A proof of the strong law of large numbers

Alex Fu

2023-01-02

Theorem 1 (Strong law of large numbers, or SLLN)**.**

Let $X_1, X_2, \ldots$ be independent and identically distributed (i.i.d.) random variables with finite mean μ. Then the sample mean $\frac{1}{n}S_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ converges to μ almost surely.

The SLLN is one of the classical limit theorems in probability, being named the *strong* law for the strength of its statement, as it asserts that the sample mean essentially always converges to the true mean. More precisely, the set of outcomes on which $\frac{1}{n}S_n \to \mu$ has probability 1.

For such a strong general statement, the hypotheses of the SLLN seem very minimal in comparison. As such, proving the SLLN is a great lesson in **managing complexity**. We will first prove simpler cases of the SLLN by introducing simplifying assumptions, then gradually work our way back up to full generality by reducing to the simpler cases.

This way of managing complexity is in fact a familiar story in probability theory. When constructing the Lebesgue measure m on $\mathbb{R}$, we first introduced the definition $m([a,b)) := b-a$ on a smaller semialgebra, then extended this to an algebra and later a σ-algebra, making sure σ-additivity was preserved each step of the way. Likewise, we first defined the Lebesgue integral for a simpler class of random variables, linear combinations of indicators of events, and step by step extended the definition to nonnegative and signed random variables.

Here, a random variable with finite first moment can be a vastly complex object in general. Fortunately, we have a few tools in our probabilistic toolkit to manage this complexity:

- Bounded random variables are easier to work with than general ones, in part because bounded r.v.s have all of their moments. We can **truncate** a given random variable X to get a bounded version, $\bar{X} = X \cdot \mathbb{1}_{|X| \le M}$, which will moreover tend to X as $M \to \infty$.
- Zero-mean or centered random variables are easier to work with, in part thanks to the convenience of not keeping track of a constant. We can **center** our r.v.s by working with $X_1 - \mu, X_2 - \mu, \ldots$, which remain independent. This allows us to assume $\mu = 0$ without loss of generality!

- Nonnegative random variables are easier to work with, for reasons we will soon see. Like we did in Lebesgue integration, we can reduce from general random variables to nonnegative ones by signed **decomposition**, writing $X = X^+ - X^-$, where $X^+ \coloneqq \max\{X, 0\}$ and $X^- \coloneqq -\min\{X, 0\}$ are both nonnegative.

Now, we start our proof of the SLLN by recalling a result used to prove the weak law of large numbers, which had a weaker statement about convergence in probability.

Lemma 1 (Markov's inequality).

Let X be a nonnegative random variable, and let $a > 0$. Then

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(X)}{a}.$$

Proof. By the linearity and monotonicity of expectation,

$$\mathbb{E}(X) = \mathbb{E}(X \cdot \mathbb{1}_{X \geq a}) + \mathbb{E}(X \cdot \mathbb{1}_{X < a}) \geq \mathbb{E}(X \cdot \mathbb{1}_{X \geq a}) \geq a \cdot \mathbb{E}(\mathbb{1}_{X \geq a}) = a \cdot \mathbb{P}(X \geq a).$$

□

Corollary 1 (Chebyshev's inequality).

Let X be any random variable with finite second moment, and let $a > 0$. Then

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq a) \leq \frac{\mathsf{var}(X)}{a^2}.$$

Proof. Apply Markov's inequality to the nonnegative $Y = |X - \mathbb{E}(X)|^2$, where $\mathbb{E}(Y) = \mathsf{var}(X)$. □

Corollary 2 (Generalized Markov's inequality).

