Elementary Mathematical Inequalities to Prove the Laws of Large Numbers

Yuchen Wang *

University of Wisconsin Madison, Madison, Wisconsin, America

* Corresponding Author Email: wang3628@wisc.edu

Abstract. This paper gives a simple and clear proof of the Weak Law of Large Numbers (WLLN) and the Strong Law of Large Numbers (SLLN). We only use basic tools from real analysis and elementary probability. First, we prove Markov's inequality, Cheby- shev's inequality, and the Borel–Cantelli lemma in a direct way. We also give a proof of Kolmogorov's maximal inequality. After that, we use these inequalities to prove the WLLN with Chebyshev's inequality and the SLLN with Kolmogorov's inequality and the Borel–Cantelli lemma. The main goal of this paper is to show that the laws of large num- bers can be proved with simple ideas, without advanced probability theory. Our work helps beginners understand these important results in a more easy and friendly way.

Keywords: Statistics, Applied mathematics, Chebyshev's inequality, Markov's inequality, Borel–Cantelli lemma, Kolmogorov's maximal inequality, WLLN, SLLN, LLN.

1. Introduction

The Laws of Large Numbers are basic and important results in probability. They say that when we take many independent observations, the average value will get close to the true expected value. [1] These laws play a key role in statistics and many real problems. There are many ways to prove them, but in this paper we choose a simple and clear method. We use classical inequalities and the Borel–Cantelli lemma to build the proofs step by step.[2]

In this work, we first prove several key tools, including Markov's inequality, Chebyshev's inequality, the Borel–Cantelli lemma, and Kolmogorov's maximal inequality. Then we show how these results help us prove the Weak Law of Large Numbers (WLLN) and the Strong Law of Large Numbers (SLLN). We focus on the case of independent and identically distributed random variables with finite variance.[3] The goal of this paper is to show that these strong results do not need very advanced ideas. With simple inequalities and clear steps, we can understand and prove the laws of large numbers. This gives beginners a friendly way to learn these fundamental results.

2. Notation and preliminaries

Let (Ω , F, e) be a probability space. All random variables in this paper are assumed to be real-valued and measurable with respect to the σ-algebra F. We only work with basic objects from classical probability theory, so the setting here is simple, transparent, and sufficient for proving.

the main results. This elementary framework allows us to focus on the probabilistic ideas themselves without introducing additional technical complications such as abstract measure- theoretic constructions or general Banach-space valued random variables.

For a random variable X with finite first moment, i.e., $E|X| < \infty$, we denote its expectation by

$$\mu = E[X].$$

Furthermore, when the second moment is finite, $E[(X-\mu)^2] < \infty$, we define its variance as

$$\text{Var}(X) = \sigma^2 = E[(X-\mu)^2].$$

These basic quantities—mean and variance—play a central role in our proofs of the laws of large numbers. They provide a measure of central tendency and spread, respectively, which are essential for controlling deviations of sums of random variables.

For an event A ∈ F, we write 1_A for its indicator function [4], which is defined by

$$\mathbf{1}_A(\omega) = \begin{cases} 1, & \text{if } \omega \in A, \\ 0, & \text{if } \omega \notin A. \end{cases}$$

The indicator function is a simple but powerful tool: it allows us to represent events as ran- dom variables and to manipulate probabilities using expectations. Indeed, using the indicator function, the probability of an event can be written as an expectation:

$$P(A) = E[1_A]. \tag{1}$$

This identity provides a direct link between probability and expectation, which is exploited extensively in many of the proofs in this paper. By converting statements about events into statements about random variables, the indicator function enables us to apply linearity of expectation, variance calculations, and classical inequalities in a uniform and systematic way. Moreover, it serves as a bridge between the combinatorial or set-theoretic view of probability and the analytical approach based on random variables and moments, which is particularly convenient when dealing with sums of independent variables, maxima, or sequences of events.

