# 3 | Laws of Large Numbers: Weak and Strong

The Weak Law of Large Numbers says that, for any sequence $X_1, X_2, \ldots$ of i.i.d. random variables with finite mean $\mathbb{E}[X_1] = \mu$ and finite variance $\mathrm{Var}[X_1] = \sigma^2$, the sample averages

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

converge to $\mu$ in probability. The proof is a simple application of Chebyshev's inequality. But, in fact, much more can be said:

*Theorem 1 (Strong Law of Large Numbers).* Under the above assumptions, the sequence of the sample means $\overline{X}_n$ converges to $\mu$ almost surely.

Before we prove this theorem, let us try to understand its meaning. Consider the act of repeatedly tossing a fair coin, such that each toss is independent of all tosses before it. The underlying probability space can be constructed as follows. Let $\Omega$ be the set of all one-sided infinite sequences $\omega = (\omega_1, \omega_2, \ldots)$ such that $\omega_i \in \{\mathtt{H}, \mathtt{T}\}$ for all $i$; let $\mathcal{F}$ be the smallest $\sigma$-field containing all sets of the form

$$E(k; i_1, \ldots, i_k; w_1, \ldots, w_k) = \{\omega \in \Omega : \omega_{i_1} = w_1, \ldots, \omega_{i_k} = w_k\} \tag{1}$$

for all $k \in \mathbb{N}$, all choices $1 \leq i_1 < \ldots < i_k$ of $k$ positive integers, and all choices of $w_1, \ldots, w_k \in \{\mathtt{H}, \mathtt{T}\}$. In other words, $\mathcal{F}$ is the smallest $\sigma$-algebra containing all events determined by the outcomes of any finite number of tosses. Finally, let the probability measure $P$ assign probability $2^{-k}$ to each event of the form (1); this uniquely defines $P$ on all of $\mathcal{F}$. Now define the following sequence of random variables:

$$X_n(\omega) \triangleq \begin{cases} 1, & \text{if } \omega_n = \mathtt{H} \\ 0, & \text{if } \omega_n = \mathtt{T} \end{cases} \tag{2}$$

We claim that $X_1, X_2, \ldots$ are i.i.d. Bernoulli(1/2) random variables. To see this, note that for any $k \in \mathbb{N}$, $1 \leq i_1 < \ldots < i_k$, and $w_1, \ldots, w_k \in \{\mathtt{H}, \mathtt{T}\}$ we have

$$P\left(X_{i_1} = w_1, \ldots, X_{i_k} = w_k\right) = P\left(E(k; i_1, \ldots, i_k; w_1, \ldots, w_k)\right) = \frac{1}{2^k}.$$

In particular, $P(X_n = 1) = P(X_n = 0) = 1/2$ for each $n$, and

$$\begin{aligned} P\left(X_{i_1} = w_1, \ldots, X_{i_k} = w_k\right) &= P\left(E(k; i_1, \ldots, i_k; w_1, \ldots, w_k)\right) \\ &= \prod_{j=1}^{k} P(X_{i_j} = w_j). \end{aligned}$$

(Exercise: repeat this construction for a biased coin with bias $p$.) Consider now the following events:

$$L^- \triangleq \left\{\omega \in \Omega : \lim_n \inf_{k \geq n} \frac{1}{k} \sum_{i=1}^{k} X_i(\omega) = \frac{1}{2}\right\}$$
$$L^+ \triangleq \left\{\omega \in \Omega : \lim_n \sup_{k \geq n} \frac{1}{k} \sum_{i=1}^{k} X_i(\omega) = \frac{1}{2}\right\}$$

(Exercise: show that $L^-$ and $L^+$ are, indeed, events, i.e., elements of $\mathcal{F}$.) If $\omega \in L^-$, then for any $\varepsilon > 0$ there exists some $N^- = N^-(\varepsilon)$, such that

$$\inf_{k \geq n} \frac{1}{k} \sum_{i=1}^{k} X_i(\omega) - \frac{1}{2} > -\varepsilon, \qquad \forall n \geq N^-$$

(note that the sequence

$$\underline{S}_n = \inf_{k \geq n} \frac{1}{k} \sum_{i=1}^{k} X_i(\omega), \qquad n = 1, 2, \ldots$$

is monotone increasing, $\underline{S}_1 \leq \underline{S}_2 \leq \ldots$, and bounded, so it has a limit). Similarly, if $\omega \in L^+$, then for any $\varepsilon > 0$ there exists some $N^+ = N^+(\varepsilon)$, such that

$$\sup_{k \geq n} \frac{1}{k} \sum_{i=1}^{k} X_i(\omega) - \frac{1}{2} < \varepsilon, \qquad \forall n \geq N^+$$