With the same assumptions as in Lemma 1, let $\phi\colon \mathbb{R}^+ \to \mathbb{R}^+$ be any monotonic function. Then

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(\phi(X))}{\phi(a)}.$$

Proof. By the definition of monotonicity, $\{X \geq a\}$ and $\{\phi(X) \geq \phi(a)\}$ are the same event. Then apply Markov's inequality for $\phi(X) \geq 0$ and $\phi(a) > 0$. □

Chebyshev's inequality was useful for the weak law of large numbers by providing an upper bound on the probability of deviating from the mean, which is precisely convergence in probability if the upper bound tends to zero. For the SLLN, we need to use one of the few results about almost sure convergence that we know of:

Lemma 2 (Borel–Cantelli lemma)**.**

Let $A_1, A_2, \ldots$ be any sequence of events. If $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$, then $\mathbb{P}(A_n \text{ i.o.}) = 0$.

$$\{A_n \text{ i.o.}\} = \{A_n \text{ occurs infinitely often}\} = \limsup_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} \bigcup_{k\geq n} A_k$$

is the set of outcomes in which infinitely many of $A_1, A_2, \ldots$ occur.

Proof. Let $B_n \coloneqq \bigcup_{k\geq n} A_k$, so that $\{A_n \text{ i.o.}\} = \bigcap_{n=1}^{\infty} B_n$ is the limiting set of the decreasing sequence $B_1 \supseteq B_2 \supseteq \cdots$. By continuity from above, $\mathbb{P}(A_n \text{ i.o.}) = \lim_{n\to\infty} \mathbb{P}(B_n)$. By the union bound,

$$\mathbb{P}(B_n) = \mathbb{P}\left(\bigcup_{k\geq n} A_k\right) \leq \sum_{k\geq n} \mathbb{P}(A_k) \to 0$$

as $n \to \infty$, as $\sum_{k\geq n} \mathbb{P}(A_k)$ is the tail of the convergent series $\sum_{k\geq 1} \mathbb{P}(A_k) < \infty$. □

The Borel–Cantelli lemma is an example of another useful philosophy in general, namely converting more *qualitative* statements to more provable *quantitative* statements, for instance, converting "$\{A_n \text{ i.o.}\}$ is an almost sure set" to "$\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$."

After reviewing our probabilistic toolkit, we are finally ready to tackle a preliminary version of the SLLN.

Lemma 3 (SLLN with fourth moment assumption)**.**

Let $X_1, X_2, \ldots$ be i.i.d. random variables with common mean μ and finite fourth moment, $\mathbb{E}\,|X_1|^4 < \infty$. Then the sample mean $\frac{1}{n}S_n$ converges to μ almost surely.

Proof. Suppose $\mu = 0$ without loss of generality. By the definition of a limit, we observe that

$$\left\{\frac{S_n}{n} \text{ converges to } 0\right\} = \bigcap_{k=1}^{\infty} \left\{\left|\frac{S_n}{n}\right| < \frac{1}{k} \text{ eventually}\right\} = \bigcap_{k=1}^{\infty} \left\{\left|\frac{S_n}{n}\right| \geq \frac{1}{k} \text{ only finitely often}\right\}.$$

To show that the probability of this intersection is 1, it suffices to show that $\mathbb{P}(|\frac{S_n}{n}| \geq \varepsilon \text{ finitely often}) = 1$ for some arbitrary $\varepsilon = \frac{1}{k}$, or equivalently $\mathbb{P}(|\frac{S_n}{n}| \geq \varepsilon \text{ i.o.}) = 0$. (This argument is a common application of the Borel–Cantelli lemma.)

Per Lemma 2, it is enough to have $\sum_{n=1}^{\infty} \mathbb{P}(|\frac{S_n}{n}| \geq \varepsilon) < \infty$. Invoking the generalized Markov's inequality for the nonnegative random variable $|\frac{S_n}{n}|$ and the monotonic function $\phi\colon x \mapsto x^4$,

$$\sum_{n=1}^{\infty} \mathbb{P}\left(\left|\frac{S_n}{n}\right| \geq \varepsilon\right) \leq \frac{1}{\varepsilon^4} \sum_{n=1}^{\infty} \frac{\mathbb{E}(S_n^4)}{n^4}.$$