In summary, the definitions and notation introduced here are intentionally simple and con- crete, yet they provide all the tools necessary for the elementary and self-contained arguments that follow.

2.1. Markov's inequality

Let Y ≥ 0 be a nonnegative random variable with finite expectation. For any a > 0,

$$\mathbb{P}(Y \geq a) \leq \frac{\mathbb{E}[Y]}{a}. \tag{2}$$

This is called Markov's inequality. It gives a simple upper bound on the probability that Y takes a large value. The idea is very basic: if the average of Y is not very big, then Y cannot be large too often. This result is one of the first tools used in probability, and it is the starting point for many other inequalities.

Since

$$Y \geq a\mathbf{1}_{\{Y \geq a\}}, \tag{3}$$

we see that Y is at least a whenever the event {Y ≥ a} happens. The function $1_{\{Y \geq a\}}$ is the indicator of this event. It takes the value 1 when Y ≥ a, and it is 0 when Y < a. This simple idea lets us compare the random variable Y with a constant times an indicator event.

Taking expectation yields

$$E[Y] \geq a \cdot P(Y \geq a). \tag{4}$$

This means the average value of Y must be at least a times the probability that Y is at least a. If the probability of being large was too big, then the expectation would also have to be large. Rearranging (4) gives (2).

Markov's inequality is very useful because it is easy to apply. It does not need many assumptions. We only require that Y is nonnegative and has a finite mean. Even though it is simple, it will help us prove more advanced results later, including Chebyshev's inequality and the laws of large numbers. It also shows an important idea in probability: extreme events cannot happen too often if the average size of a random variable is controlled.

2.2. Chebyshev's inequality

Let X be a random variable with mean μ and finite variance σ^2. For every ε > 0,

$$\mathbb{P}(|X - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}. \tag{5}$$

Apply Markov's inequality to

$$Y = (X - \mu)^2 \quad \geq 0 \tag{6}$$

with $a = \varepsilon^2$. Then

$$\mathbb{P}(|X - \mu| \geq \varepsilon) = \mathbb{P}((X - \mu)^2 \geq \varepsilon^2) \leq \frac{\mathbb{E}[(X - \mu)^2]}{\varepsilon^2} = \frac{\sigma^2}{\varepsilon^2}, \tag{7}$$

which is (5).

3. Weak law of large numbers (WLLN)

Let $X_1, X_2, \ldots$ be i.i.d. random variables with finite mean $\mu = E[X_1]$ and finite variance $\sigma^2 = \mathrm{Var}(X_1)$. Define the sample mean

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i. \tag{8}$$

The weak law of large numbers states that, as the number of observations grows, the sample mean converges in probability to the true mean. More precisely, for every $\varepsilon > 0$,

$$\lim_{n \to \infty} \mathbb{P}(|\overline{X}_n - \mu| \geq \varepsilon) = 0, \tag{9}$$

which can also be expressed in the probabilistic convergence notation as

$$\overline{X}_n \xrightarrow{\mathbb{P}} \mu. \tag{10}$$

This convergence in probability captures the intuitive idea that for a large number of inde- pendent observations, the probability of a significant deviation of the sample mean from the population mean becomes arbitrarily small.

To see why this holds, first note that linearity of expectation gives

$$\mathbb{E}[\overline{X}_n] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[X_i] = \mu. \tag{11}$$

This shows that the sample mean is an unbiased estimator of the population mean. Further- more, independence and identical distribution imply

$$\mathrm{Var}(\overline{X}_n) = \mathrm{Var}\Big(\frac{1}{n} \sum_{i=1}^{n} X_i\Big) = \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}(X_i) = \frac{\sigma^2}{n}. \tag{12}$$

Thus, as the number of samples increases, the variance of the sample mean decreases inversely proportional to the sample size. This is the key quantitative insight behind the weak law.

