2.10. Strong law of large numbers

If X_n are i.i.d with finite mean, then the weak law asserts that $n^{-1}S_n \xrightarrow{P} \mathbf{E}[X_1]$. The strong law strengthens it to almost sure convergence.

Theorem 2.36 (Kolmogorov's SLLN). *Let X_n be i.i.d with $\mathbf{E}[|X_1|] < \infty$. Then, as $n \to \infty$, we have $\frac{S_n}{n} \xrightarrow{a.s.} \mathbf{E}[X_1]$.*

The proof of this theorem is somewhat complicated. First of all, we should ask if WLLN implies SLLN? From Lemma 2.27 we see that this can be done if $\mathbf{P}\left(|n^{-1}S_n - \mathbf{E}[X_1]| > \delta\right)$ is summable, for every $\delta > 0$. Even assuming finite variance $\mathrm{Var}(X_1) = \sigma^2$, Chebyshev's inequality only gives a bound of $\sigma^2\delta^{-2}n^{-1}$ for this probability and this is not summable. Since this is at the borderline of summability, if we assume that p^{th} moment exists for some $p > 2$, we may expect to carry out this proof. Suppose we assume that $\alpha_4 := \mathbf{E}[X_1^4] < \infty$ (of course 4 is not the smallest number bigger than 2, but how do we compute $\mathbf{E}[|S_n|^p]$ in terms of moments of X_1 unless p is an even integer?). Then, we may compute that (assume $\mathbf{E}[X_1] = 0$ wlog)

$$\mathbf{E}\left[S_n^4\right] = n^2(n-1)^2\sigma^4 + n\alpha_4 = O(n^2).$$

Thus $\mathbf{P}\left(|n^{-1}S_n| > \delta\right) \le n^{-4}\delta^{-4}\mathbf{E}[S_n^4] = O(n^{-2})$ which is summable, and by Lemma 2.27 we get the following weaker form of SLLN.

Theorem 2.37. *Let X_n be i.i.d with $\mathbf{E}[|X_1|^4] < \infty$. Then, $\frac{S_n}{n} \xrightarrow{a.s.} \mathbf{E}[X_1]$ as $n \to \infty$.*

Now we return to the serious question of proving the strong law under first moment assumptions. The presentation of the following proof is adapted from a blog article of Terence Tao.

PROOF. **Step 1:** It suffices to prove the theorem for integrable non-negative r.v, because we may write $X = X_+ - X_-$ and note that $S_n = S_n^+ - S_n^-$. (Caution: Don't also assume zero mean in addition to non-negativity!). Henceforth, we assume that $X_n \ge 0$ and $\mu = \mathbf{E}[X_1] < \infty$. One consequence is that

$$\frac{S_{N_1}}{N_2} \le \frac{S_n}{n} \le \frac{S_{N_2}}{N_1} \quad \text{if } N_1 \le n \le N_2. \tag{2.10}$$

Step 2: The second step is to prove the following claim. To understand the big picture of the proof, you may jump to the third step where the strong law is deduced using this claim, and then return to the proof of the claim.

Claim 2.38. *Fix any $\lambda > 1$ and define $n_k := \lfloor \lambda^k \rfloor$. Then, $\frac{S_{n_k}}{n_k} \xrightarrow{a.s.} \mathbf{E}[X_1]$ as $k \to \infty$.*

Proof of the claim Fix j and for $1 \le k \le n_j$ write $X_k = Y_k + Z_k$ where $Y_k = X_k \mathbf{1}_{X_k \le n_j}$ and $Z_k = X_k \mathbf{1}_{X_k > n_j}$ (why we chose the truncation at n_j is not clear at this point). Then, let J_δ be large enough so that for $j \ge J_\delta$, we have $\mathbf{E}[Z_1] \le \delta$. Let $S_{n_j}^Y = \sum_{k=1}^{n_j} Y_k$ and $S_{n_j}^Z = \sum_{k=1}^{n_j} Z_k$. Since $S_{n_j} = S_{n_j}^Y + S_{n_j}^Z$ and $\mathbf{E}[X_1] = \mathbf{E}[Y_1] + \mathbf{E}[Z_1]$, we get

