# The weak and strong laws of large numbers

Jordan Bell
jordan.bell@gmail.com
Department of Mathematics, University of Toronto

May 29, 2015

## 1 Introduction

Using Egorov's theorem, one proves that if a sequence of random variables $X_n$ **converges almost surely** to a random variable $X$ then $X_n$ **converges in probability** to $X$.[1]

Let $X_n$ be a sequence of $L^1$ random variables, $n \geq 1$. **A weak law of large numbers** is a statement that

$$\frac{1}{n}\sum_{k=1}^{n}(X_k - E(X_k)) \tag{1}$$

converges in probability to 0. **A strong law of large numbers** is a statement that (1) converges almost surely to 0. Thus, if the hypotheses assumed on the sequence of random variables are the same, a strong law implies a weak law.

We shall prove the weak law of large numbers for a sequence of independent identically distributed $L^1$ random variables, and the strong law of large for the same hypotheses. We give separate proofs for these theorems as an occasion to inspect different machinery, although to establish the weak law it thus suffices to prove the strong law. One reason to distinguish these laws is for cases when we impose different hypotheses.[2]

We also prove **Markov's weak law of large numbers**, which states that if $X_n$ is a sequence of $L^2$ random variables that are pairwise uncorrelated and

$$\frac{1}{n^2}\sum_{n=k}^{n}\operatorname{Var}(X_k) \to 0,$$

then $\frac{1}{n}\sum_{k=1}^{n}(X_k - E(X_k))$ converges to 0 in $L^2$, from which it follows using Chebyshev's inequality that it converges in probability to 0. (We remark that

[1] V. I. Bogachev, *Measure Theory*, volume I, p. 111, Theorem 2.2.3; `http://individual.utoronto.ca/jordanbell/notes/L0.pdf`, p. 3, Theorem 3.

[2] cf. Jordan M. Stoyanov, *Counterexamples in Probability*, third ed., p. 163, §15.3; Dieter Landers and Lothar Rogge, *Identically distributed uncorrelated random variables not fulfilling the WLLN*, Bull. Korean Math. Soc. **38** (2001), no. 3, 605–610.

a sequence of $L^2$ random variables converging in $L^2$ to 0 does not imply that it converges almost surely to 0, although there is indeed a subsequence that converges almost surely to 0.[3])

If $(\Omega, \mathscr{F}, P)$ is a probability space, $(Y, \mathscr{A})$ is a measurable space, and $T : (\Omega, \mathscr{F}) \to (Y, \mathscr{A})$ is measurable, the **pushforward measure of $P$ by $T$** is

$$(T_*P)(A) = P(T^{-1}(A)), \qquad A \in \mathscr{A}.$$

Then $(Y, \mathscr{A}, T_*P)$ is a probability space. We remind ourselves of the **change of variables theorem**, which we shall use.[4] Let $f : Y \to \mathbb{R}$ be a function. On the one hand, if $f \in L^1(T_*P)$ then $f \circ T \in L^1(P)$ and

$$\int_Y f d(T_*P) = \int_\Omega f \circ T dP. \tag{2}$$

On the other hand, if $f$ is $T_*P$-measurable and $f \circ T \in L^1(P)$, then $f \in L^1(T_*P)$ and

$$\int_Y f d(T_*P) = \int_\Omega f \circ T dP.$$

## 2 The weak law of large numbers

Suppose that $X_n : (\Omega, \mathscr{F}, P) \to \mathbb{R}$, $n \geq 1$, are independent identically distributed $L^1$ random variables, and write

$$S_n = \sum_{k=1}^n X_k,$$

for which $E(S_n) = \sum_{k=1}^n E(X_k) = nE(X_1)$.

**Lemma 1.** *If $X_n$ are $L^2$, then for any $\lambda > 0$ and for any $n \geq 1$,*

$$P\left(\left|\frac{S_n}{n} - E(X_1)\right| \geq \lambda\right) \leq \frac{E(|X_1 - E(X_1)|^2)}{n\lambda^2}.$$

*Proof.* Using

$$S_n^2 = \sum_{k=1}^n X_k^2 + \sum_{j \neq k} X_j X_k,$$

[3] `http://individual.utoronto.ca/jordanbell/notes/L0.pdf`, p. 4, Theorem 5.

[4] Charalambos D. Aliprantis and Kim C. Border, *Infinite Dimensional Analysis: A Hitchhikers Guide*, third ed., p. 485, Theorem 13.46.

we have

$$\begin{aligned}
E\left(\left|\frac{S_n}{n} - E(X_1)\right|^2\right) &= E\left(\sum_{k=1}^n \frac{X_k^2}{n^2} + \sum_{j\neq k} \frac{X_j X_k}{n^2} - 2E(X_1)\frac{S_n}{n} + E(X_1)^2\right) \\
&= \sum_{k=1}^n \frac{E(X_k^2)}{n^2} + \sum_{j\neq k} \frac{E(X_j X_k)}{n^2} - 2E(X_1)^2 + E(X_1)^2 \\
&= \sum_{k=1}^n \frac{E(X_1^2)}{n^2} + \sum_{j\neq k} \frac{E(X_j)E(X_k)}{n^2} - E(X_1)^2 \\
&= \sum_{k=1}^n \frac{E(X_1^2)}{n^2} + \sum_{j\neq k} \frac{E(X_1)^2}{n^2} + E(X_1)^2 \\
&= \frac{E(X_1^2)}{n} + \frac{E(X_1)^2}{n}.
\end{aligned}$$