Now, we see that $\mathbb{E}(S_n^4) \in O(n^2)$ asymptotically: by the linearity of expectation,

$$\begin{aligned}\mathbb{E}\big((X_1 + \cdots + X_n)^4\big) &= \sum_{i=1}^{n} \mathbb{E}\big(X_i^4\big) + \binom{4}{2} \sum_{i<j} \mathbb{E}\big(X_i^2 X_j^2\big) \\ &= n\,\mathbb{E}\big(X_1^4\big) + 3n(n-1)\,\mathbb{E}\big(X_1^2 X_2^2\big).\end{aligned}$$

The other terms in the expanded sum, which have the form $X_i X_j^3$, $X_i X_j X_k^2$, and $X_i X_j X_k X_\ell$, all have zero expected value, because $X_1, X_2, \ldots$ are zero-mean and independent, and these terms all contain an r.v. with exponent 1. (By independence, $\mathbb{E}(X_i^1 \cdots) = \mathbb{E}(X_i^1)\,\mathbb{E}(\cdots) = 0$.) Therefore

$$\sum_{n=1}^{\infty} \mathbb{P}\left(\left|\frac{S_n}{n}\right| \geq \varepsilon\right) \leq \frac{1}{\varepsilon^4} \sum_{n=1}^{\infty} \frac{Cn^2}{n^4} < \infty,$$

and we are done per the previous remarks. □

Why the fourth moment?

- The invocation of the Borel–Cantelli lemma is fairly natural; the lemma is one of the few general results whose conclusion is that an event is almost sure, which is also the conclusion of the SLLN. Moreover, its statement about events occurring infinitely often makes it uniquely suited to convert a qualitative statement of *convergence* into a more quantitative one of *summability*.
- With the new goal of showing that a numerical series $\sum_{n=1}^{\infty} \mathbb{P}(A_n)$ is finite, Markov's inequality is a natural choice here, being one of the most common upper bounds on a probability (of deviation). This also turns a question of summability into an even finer question on asymptotic rate of decay.
- The choice of the monotonic function $\phi(x) = x^k$ is also quite natural: by the linearity of expectation and independence, $\mathbb{E}(S_n^k)$ expands into a sum of products of moments of X_i, which gives us control through our finite moment assumptions. Still, the question remains: why $k = 4$?

The key is that the Markov bound depends on $\mathbb{E}(|S_n|^k)$, not $\mathbb{E}(S_n^k)$, at least for odd k. If k is even, then $\mathbb{E}(|S_n|^k) = \mathbb{E}(S_n^k)$ allows for expanding as above. But, otherwise, we only have a rough bound given by the triangle inequality: for example, when $k = 3$,

$$\mathbb{E}|S_n|^3 \leq \sum_{i=1}^{n} \mathbb{E}|X_i|^3 + 3\sum_{i<j} \mathbb{E}|X_i|^2\,\mathbb{E}|X_j| + C\sum_{i<j<k} \mathbb{E}|X_i|\,\mathbb{E}|X_j|\,\mathbb{E}|X_k| \in \Theta(n^3).$$

In the process, we also lose the advantage of the X_i being zero-mean and independent. The SLLN is, in some sense, a statement about *cancellation*: i.i.d. samples fluctuate about their mean, sometimes below, sometimes above, but on average these fluctuations will cancel out. Taking away the sign in $\mathbb{E}|X_i|$, then, is undesirable. Finally, we take $k = 4$ because the first even k does not succeed:

$$\mathbb{E}|S_n|^2 = \mathbb{E}(S_n^2) = \sum_{i=1}^{n} \mathbb{E}(X_i^2) + 2\sum_{i<j} \mathbb{E}(X_i)\,\mathbb{E}(X_j) = n\,\mathbb{E}(X_1^2) \in \Theta(n),$$

but $n^{-2}\,\mathbb{E}(S_n^2) \in \Theta(\frac{1}{n})$ is not summable for the purposes of Borel–Cantelli. $\Theta(\frac{1}{n^2})$, however, is.