$$\begin{aligned}
\mathbf{P}\left(\left|\frac{S_{n_j}}{n_j} - \mathbf{E}[X_1]\right| > 2\delta\right) &\le \mathbf{P}\left(\left|\frac{S_{n_j}^Y}{n_j} - \mathbf{E}[Y_1]\right| + \left|\frac{S_{n_j}^Z}{n_j} - \mathbf{E}[Z_1]\right| > 2\delta\right) \\
&\le \mathbf{P}\left(\left|\frac{S_{n_j}^Y}{n_j} - \mathbf{E}[Y_1]\right| > \delta\right) + \mathbf{P}\left(\left|\frac{S_{n_j}^Z}{n_j} - \mathbf{E}[Z_1]\right| > \delta\right) \\
&\le \mathbf{P}\left(\left|\frac{S_{n_j}^Y}{n_j} - \mathbf{E}[Y_1]\right| > \delta\right) + \mathbf{P}\left(\frac{S_{n_j}^Z}{n_j} \neq 0\right). \qquad (2.11)
\end{aligned}$$

We shall show that both terms in (2.11) are summable over j. The first term can be bounded by Chebyshev's inequality

$$\mathbf{P}\left(\Big| \frac{S^Y_{n_j}}{n_j} - \mathbf{E}[Y_1] \Big| > \delta \right) \leq \frac{1}{\delta^2 n_j} \mathbf{E}[Y_1^2] = \frac{1}{\delta^2 n_j} \mathbf{E}[X_1^2 \mathbf{1}_{X_1 \leq n_j}]. \tag{2.12}$$

while the second term is bounded by the union bound

$$\mathbf{P}\left(\frac{S^Z_{n_j}}{n_j} \neq 0 \right) \leq n_j \mathbf{P}(X_1 > n_j). \tag{2.13}$$

The right hand sides of (2.12) and (2.13) are both summable. To see this, observe that for any positive x, there is a unique k such that $n_k < x \leq n_{k+1}$, and then

$$(a)\ \sum_{j=1}^{\infty} \frac{1}{n_j} x^2 \mathbf{1}_{x \leq n_j} \leq x^2 \sum_{j=k+1}^{\infty} \frac{1}{\lambda^j} \leq C_\lambda x. \qquad (b)\ \sum_{j=1}^{\infty} n_j \mathbf{1}_{x > n_j} \leq \sum_{j=1}^{k} \lambda^j \leq C_\lambda x. \tag{2.14}$$

Here, we may take $C_\lambda = \frac{\lambda}{\lambda - 1}$, but what matters is that it is some constant depending on λ (but not on x). We have glossed over the difference between $\lfloor \lambda^j \rfloor$ and λ^j but you may check that it does not matter (perhaps by replacing C_λ with a larger value). Setting $x = X_1$ in the above inequalities (a) and (b) and taking expectations, we get

$$\sum_{j=1}^{\infty} \frac{1}{n_j} \mathbf{E}[X_1^2 \mathbf{1}_{X_1 \leq n_j}] \leq C_\lambda \mathbf{E}[X_1]. \qquad \sum_{j=1}^{\infty} n_j \mathbf{P}(X_1 > n_j) \leq C_\lambda \mathbf{E}[X_1].$$

As $\mathbf{E}[X_1] < \infty$, the probabilities on the left hand side of (2.12) and (2.13) are summable in j, and hence it also follows that $\mathbf{P}\left(\Big| \frac{S_{n_j}}{n_j} - \mathbf{E}[X_1] \Big| > 2\delta \right)$ is summable. This happens for every $\delta > 0$ and hence Lemma 2.27 implies that $\frac{S_{n_j}}{n_j} \overset{a.s.}{\to} \mathbf{E}[X_1]$ a.s. This proves the claim.

Step 3: Fix $\lambda > 1$. Then, for any n, find k such that $\lambda^k < n \leq \lambda^{k+1}$, and then, from (2.10) we get

$$\frac{1}{\lambda} \mathbf{E}[X_1] \leq \liminf_{n \to \infty} \frac{S_n}{n} \leq \limsup_{n \to \infty} \frac{S_n}{n} \leq \lambda \mathbf{E}[X_1], \text{ almost surely.}$$

Take intersection of the above event over all $\lambda = 1 + \frac{1}{m}$, $m \geq 1$ to get $\lim_{n \to \infty} \frac{S_n}{n} = \mathbf{E}[X_1]$ a.s. ∎

2.11. Kolmogorov's zero-one law

We saw that in strong law the limit of $n^{-1} S_n$ turned out to be constant, while a priori, it could well have been random. This is a reflection of the following more general and surprising fact.

