t-distribution of parameter $1 < k \leq 2$, the LLN *does not hold true* when one embeds the spaces $(\Omega, \mathscr{F}, \mathbb{P})^{\otimes n}$ in the triangular way; this is a typical phenomenom for many other heavy tailed random variables. We illustrate this situation with simulations when X is a Student random variable of parameter 2 in Figure 1.

Summary. We have seen that, in order to use the (strong) LLN to justify that $\overline{x}_n$ is a reasonable approximation of $\mathbb{E}(X)$, it is important to take care of the way we gathered all the spaces $(\Omega, \mathscr{F}, \mathbb{P})^{\otimes n}$ into a large one. But this leads to the following puzzling question: Imagine the variable X of interest is such that $\mathbb{E}|X| < \infty$ and $\mathbb{E}(X^2) = \infty$. If one obtains *one* sample $x_1, \ldots, x_n$ from experiments, then can we just decide that it comes from a linear construction instead of a triangular one? The LLN thought in the linear way states that $\overline{x}_n \to \mathbb{E}(X)$ a.s. as $n \to \infty$, but if we think the triangular way then $\overline{x}_n \nrightarrow \mathbb{E}(X)$ a.s. as $n \to \infty$. Obviously when we obtain (finite) data there is no way to know which way applies; this is a matter of belief (or modelisation). And what if we have two independent samples $x_1, \ldots, x_n$ and $x'_1, \ldots, x'_m$? Depending on the way we have gathered these spaces,

the LLN holds true or does not.

Note however that, if $\mathbb{E}(X^2) < \infty$, then (C) holds true and the first Borel-Cantelli lemma yields that the LLN holds true whatever the way you construct the big space $(\Omega^*, \mathscr{F}^*, \mathbb{P}^*)$, so that the aforementioned problem is solved anyway. But if $\mathbb{E}(X^{2p}) = \infty$ for some $p \geq 1$ then approximating $\mathbb{E}(X^p)$ by empirical means leads to the same problematic dichotomy.

The other ways? Both linear and triangular ways are in fact two examples of projective limits of probability spaces. As a matter of fact, the existence of a space $(\Omega^*, \mathscr{F}^*, \mathbb{P}^*)$ is equivalent to the existence of a family of measurable mappings $\pi_{n,m} : (\Omega, \mathscr{F}, \mathbb{P})^{\otimes m} \to (\Omega, \mathscr{F}, \mathbb{P})^{\otimes n}$ satisfying for every $n \leq m \leq p$ the compatibility relations $\pi_{n,m} \circ \pi_{m,p} = \pi_{n,p}$ and $\pi_{m,m} = \mathrm{id}$ and furthermore $\mathbb{P}^{\otimes m} \circ \pi_{n,m}^{-1} = \mathbb{P}^{\otimes n}$; and there are many ways to do so. The triangular way is the worst case of such projective limits in the sense that one can always embed a projective limit as a restriction of this one. But which projective limit leads to a good old (strong) LLN when $\mathbb{E}|X| < \infty$ but $\mathbb{E}|X|^{1+\delta} = \infty$ for some $0 < \delta < 1$?

One way to solve this conceptual problem. After all, one may decide that the notion of convergence of random variables in the almost sure sense is a degenerated notion of convergence, at least from an application perspective. Another hint from the theoretical side is that the almost sure convergence is not metrizable, and thus in a sense not a natural notion of convergence. We may then come back to the weaker notion of convergence in probability, since it has the advantage that it does not require to embed all the random variables in the same probability space and thus make the previous puzzling (at least to me) discussion unproblematic. In particular concentration inequalities become even more attractive for applications. This would however rules out classical notions of statistics such as the one of consistent estimators.

Beyond independence. To make a good approximation of $\mathbb{E}(X)$ by its empirical mean $\overline{x}_n$, we used the assumption of *independence* to justify that the data sample is sufficiently representative of all possible behaviors of the random variable X; the famous theorem known as the LLN, when it holds, provides the theoretical landmark to justify that claim. There are however other notions of *being sufficiently representative* as one can find in dynamical systems, such as Markov chains or stationary time series, where the theoretical guaranties are provided by ergodic theorem(s), see any elementary textbook on that subjet. More recently, it has been understood that one can also find strongly correlated random variables that explore even better the typical states of a random phenomenon, such as determinantal point processes, see Bardenet & Hardy [2020].

Acknowledgments. The author has annoyed several people from l'équipe de Probabilité et Statistiques de l'Université de Lille and also his friend Djalil Chafaï with this problem and this note benefited from their precious insight.

References

R. Bardenet and A. Hardy. Monte Carlo with determinantal point processes. *Annals of Applied Probability.* Volume 30, Number 1 (2020), 368-417.

P. Erdös. On a theorem of Hsu and Robbins. *The Annals of Mathematical Statistics.* 286–291 (1949).

P. L. Hsu and H. Robbins. Complete convergence and the law of large numbers. *Proceedings of the National Academy of Sciences of the United States of America*, 33(2), 25 (1947).