Applying Chebyshev's inequality, which provides an upper bound on the probability of deviations from the mean in terms of the variance, we obtain for any $\varepsilon > 0$,

$$\mathbb{P}(|\overline{X}_n - \mu| \geq \varepsilon) \leq \frac{\mathrm{Var}(\overline{X}_n)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2}. \tag{13}$$

The right-hand side clearly tends to zero as $n \to \infty$, confirming that the probability of observing a large deviation of the sample mean from the true mean vanishes in the large-sample limit. This establishes the weak law of large numbers in a straightforward and self-contained manner.

The WLLN is a fundamental result in probability theory because it formalizes the intuitive principle that the average of many independent measurements stabilizes around the expected value. It is widely used in statistics, stochastic processes, and applied probability, serving as a foundational justification for replacing population parameters with empirical averages in practical situations.[5]

4. Borel–Cantelli lemma

4.1. Borel–Cantelli

Let $(A_n)_{n\geq 1}$ be sequences in (Ω, F, P).

(i) If $\Sigma_{n=1}^{\infty} P(A_n) < \infty$, then

$$\mathbb{P}\Big(\limsup_{n\to\infty} A_n\Big) = 0. \tag{14}$$

(ii) If the sequences (A_n) are independent and $\Sigma_{n=1}^{\infty} P(A_n) = \infty$, then

$$\mathbb{P}\Big(\limsup_{n\to\infty} A_n\Big) = 1. \tag{15}$$

4.2. Proof of (i)

For each $m \geq 1$ we have the union bound

$$\mathbb{P}\Big(\bigcup_{n=m}^{\infty} A_n\Big) \leq \sum_{n=m}^{\infty} \mathbb{P}(A_n). \tag{16}$$

By assumption the right-hand side tends to 0 as $m \to \infty$. Since

$$\limsup_{n\to\infty} A_n = \bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} A_n, \tag{17}$$

and probability is continuous from above for a decreasing sequence of sets, we obtain

$$\mathbb{P}\Big(\limsup_{n\to\infty} A_n\Big) = \lim_{m\to\infty} \mathbb{P}\Big(\bigcup_{n=m}^{\infty} A_n\Big) = 0, \tag{18}$$

which is (14).

4.3. Proof of (ii)

Assume mutual independence of the events (A_n) and $\Sigma_{n=1}^{\infty} p_n = \infty$ where $p_n = P(A_n)$. Fix $m \geq 1$ and consider

$$B_m := \bigcap_{n=m}^{\infty} A_n^c, \tag{19}$$

the event that none of $A_m, A_{m+1}, \ldots$ occur. For any finite $N \geq m$, independence gives

$$\mathbb{P}\Big(\bigcap_{n=m}^{N} A_n^c\Big) = \prod_{n=m}^{N} (1 - p_n). \tag{20}$$

Using $\log(1 - x) \leq -x$ for $x \in [0, 1)$ we get

$$\sum_{n=m}^{\infty} \log(1 - p_n) \leq -\sum_{n=m}^{\infty} p_n = -\infty, \tag{21}$$

since the series of p_n diverges. Therefore

$$\prod_{n=m}^{\infty} (1 - p_n) = \exp\Big(\sum_{n=m}^{\infty} \log(1 - p_n)\Big) = 0, \tag{22}$$

and hence $P(B_m) = 0$. But

$$\left(\limsup_{n\to\infty} A_n\right)^c = \bigcup_{m=1}^{\infty} B_m, \tag{23}$$

so P ((limsup A_n)^c) = 0 and P(limsup A_n) = 1, which is (15).

5. Kolmogorov's maximal inequality

Kolmogorov's maximal inequality provides a control on the maximum of partial sums of independent zero-mean variables. [6] This inequality is very useful in probability theory. In par- ticular, it is a key tool for proving the strong law of large numbers (SLLN) when the random variables have finite variance. The inequality allows us to bound the probability that the partial sums become large, which is crucial for controlling the fluctuations of sums of random variables.