(note that the sequence

$$\overline{S}_n = \sup_{k \geq n} \frac{1}{k} \sum_{i=1}^{k} X_i(\omega), \qquad n = 1, 2, \ldots$$

is monotone decreasing, $\overline{S}_1 \geq \overline{S}_2 \geq \ldots$, and bounded, so it has a limit). If $\omega \in L^- \cap L^+$, then for all $n \geq \max\{N^-, N^+\}$ we will have

$$\frac{1}{2} - \varepsilon < \inf_{k \geq n} \frac{1}{k} \sum_{i=1}^{k} X_i(\omega) \leq \sup_{k \geq n} \frac{1}{k} \sum_{i=1}^{k} X_i(\omega) < \frac{1}{2} + \varepsilon,$$

where, for instance,

$$\sup_{k \geq n} \frac{1}{k} \sum_{i=1}^{k} X_i(\omega) = \sup_{k \geq n} \overline{X}_k(\omega)$$

is the largest possible fraction of heads in any sequence of $n$ or more tosses. In other words, if $\omega \in L^- \cap L^+$, then, provided you toss the coin enough times, the fraction of heads will stay arbitrarily close to $1/2$. Now, another way of writing down the definitions of $L^-$ and $L^+$ is

$$L^- = \left\{ \omega \in \Omega : \liminf_{n \to \infty} \overline{X}_n(\omega) = 1/2 \right\}$$
$$L^+ = \left\{ \omega \in \Omega : \limsup_{n \to \infty} \overline{X}_n(\omega) = 1/2 \right\}$$

(consult the course notes for the definitions of liminf and limsup). Since a sequence $\{a_n\}$ of real numbers converges if and only if $\liminf_n a_n = \limsup_n a_n$, we have

$$\left\{ \omega \in \Omega : \overline{X}_n(\omega) \to \frac{1}{2} \right\} = L^- \cap L^+.$$

Using this and the Strong Law of Large Numbers, we see that $P(\overline{X}_n \to 1/2) = P(L^- \cap L^+) = 1$. Therefore, almost sure convergence of the sample means $\overline{X}_n$ to $1/2$ is a statement about long-term *stability* of the patterns of `H`'s and `T`'s in sufficiently long sequences of independent tosses of a fair coin: provided you keep tossing the coin long enough, you are all but assured to see the fraction of heads settling down to its expected value, namely $1/2$.

## Proof of the Strong Law for bounded random variables

We will prove Theorem 1 under an additional assumption that the variables $X_1, X_2, \ldots$ are bounded with probability one, i.e., there exist some $-\infty < a \leq b < \infty$, such that $P(a \leq X_1 \leq b) = 1$. Our strategy will be as follows: We will first show that, for any $\varepsilon > 0$,

$$\sum_{n=1}^{\infty} P\left( \left| \overline{X}_n - \mu \right| \geq \varepsilon \right) < \infty. \tag{3}$$

Using this and the Borel–Cantelli lemma, we will then conclude that $\overline{X}_n \xrightarrow{\text{a.s.}} \mu$. In fact, we will prove a lot more than just (3): we will also show that the probability that $|\overline{X}_n - \mu| \geq \varepsilon$ decays *exponentially fast* with $n$.

At the heart of the proof lies a very useful bounding technique, which is typically referred to as the *Chernoff–Hoeffding technique* (although it seems to go back at least to the work of S.N. Bernstein in 1927). The idea is as

follows. Consider a random variable $Z$, and suppose that we wish to bound the probability that $Z \geq t$ for some $t > 0$. Observe that for any $s > 0$ we have

$$P(Z \geq t) = P(e^{sZ} \geq e^{st}) \leq e^{-st}\,\mathbb{E}[e^{sZ}], \tag{4}$$

where the first step uses the monotonicity of the exponential function, and the second step uses Markov's inequality. Observe, by the way, that $\mathbb{E}[e^{sZ}]$ is the moment generating function $\Psi_Z(s)$. The trick is to choose an $s > 0$ that would make the right-hand side of (4) suitably small. In fact, since (4) holds *simultaneously* for all $s > 0$, the best thing to do would be to minimize the right-hand side over all such $s$:

$$P(Z \geq t) \leq \inf_{s>0} e^{-st}\Psi_Z(s).$$

However, a good upper bound on the moment generating function $\Psi_Z$ is often sufficient. One such bound was derived by Hoeffding for the case when $Z$ is bounded with probability 1:

*Lemma 1 (Hoeffding).* Let $Z$ be a random variable that satisfies $P(a \leq Z \leq b) = 1$ for some $-\infty < a \leq b < \infty$. Then

$$\Psi_Z(s) = \mathbb{E}[e^{sZ}] \leq e^{\mathbb{E}[Z]} e^{s^2(b-a)^2/8}. \tag{5}$$

The proof of this lemma uses convexity, and can be found in a variety of places.