Lemma 3 is a great start, but we can upgrade to a statement with even weaker assumptions by considering a "fast" convergent **subsequence**.

Lemma 4 (SLLN with second moment assumption)**.**

The SLLN (Theorem 1) holds under the additional assumption that the random variables have finite second moment, $\mathbb{E}(|X_1|^2) < \infty$.

Proof. The argument is largely the same as above, but the details of justifying Borel–Cantelli will differ. We lose access to generalized Markov's inequality with $\phi(x) = x^4$, but we still have Chebyshev's inequality by the second moment assumption. Let k_n be a sequence such that $\sum_{n=1}^{\infty} \frac{1}{k_n} < \infty$. Then

$$\sum_{n=1}^{\infty} \mathbb{P}\left(\left|\frac{S_{k_n}}{k_n}\right| \geq \varepsilon\right) \leq \frac{1}{\varepsilon^2} \sum_{n=1}^{\infty} \frac{\mathbb{E}(S_{k_n}^2)}{k_n^2} = \frac{1}{\varepsilon^2} \sum_{n=1}^{\infty} \frac{k_n \, \mathbb{E}(X_1^2)}{k_n^2} = C \sum_{n=1}^{\infty} \frac{1}{k_n} < \infty$$

by Chebyshev's inequality, which shows that the subsequence $(\frac{1}{k_n} S_{k_n})_{n=1}^{\infty}$ converges to 0 almost surely. The key now is to bound the **oscillations** of the sequence between consecutive points in the subsequence. Combined with the fact that $\frac{1}{k_n} S_{k_n} \to 0$, we claim that if

$$\frac{1}{k_n} \Delta_n \coloneqq \sup_{k_n \leq \ell \leq k_{n+1}} \frac{|S_\ell - S_{k_n}|}{k_n} \to 0,$$

then the full sequence $\frac{1}{n} S_n \to 0$ almost surely as well. By the triangle inequality,

$$\frac{|S_\ell|}{\ell} \leq \frac{|S_{k_n}| + \Delta_n}{\ell} = \frac{k_n}{\ell} \cdot \frac{|S_{k_n}| + \Delta_n}{k_n}$$

for all $k_n \leq \ell \leq k_{n+1}$. As $n \to \infty$, both $\frac{1}{k_n} \Delta_n$ and $\frac{1}{k_n} |S_{k_n}|$ converge to 0, and $\frac{k_n}{\ell}$ is bounded above by 1, which shows that $\frac{1}{\ell} |S_\ell| \to 0$ almost surely as well.

Lastly, let us show the hypothesis of the claim. By the union bound, then Chebyshev's inequality,

$$\mathbb{P}\left(\left|\frac{\Delta_n}{k_n}\right| \geq \varepsilon\right) \leq \sum_{\ell=k_n}^{k_{n+1}} \mathbb{P}\left(\frac{|S_\ell - S_{k_n}|}{k_n} \geq \varepsilon\right) \leq \frac{1}{\varepsilon^2} \sum_{\ell=k_n}^{k_{n+1}} \frac{(\ell - k_n) \, \mathbb{E}(X_1^2)}{k_n^2}.$$

The rest is simply algebra. For convenience, let us choose $k_n = n^2$. Then

$$\cdots \leq C \sum_{\ell=k_n}^{k_{n+1}} \frac{k_{n+1} - k_n}{k_n^2} = C \left(\frac{k_{n+1} - k_n}{k_n}\right)^2 = C \left(\frac{2n+1}{n^2}\right)^2$$

is summable, which shows that $\frac{1}{k_n} \Delta_n \to 0$ almost surely. □

Before we move onto a proof of the SLLN with only the assumption of a finite first moment, we will add one last quantitative result to our toolkit.

Lemma 5 (Tail-sum approximation).