Let Y_1,..., Yn be independent random variables with E[Y_i] = 0 and E[Y_i^2] < ∞. Define the partial sums

$$S_k = \sum_{i=1}^{k} Y_i, \qquad S_0 = 0. \tag{24}$$

Then for any λ > 0, the probability that the maximum absolute value of these partial sums exceeds λ can be bounded as

$$\mathbb{P}\Big(\max_{1\le k\le n} |S_k| \ge \lambda\Big) \le \frac{\mathbb{E}[S_n^2]}{\lambda^2} = \frac{\sum_{i=1}^{n}\mathbb{E}[Y_i^2]}{\lambda^2}. \tag{25}$$

This gives a precise estimate of the likelihood of large deviations of partial sums from zero. Let

$$A := \{\max_{1\le k\le n} |S_k| \ge \lambda\}. \tag{26}$$

Define the (optional) stopping time with respect to the natural filtration [7]:

$$\tau(\omega) = \min\{1 \le k \le n : |S_k(\omega)| \ge \lambda\} \tag{27}$$

on A, and set τ(ω) = n + 1 on A^c. For convenience, define

$$T = \tau \wedge n, \tag{28}$$

so that 1 ≤ T ≤ n on A and T = n on A^c. This construction helps to isolate the first time the partial sum becomes large.

On the event {τ ≤ n} we have $|S_T| \ge \lambda$. Consider the decomposition

$$S_n = S_T + (S_n - S_T). \tag{29}$$

Squaring both sides, multiplying by $\mathbf{1}_A$, and taking expectation gives [8]

$$E[S_n^2\mathbf{1}_A] = E[S_T^2\mathbf{1}_A] + E[(S_n - S_T)^2\mathbf{1}_A] + 2E[S_T(S_n - S_T)\mathbf{1}_A]. \tag{30}$$

We claim that the cross term vanishes. Indeed, by the tower property of conditional expectation and the independence of the increments,

$$E[S_T(S_n - S_T)\mathbf{1}_A] = E[\mathbf{1}_A S_T E[S_n - S_T \mid F_T]] = 0, \tag{31}$$

because given F_T , the increment $S_n - S_T = \sum_{i=T+1}^{n} Y_i$ has conditional mean zero. This shows that the contribution of the cross term is exactly zero.

Furthermore, since $(S_n - S_T)^2\mathbf{1}_A \ge 0$, we have

$$E[S_n^2\mathbf{1}_A] \ge E[S_T^2\mathbf{1}_A] \ge \lambda^2 P(A), \tag{32}$$

because on A we know $|S_T| \geq \lambda$ and hence $S_T^2 \mathbf{1}_A \geq \lambda^2 \mathbf{1}_A$. Therefore, we can write

$$\mathbb{P}(A) \leq \frac{\mathbb{E}[S_n^2 \mathbf{1}_A]}{\lambda^2} \leq \frac{\mathbb{E}[S_n^2]}{\lambda^2}, \tag{33}$$

which reproduces (25). Finally, independence and zero mean imply that all cross termsin $E[S_n^2]$ vanish, so

$$\mathbb{E}[S_n^2] = \sum_{i=1}^{n} \mathbb{E}[Y_i^2]. \tag{34}$$

This completes the proof and illustrates clearly how the maximal inequality provides a bound on the probability of large deviations of partial sums.

6. Strong law of large numbers (SLLN) for finite-variance i.i.d. case

We now give a complete proof of the strong law of large numbers for i.i.d. random variables with finite variance. The argument uses a standard block-decomposition approach, Kolmogorov maximal inequality, and the Borel–Cantelli lemma. These tools allow us to control the devi- ations of partial sums and extend convergence from carefully chosen subsequences to the full sequence.