Definition 2.39. Let $\mathcal{F}_n$ be sub-sigma algebras of $\mathcal{F}$. Then the tail σ-algebra of the sequence $\mathcal{F}_n$ is defined to be $\mathcal{T} := \cap_n \sigma(\cup_{k \geq n} \mathcal{F}_k)$. For a sequence of random variables $X_1, X_2, \ldots$, the tail sigma algebra is the tail of the sequence $\sigma(X_n)$.

We also say that a σ-algebra is trivial (w.r.t a probability measure) if $\mathbf{P}(A)$ equals 0 or 1 for every A in the sig-algebra.

Theorem 2.40 (Kolmogorov's zero-one law). *Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space.*

(1) *If $\mathcal{F}_n$ is a sequence of independent sub-sigma algebras of $\mathcal{F}$, then the tail sig-algebra is trivial.*

(2) *If X_n are independent random variables, and A is a tail event, then* $\mathbf{P}(A)$ *is* 0 *or* 1 *for every* $A \in \mathcal{T}$.

PROOF. The second statement follows immediately from the first. To prove the first, define $\mathcal{T}_n := \sigma(\cup_{k>n}\mathcal{F}_k)$. Then, $\mathcal{F}_1,\dots,\mathcal{F}_n,\mathcal{T}_n$ are independent. Hence, $\mathcal{F}_1,\dots,\mathcal{F}_n,\mathcal{T}$ are independent. Since this is true for every n, we see that $\mathcal{T},\mathcal{F}_1,\mathcal{F}_2,\dots$ are independent. Hence, $\mathcal{T}$ and $\sigma(\cup_n\mathcal{F}_n)$ are independent. But $\mathcal{T} \subset \sigma(\cup_n\mathcal{F}_n)$, hence, $\mathcal{T}$ is independent of itself. This implies that for any $A \in \mathcal{T}$, we must have $\mathbf{P}(A)^2 = \mathbf{P}(A \cap A) = \mathbf{P}(A)$ which forces $\mathbf{P}(A)$ to be 0 or 1. ■

Exercise 2.41. Let X_i be independent random variables. Which of the following random variables must necessarily be constant almost surely? $\limsup X_n$, $\liminf X_n$, $\limsup n^{-1}S_n$, $\liminf S_n$.

An application: This application is really an excuse to introduce a beautiful object of probability. Consider the lattice $\mathbb{Z}^2$, points of which we call vertices. By an edge of this lattice we mean a pair of adjacent vertices $\{(x,y),(p,q)\}$ where $x = p, |y-q| = 1$ or $y = q, |x-p| = 1$. Let E denote the set of all edges. X_e, $e \in E$ be i.i.d Ber(p) random variables indexed by E. Consider the subset of all edges e for which $X_e = 1$. This gives a random subgraph of $\mathbb{Z}^2$ called the *bond percolation at level* p. We denote the subgraph by G_ω.t

Question: What is the probability that in the percolation subgraph, there is an infinite connected component?

Let $A = \{\omega : G_\omega \text{ has an infinite connected component}\}$. If there is an infinite component, changing X_e for finitely many e cannot destroy it. Conversely, if there was no infinite cluster to start with, changing X_e for finitely many e cannot create one. In other words, A is a tail event for the collection X_e, $e \in E$! Hence, by Kolmogorov's 0-1 law, $\mathbf{P}_p(A)$ is equal to 0 or 1. Is it 0 or is it 1?

In pathbreaking work, it was proved by 1980s that $\mathbf{P}_p(A) = 0$ if $p \le \frac{1}{2}$ and $\mathbf{P}_p(A) = 1$ if $p > \frac{1}{2}$.

The same problem can be considered on $\mathbb{Z}^3$, keeping each edge with probability p and deleting it with probability $1-p$, independently of all other edges. It is again known (and not too difficult to show) that there is some number $p_c \in (0,1)$ such that $\mathbf{P}_p(A) = 0$ if $p < p_c$ and $\mathbf{P}_p(A) = 1$ if $p > p_c$. The value of p_c is not known, and more importantly, it is not known whether $\mathbf{P}_{p_c}(A)$ is 0 or 1!

2.12. The law of iterated logarithm

If $a_n \uparrow \infty$, then the reasoning in the previous section applies and $\limsup a_n^{-1}S_n$ is constant a.s. This motivates the following natural question.

Question: Let X_i be i.i.d random variables taking values ± 1 with equal probability. Find a_n so that $\limsup \frac{S_n}{a_n} = 1$ a.s.