Let X be a nonnegative random variable, and let $p \geq 1$. Then, asymptotically in p,

$$\sum_{k=1}^{\infty} k^{p-1}\,\mathbb{P}(X \geq k) \lesssim \mathbb{E}(X^p) \lesssim 1 + \sum_{k=1}^{\infty} k^{p-1}\,\mathbb{P}(X \geq k).$$

Proof. Omitted. See lower and upper Riemann sums for the Riemann integral. □

The following proof will require all of the tools and techniques that we have seen so far, culminating in a final demonstration of managing complexity and transforming qualitative statements into quantitative statements, two of the key takeaways here.

Proof of Theorem 1. As before, assume that $X_1, X_2, \ldots$ are zero-mean without loss of generality.

- If the SLLN holds for $X_1^+, X_2^+, \ldots$ and $X_1^-, X_2^-, \ldots$, then $\frac{1}{n}S_n = \frac{1}{n}S_n^+ - \frac{1}{n}S_n^- \to 0$ almost surely as well. As such, we can assume nonnegativity without loss of generality, e.g. taking $X_1^+, X_2^+, \ldots$. However, we lose zero-meanness by doing so: we now want to show that $\frac{1}{n}S_n \to \mu = \mathbb{E}(X_1)$.

- In order to invoke Lemma 4, we work with the truncated random variables $Y_i = X_i \cdot \mathbb{1}_{X_i \leq i}$, writing $\bar{S}_n = \sum_{i=1}^n Y_i$. It now suffices to prove $\frac{1}{n}\bar{S}_n \to \mu$. By Lemma 2 and Lemma 5, $X_i = Y_i$ eventually almost surely, i.e. $\mathbb{P}(X_i \neq Y_i \text{ i.o.}) = 0$:

  $$\sum_{i=1}^{\infty} \mathbb{P}(X_i \neq Y_i) = \sum_{i=1}^{\infty} \mathbb{P}(X_i > i) \approx \mathbb{E}(X_1) < \infty.$$

  In other words, $\frac{1}{n}|S_n - \bar{S}_n| \to 0$ almost surely, since the sums will only differ in finitely many terms. Then $\frac{1}{n}\bar{S}_n \to \mu$ implies $\frac{1}{n}S_n \to \mu$ by the triangle inequality.

- By Lemma 4 for $Y_1, Y_2, \ldots$, given a sequence $(k_n)_{n=1}^{\infty}$ along which

  $$\sum_{n=1}^{\infty} \frac{\mathsf{var}(\bar{S}_{k_n})}{k_n^2} < \infty,$$

  we have $\frac{1}{k_n}|\bar{S}_{k_n} - \mathbb{E}(\bar{S}_{k_n})| \to 0$ almost surely. The truncated mean $\frac{1}{k_n}\,\mathbb{E}(\bar{S}_{k_n}) = \mathbb{E}(Y_1)$ might not equal μ, but by the dominated convergence theorem, $\mathbb{E}(Y_i) \to \mathbb{E}(X_1) = \mu$. As such,

  $$\frac{\mathbb{E}(\bar{S}_{k_n})}{k_n} = \frac{1}{k_n}\sum_{i=1}^{k_n} \mathbb{E}(Y_i) \to \mu.$$

  The running average of a convergent sequence converges to the same limit. From here, $\frac{1}{k_n}\bar{S}_{k_n} \to \mu$ almost surely as well.