Let $(X_n)_{n\geq 1}$ be i.i.d. random variables with

$$\mu = E[X_1], \qquad \sigma^2 = \mathrm{Var}(X_1) < \infty.$$

Our goal is to show that the sequence of sample means

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{35}$$

converges almost surely to the expected value μ, i.e.,

$$\overline{X}_n \xrightarrow{\text{a.s.}} \mu.$$

To simplify notation, define the centered variables

$$Y_i = X_i - \mu, \tag{36}$$

so that $E[Y_i] = 0$ and $\mathrm{Var}(Y_i) = \sigma^2$. Let

$$S_n = \sum_{i=1}^{n} Y_i, \tag{37}$$

so that $\overline{X}_n - \mu = S_n/n$. We will show that $S_n/n \to 0$ almost surely, which immediately implies (35).

First, consider the dyadic subsequence $n = 2^k$. By Chebyshev's inequality, for any $\varepsilon > 0$,

$$\mathbb{P}\big(|S_{2^k}| \geq \varepsilon 2^k\big) = \mathbb{P}\big(|\overline{X}_{2^k} - \mu| \geq \varepsilon\big) \leq \frac{\mathrm{Var}(S_{2^k})}{\varepsilon^2 2^{2k}} = \frac{2^k \sigma^2}{\varepsilon^2 2^{2k}} = \frac{\sigma^2}{\varepsilon^2 2^k}. \tag{38}$$

Since

$$\sum_{k=1}^{\infty} \frac{1}{2^k} < \infty, \tag{39}$$

the Borel–Cantelli lemma (part (i)) implies that with probability one only finitely many of the events $\{|S_{2k}| \geq \varepsilon 2^k\}$ occur. Therefore, along the dyadic subsequence,

$$\frac{S_{2^k}}{2^k} \to 0 \quad \text{a.s.} \tag{40}$$

This shows that the sample means converge along the subsequence $n = 2^k$.

Next, to extend the convergence from the dyadic subsequence to the full sequence, we need to control the maximum fluctuations within each dyadic block. For k ≥ 1, define

$$M_k := \max_{2^k < n \le 2^{k+1}} |S_n - S_{2^k}|. \tag{41}$$

Observe that for $2^k < n \le 2^{k+1}$,

$$\frac{|S_n|}{n} \le \frac{|S_{2^k}|}{2^k} \cdot \frac{2^k}{n} + \frac{M_k}{2^k} \cdot \frac{2^k}{n} \le \frac{|S_{2^k}|}{2^k} + \frac{M_k}{2^k}. \tag{42}$$

Hence, if $|S_{2^k}|/2^k \to 0$ and $M_k/2^k \to 0$ almost surely, then it follows that $S_n/n \to 0$ almost surely.

To bound $P(M_k \ge \varepsilon 2^k)$, we apply Kolmogorov's maximal inequality to the block increments $Y_{2^k+1}, \ldots, Y_{2^{k+1}}$. Define

$$T_k(m) = \sum_{i=2^k+1}^{2^k+m} Y_i \tag{43}$$

as the partial sums within the block (so $T_k(0) = 0$ and $T_k(2^k) = S_{2^{k+1}} - S_{2^k}$). Then

$$\mathbb{P}(M_k \ge \varepsilon 2^k) = \mathbb{P}\Big(\max_{1 \le m \le 2^k} |T_k(m)| \ge \varepsilon 2^k \Big) \le \frac{\mathbb{E}[T_k(2^k)^2]}{\varepsilon^2 2^{2k}}. \tag{44}$$

By independence and identical distribution [10],

$$E[T_k(2^k)^2] = 2^k \sigma^2, \tag{45}$$

so that

$$\mathbb{P}(M_k \ge \varepsilon 2^k) \le \frac{2^k \sigma^2}{\varepsilon^2 2^{2k}} = \frac{\sigma^2}{\varepsilon^2 2^k}. \tag{46}$$

Since

$$\sum_{k \ge 1} \mathbb{P}(M_k \ge \varepsilon 2^k) < \infty, \tag{47}$$

the Borel–Cantelli lemma implies that only finitely many of the events $\{M_k \ge \varepsilon 2^k\}$ occur[**?**], i.e.,

$$\frac{M_k}{2^k} \to 0 \quad \text{a.s.} \tag{48}$$

Combining (40) and (48), we conclude that

$$\frac{S_n}{n} \to 0 \quad \text{a.s.}, \quad \text{which implies} \quad \bar{X}_n \to \mu \quad \text{a.s.}$$

This completes the proof of the SLLN under the finite-variance assumption, illustrating how the block decomposition, maximal inequality, and Borel–Cantelli lemma together yield almost sure convergence.