The question is about the growth rate of sums of random independent ± 1s. We know that $n^{-1}S_n \overset{a.s.}{\to} 0$ by the SLLN, hence, $a_n = n$ is "too much". What about n^α. Applying Hoeffding's inequality (proved in the next section), we see that $\mathbf{P}(n^{-\alpha}S_n > t) \le \exp\{-\frac{1}{2}t^2n^{2\alpha-1}\}$. If $\alpha > \frac{1}{2}$, this is a summable sequence for any $t > 0$, and therefore $\mathbf{P}(n^{-\alpha}S_n > t \text{ i.o.}) = 0$. That is $\limsup n^{-\alpha}S_n \overset{a.s.}{\to} 0$ for $\alpha > \frac{1}{2}$. What about $\alpha = \frac{1}{2}$? One can show that $\limsup n^{-\frac{1}{2}}S_n = +\infty$ a.s, which means that $\sqrt{n}$ is too slow compared to S_n. So the right answer is larger than $\sqrt{n}$ but smaller than $n^{\frac{1}{2}+\epsilon}$ for any $\epsilon > 0$. The sharp answer, due to Khinchine is a crown jewel of probability theory!

Result 2.42 (Khinchine's law of iterated logarithm). Let X_i be i.i.d with zero mean and finite variance $\sigma^2 = 1$ (without loss of generality). Then,

$$\limsup_{n\to\infty} \frac{S_n}{\sqrt{2n\log\log n}} = +1 \text{ a.s.}$$

In fact the set of all limit points of the sequence $\left\{\frac{S_n}{\sqrt{2n\log\log n}}\right\}$ is almost surely equal to the interval $[-1,1]$.

We skip the proof of LIL, because it is a bit involved, and there are cleaner ways to deduce it using Brownian motion (in this or a later course).

Exercise 2.43. Let X_i be i.i.d random variables taking values ± 1 with equal probability. Show that $\limsup_{n\to\infty} \frac{S_n}{\sqrt{2n\log\log n}} \le 1$, almost surely.

2.13. Hoeffding's inequality

If X_n are i.i.d with finite mean, then we know that the probability for S_n/n to be more than δ away from its mean, goes to zero. How fast? Assuming finite variance, we saw that this probability decays at least as fast as n^{-1}. If we assume higher moments, we can get better bounds, but always polynomial decay in n. Here we assume that X_n are bounded a.s, and show that the decay is like a Gaussian.

Lemma 2.44. (Hoeffding's inequality). *Let $X_1,\dots,X_n$ be independent, and assume that $|X_k| \le d_k$ w.p.1. For simplicity assume that $\mathbf{E}[X_k] = 0$. Then, for any $n \ge 1$ and any $t > 0$,*

$$\mathbf{P}(|S_n| \ge t) \le 2\exp\left\{-\frac{t^2}{2\sum_{i=1}^n d_i^2}\right\}.$$

Remark 2.45. The boundedness assumption on X_ks is essential. That $\mathbf{E}[X_k] = 0$ is for convenience. If we remove that assumption, note that $Y_k = X_k - \mathbf{E}[X_k]$ satisfy the assumptions of the theorem, except that we can only say that $|Y_k| \le 2d_k$ (because $|X_k| \le d_k$ implies that $|\mathbf{E}[X_k]| \le d_k$ and hence $|X_k - \mathbf{E}[X_k]| \le 2d_k$). Thus, applying the result to Y_ks, we get

$$\mathbf{P}(|S_n - \mathbf{E}[S_n]| \ge t) \le 2\exp\left\{-\frac{t^2}{8\sum_{i=1}^n d_i^2}\right\}.$$

PROOF. Without loss of generality, take $\mathbf{E}[X_k] = 0$. Now, if $|X| \le d$ w.p.1, and $\mathbf{E}[X] = 0$, by convexity of exponential on $[-1,1]$, we write for any $\lambda > 0$

$$e^{\lambda X} \le \frac{1}{2}\left(\left(1+\frac{X}{d}\right)e^{-\lambda d} + \left(1-\frac{X}{d}\right)e^{\lambda d}\right).$$

Therefore, taking expectations we get $\mathbf{E}[\exp\{\lambda X\}] \le \cosh(\lambda d)$. Take $X = X_k$, $d = d_k$ and multiply the resulting inequalities and use independence to get $\mathbf{E}[\exp\{\lambda S_n\}] \le \prod_{k=1}^n \cosh(\lambda d_k)$. Apply the elementary inequality $\cosh(x) \le \exp(x^2/2)$ to get

$$\mathbf{E}[\exp\{\lambda S_n\}] \le \exp\left\{\frac{1}{2}\lambda^2 \sum_{k=1}^n d_k^2\right\}.$$