- We have yet to show that $(k_n)_{n=1}^{\infty}$ satisfying the summability condition above exists. Let us write

  $$\sum_{n=1}^{\infty} \frac{1}{k_n^2}\,\mathsf{var}(\bar{S}_{k_n}) \leq \sum_{n=1}^{\infty}\left[\frac{1}{k_n^2}\sum_{i=1}^{k_n}\mathbb{E}\big(Y_i^2\big)\right] \approx \sum_{n=1}^{\infty}\sum_{i=1}^{k_n}\sum_{j=1}^{i}\frac{j\,\mathbb{P}(X_1 > j)}{k_n^2}$$

by Lemma 5. The innermost sum ranges from $j = 1$ to i as Y_i is X_i truncated at i; for $1 \le j \le i$, $\mathbb{P}(Y_i > j) = \mathbb{P}(X_i > j) = \mathbb{P}(X_1 > j)$. Now, the term $j\,\mathbb{P}(X_1 > j)$ appears once for each $i \ge j$, i.e. $k_n - j + 1$ times for each $k_n \ge j$. Exchanging the summations by Tonelli's theorem,

$$\cdots = \sum_{j=1}^{\infty} \left[j\,\mathbb{P}(X_1 > j) \sum_{n:k_n \ge j} \frac{k_n - j + 1}{k_n^2} \right] \le \sum_{j=1}^{\infty} \left[j\,\mathbb{P}(X_1 > j) \sum_{n:k_n \ge j} \frac{1}{k_n} \right].$$

If we choose $k_n = \lceil \alpha^n \rceil$ for $\alpha > 1$, then $\sum_{n:k_n \ge j} \frac{1}{k_n} \approx \frac{1}{j}$, and the above is $\approx \sum_{j=1}^{\infty} \mathbb{P}(X_1 > j) \approx \mathbb{E}(X_1) < \infty$.

- Finally, let us upgrade from the subsequential convergence of $\frac{1}{k_n}\bar{S}_{k_n} \to \mu$ to $\frac{1}{n}\bar{S}_n \to \mu$ along the full sequence. The proof is much easier than the one in the proof of Lemma 4: by the nonnegativity of the Y_i, the sums $\bar{S}_n$ form a monotone sequence, which allows us to construct the sandwich

$$\frac{\bar{S}_{k_n}}{k_n} \cdot \frac{k_n}{k_{n+1}} = \frac{\bar{S}_{k_n}}{k_{n+1}} \le \frac{\bar{S}_\ell}{\ell} \le \frac{\bar{S}_{k_{n+1}}}{k_n} = \frac{\bar{S}_{k_{n+1}}}{k_{n+1}} \cdot \frac{k_{n+1}}{k_n}$$

for $k_n \le \ell \le k_{n+1}$. Taking the limit, we find that

$$\mu \cdot \alpha^{-1} \le \liminf_{\ell \to \infty} \frac{\bar{S}_\ell}{\ell} \le \limsup_{\ell \to \infty} \frac{\bar{S}_\ell}{\ell} \le \mu \cdot \alpha.$$

Sending $\alpha \to 1^+$, we are done.

□

In summary, we reduced the strong law of large numbers for general random variables with first moment to the simpler case of convergence along a subsequence of nonnegative bounded random variables, which we proved using the Borel–Cantelli lemma, Chebyshev's inequality, and the tail-sum approximation. Then, we upgraded to convergence along the full sequence by using monotonicity to control oscillations, lifted the boundedness condition by showing that truncation has zero effect in the limit, and found the conclusion for signed random variables by putting together the nonnegative parts.

This is far from the end of the journey. It is a straightforward exercise to extend the SLLN to integrable random variables, where $\mathbb{E}(X_1)$ may be $\pm\infty$, and to the more general case of $\frac{1}{N}S_N \to \mu$, where N is a r.v. itself tending to ∞ almost surely. We can also gain a more quantitative understanding of the sample sum: $n^{-(0.5+\varepsilon)}S_n \to 0$; $n^{-1/2}\log(n)^{-c}S_n \to 0$; and $\limsup(2n\log\log n)^{-1/2}S_n = 1$ almost surely give far more precise descriptions of the asymptotic rate of S_n. And, various *Central Limit Theorems* (CLTs) characterize the exact distribution of $n^{-1/2}S_n$ or similar, a topic we will leave for another time.

This note was adapted largely from Professor Shirshendu Ganguly's fall 2022 offering of Math C218A / Stat C205A at UC Berkeley.

■