7. Summary

We have presented elementary and self-contained proofs of both the weak law of large numbers (WLLN) and the strong law of large numbers (SLLN) for i.i.d. random variables with finite variance. Our primary goal was to demonstrate that these foundational results in probability theory can be fully understood using basic ideas and standard techniques, without resorting to advanced machinery from

modern probability theory. In particular, we have avoided tools such as martingale convergence theorems, characteristic functions, or measure-theoretic limit theorems, which, although powerful, can obscure the underlying intuition for beginners.

Instead, our approach relies on a sequence of clear and simple steps that illustrate the core reasoning behind the laws. The main ingredients are classical inequalities such as Markov's and Chebyshev's, the Borel–Cantelli lemma, and Kolmogorov's maximal inequality. These tools provide a way to control the deviations of sums of independent random variables, estimate probabilities of large fluctuations, and extend convergence from selected subsequences to the full sequence. By carefully combining these classical results, we were able to prove the WLLN and the SLLN in a straightforward and transparent manner, highlighting the essential ideas without introducing unnecessary technicalities.

This method offers several pedagogical advantages. It allows readers to clearly follow the logical structure of the proofs and understand how the convergence of sample means emerges from independence and finite variance. It also gives beginners a friendly and accessible route to learning these fundamental results, helping them appreciate how elementary arguments can lead to deep theorems. Moreover, although our setting is simple, the underlying ideas are flexible and can be applied in more general contexts. For example, when higher moments exist, the same techniques can be adapted to obtain faster rates of convergence, more refined probability bounds, or extensions to more general sequences of random variables.

Finally, our work illustrates a broader lesson in probability theory: even classical and seem- ingly simple results, such as the WLLN and SLLN, can be fully understood and rigorously proved using only basic, transparent tools. This emphasizes that deep theorems are often accessible through careful and clean arguments, and that foundational principles such as in- dependence, variance control, and classical inequalities are sufficient to reveal the essential behavior of sums of random variables. Looking ahead, these elementary techniques can be extended to study generalized forms of the laws of large numbers, dependent random variables, or random processes with more complex structures. They may also serve as a stepping stone for exploring limit theorems in more abstract settings, providing a solid conceptual foundation for further research in probability theory and its applications. We hope that this exposition will not only clarify the classical laws of large numbers for readers but also inspire them to explore further results in probability with confidence in elementary techniques.

References

[1] Billingsley, P., Probability and Measure, 3rd ed., Wiley, 1995.

[2] Durrett, R., Probability: Theory and Examples, 5th ed., Cambridge University Press, 2019.

[3] Durrett, R., Probability: Theory and Examples, 5th ed., Cambridge University Press, 2019.

[4] Feller, W., An Introduction to Probability Theory and Its Applications, Vol. II, Wiley, 1971.

[5] Chung, K. L., A Course in Probability Theory, 3rd ed., Academic Press, 2001.

[6] Lo`eve, M., Probability Theory I, 4th ed., Springer, 1977.

[7] Kallenberg, O., Foundations of Modern Probability, 2nd ed., Springer, 2002.

[8] Shiryaev, A. N., Probability, 2nd ed., Springer, 1996.

[9] Gnedenko, B. V. and Kolmogorov, A. N., Limit Distributions for Sums of Independent Random Variables, Addison-Wesley, 1954.

[10] Williams, D., Probability with Martingales, Cambridge University Press, 1991.