From Markov's inequality we thus get $\mathbf{P}(S_n > t) \leq e^{-\lambda t}\mathbf{E}[e^{\lambda S_n}] \leq \exp\{-\lambda t + \frac{1}{2}\lambda^2 \sum_{k=1}^n d_k^2\}$. Optimizing this over λ gives the choice $\lambda = \frac{t}{\sum_{k=1}^n d_k^2}$ and the inequality

$$\mathbf{P}(S_n \geq t) \leq \exp\left\{-\frac{t^2}{2\sum_{i=1}^n d_i^2}\right\}.$$

Working with $-X_k$ gives a similar inequality for $\mathbf{P}(-S_n > t)$ and adding the two we get the statement in the lemma. ■

The power of Hoeffding's inequality is that it is not an asymptotic statement but valid for every finite n and finite t. Here are two consequences. Let X_i be i.i.d bounded random variables with $\mathbf{P}(|X_1| \leq d) = 1$.

(1) (**Large deviation regime**) Take $t = n\delta$ to get

$$\mathbf{P}\left(|\frac{1}{n}S_n - \mathbf{E}[X_1]| \geq u\right) = \mathbf{P}(|S_n - \mathbf{E}[S_n]| \geq u) \leq 2\exp\left\{-\frac{u^2}{8d^2}n\right\}.$$

This shows that for bounded random variables, the probability for the sample sum S_n to deviate by an order n amount from its mean decays exponentially in n. This is called the *large deviation regime* because the order of the deviation is the same as the typical order of the quantity we are measuring.

(2) (**Moderate deviation regime**) Take $t = u\sqrt{n}$ to get

$$\mathbf{P}(|S_n - \mathbf{E}[S_n]| \geq \delta) \leq 2\exp\left\{-\frac{u^2}{8d^2}\right\}.$$

This shows that S_n is within a window of size $\sqrt{n}$ centered at $\mathbf{E}[S_n]$. In this case the probability is not decaying with n, but the window we are looking at is of a smaller order namely, $\sqrt{n}$, as compared to S_n itself, which is of order n. Therefore this is known as *moderate deviation regime*. The inequality also shows that the tail probability of $(S_n - \mathbf{E}[S_n])/\sqrt{n}$ is bounded by that of a Gaussian with variance d. More generally, if we take $t = un^\alpha$ with $\alpha \in [1/2, 1)$, we get $\mathbf{P}(|S_n - \mathbf{E}[S_n]| \geq un^\alpha) \leq 2e^{-\frac{u^2}{2}n^{2\alpha-1}}$

As Hoeffding's inequality is very general, and holds for all finite n and t, it is not surprising that it is not asymptotically sharp. For example, CLT will show us that $(S_n - \mathbf{E}[S_n])/\sqrt{n} \xrightarrow{d} N(0,\sigma^2)$ where $\sigma^2 = \mathrm{Var}(X_1)$. Since $\sigma^2 < d$, and the $N(0,\sigma^2)$ has tails like $e^{-u^2/2\sigma^2}$, Hoeffding's is asymptotically (as $u \to \infty$) not sharp in the moderate regime. In the large deviation regime, there is well studied theory. A basic result there says that $\mathbf{P}(|S_n - \mathbf{E}[S_n]| > nu) \approx e^{-nI(u)}$, where the function $I(u)$ can be written in terms of the moment generating function of X_1. It turns out that if $|X_i| \leq d$, then $I(u)$ is larger than $u^2/2d$ which is what Hoeffding's inequality gave us. Thus Hoeffding's is asymptotically (as $n \to \infty$) not sharp in the large deviation regime.

2.14. Random series with independent terms

In law of large numbers, we considered a sum of n terms scaled by n. A natural question is to ask about convergence of infinite series with terms that are independent random variables. Of course $\sum X_n$ will not converge if X_i are i.i.d (unless $X_i = 0$ a.s!). Consider an example.

Example 2.46. Let a_n be i.i.d with finite mean. Important examples are $a_n \sim N(0,1)$ or $a_n = \pm 1$ with equal probability. Then, define $f(z) = \sum_n a_n z^n$. What is the radius of convergence of this series? From the formula for radius of convergence