Let us center the random variables $\{X_i\}$ by defining $Z_i = X_i - \mu$ for all $i$. Then $\overline{X}_n - \mu = n^{-1}\sum_{i=1}^n Z_i$, and

$$\begin{aligned}
P\left(\overline{X}_n - \mu \geq \varepsilon\right) &= P\left(\sum_{i=1}^n Z_i \geq n\varepsilon\right) \\
&\leq e^{-sn\varepsilon}\,\mathbb{E}\left[e^{s\sum_{i=1}^n Z_i}\right] \\
&= e^{-sn\varepsilon}\,\mathbb{E}\left[\prod_{i=1}^n e^{sZ_i}\right] \\
&= e^{-sn\varepsilon}\prod_{i=1}^n \mathbb{E}\left[e^{sZ_i}\right] \\
&= e^{-sn\varepsilon}\left(\Psi_{Z_1}(s)\right)^n,
\end{aligned}$$

where in the second line we have used the Chernoff–Hoeffding trick, and the last two steps use the fact that $Z_1, Z_2, \ldots$ are i.i.d. Since $X_1 \in [a, b]$ with probability one, $Z_1 \in [a - \mu, b - \mu]$ with probability one, and $\mathbb{E}[Z_1] = 0$. Thus, applying Lemma 1, we get

$$\Psi_{Z_1}(s) \leq e^{s^2(b-a)^2/8}.$$

Consequently,

$$P\left(\overline{X}_n - \mu \geq \varepsilon\right) \leq \exp\left(-sn\varepsilon + \frac{ns^2(b-a)^2}{8}\right), \qquad \forall s > 0. \tag{6}$$

If we optimize the right-hand side of (6) with respect to $s$, we get

$$P\left(\overline{X}_n - \mu \geq \varepsilon\right) \leq \exp\left(-\frac{2n\varepsilon^2}{(b-a)^2}\right). \tag{7}$$

Applying the same ideas to $-\overline{X}_n$, we get

$$P\left(\overline{X}_n - \mu \leq -\varepsilon\right) \leq \exp\left(-\frac{2n\varepsilon^2}{(b-a)^2}\right). \tag{8}$$

Therefore, using the union bound together with (7) and (8) gives us

$$\begin{aligned} P\left(\left|\overline{X}_n - \mu\right| \geq \varepsilon\right) &\leq P\left(\overline{X}_n - \mu \geq \varepsilon\right) + P\left(\overline{X}_n - \mu \leq -\varepsilon\right) \\ &\leq 2\exp\left(-\frac{2n\varepsilon^2}{(b-a)^2}\right). \end{aligned} \tag{9}$$

From this, we immediately get the Weak Law (but with a much better estimate of the rate of convergence than what Chebyshev's inequality would give you); moreover, because (3) holds, the sequence $\overline{X}_n$ converges to $\mu$ a.s., so we get the Strong Law as well.

**Back to coin tossing.** If we return to our coin tossing example, we can now answer the following question: how many times do you need to toss a coin with unknown bias $p$ to guarantee that the fraction of heads is within $\varepsilon > 0$ of $p$ with probability at least $1-\delta$? The answer is simple: using (9) with $a = 0$ and $b = 1$, we get

$$P\left(\left|\overline{X}_n - p\right| \geq \varepsilon\right) \leq 2e^{-2n\varepsilon^2}.$$

If we want the right-hand side to be no more than $\delta$, then

$$n = \left\lceil \frac{1}{2\varepsilon^2} \ln \frac{2}{\delta} \right\rceil$$

tosses will suffice. Let's compare this with the bound you get from Chebyshev's inequality. Since we don't know the bias $p$ ahead of time, we cannot use this information, but we can always upper-bound the variance of $\overline{X}_n - p$ by $1/n$:

$$P\left(\left|\overline{X}_n - p\right| \geq \varepsilon\right) \leq \frac{1}{n\varepsilon^2} \qquad \Longrightarrow \qquad n = \left\lceil \frac{1}{\varepsilon^2 \delta} \right\rceil \text{ tosses are needed}$$

The lower bound we get using the Chernoff–Hoeffding method is much better!