On the other hand,

$$E(|X_1 - E(X_1)|^2) = E(X_1^2 - 2E(X_1)X_1 + E(X_1)^2) = E(X_1^2) + E(X_1)^2.$$

So

$$E\left(\left|\frac{S_n}{n} - E(X_1)\right|^2\right) = \frac{1}{n}E(|X_1 - E(X_1)|^2).$$

Using this and Chebyshev's inequality,

$$\begin{aligned}
P\left(\left|\frac{S_n}{n} - E(X_1)\right| \geq \lambda\right) &\leq \frac{1}{\lambda^2}E\left(\left|\frac{S_n}{n} - E(X_1)\right|^2\right) \\
&= \frac{1}{n\lambda^2}E(|X_1 - E(X_1)|^2),
\end{aligned}$$

proving the claim. □

We now prove the **weak law of large numbers**, which states that if $X_n$ are independent identically distributed $L^1$ random variables each with mean 0, then $\frac{S_n}{n}$ converges in probability to $E(X_1)$. $Z_n = X_n - E(X_n) = X_n - E(X_1)$ are independent and identically distributed $L^1$ random variables with mean 0, and if $Z_n$ converges in probability to 0 then $X_n = Z_n + E(X_1)$ converges in probability to $E(X_1)$, showing that it suffices to prove the theorem when $E(X_1) = 0$.[5]

**Theorem 2** (Weak law of large numbers)**.** *Suppose that $E(X_1) = 0$. For each $\epsilon > 0$,*

$$P\left(\left|\frac{S_n}{n}\right| \geq \epsilon\right) \to 0, \qquad n \to \infty.$$

[5] Allan Gut, *Probability: A Graduate Course*, second ed., p. 270, Theorem 6.3.1, and p. 121, Theorem 3.1.5.

*Proof.* For $n \geq 1$ and $1 \leq k \leq n$, let

$$Y_{n,k} = 1_{\{|X_k| \leq n\epsilon^3\}} X_k = f \circ X_k,$$

where $f(x) = 1_{[-n\epsilon^3, n\epsilon^3]}(x) \cdot x$. Because $X_1, \ldots, X_k$ are independent and identically distributed, so are $Y_{n,1}, \ldots, Y_{n,n}$.[6] Moreover, $E(Y_{n,k}^2) \leq (n\epsilon^3)^2$, so each $Y_{n,k}$ belongs to $L^2$. Let

$$T_n = \sum_{k=1}^{n} Y_{n,k},$$

for which $E(T_n) = nE(Y_{n,1})$.

If $\omega \in \bigcap_{k=1}^{n} \{|X_k| \leq n\epsilon^3\}$, then

$$T_n(\omega) = \sum_{k=1}^{n} Y_{n,k}(\omega) = \sum_{k=1}^{n} X_k(\omega) = S_n(\omega),$$

and using this and Lemma 1, for $t > 0$ we have

$$\begin{aligned}
P(|S_n - E(T_n)| \geq t) &= P\left(\{|S_n - E(T_n)| \geq t\} \cap \bigcap_{k=1}^{n} \{|X_k| \leq n\epsilon^3\}\right) \\
&\quad + P\left(\{|S_n - E(T_n)| \geq t\} \cap \bigcup_{k=1}^{n} \{|X_k| > n\epsilon^3\}\right) \\
&\leq P(|T_n - E(T_n)| \geq t) + P\left(\bigcup_{k=1}^{n} \{|X_k| > n\epsilon^3\}\right) \\
&= P\left(\left|\frac{T_n}{n} - E(Y_{n,1})\right| \geq \frac{t}{n}\right) + P\left(\bigcup_{k=1}^{n} \{|X_k| > n\epsilon^3\}\right) \\
&\leq \frac{E(|Y_{n,1} - E(Y_{n,1})|^2)}{n \cdot \left(\frac{t}{n}\right)^2} + \sum_{k=1}^{n} P(|X_k| > n\epsilon^3) \\
&= \frac{n}{t^2} \cdot (E(Y_{n,1}^2) - E(Y_{n,1})^2) + nP(|X_1| > n\epsilon^3) \\
&\leq \frac{n}{t^2} \cdot E(Y_{n,1}^2) + nP(|X_1| > n\epsilon^3).
\end{aligned}$$

For $t = n\epsilon$ this is

$$\begin{aligned}
P(|S_n - E(T_n)| \geq n\epsilon) &\leq \frac{1}{n\epsilon^2} \cdot E(Y_{n,1}^2) + nP(|X_1| > n\epsilon^3) \\
&= \frac{1}{n\epsilon^2} \cdot E(1_{\{|X_1| \leq n\epsilon^3\}} X_1^2) + nP(|X_1| > n\epsilon^3) \\
&\leq \frac{1}{n\epsilon^2} \cdot E(1_{\{|X_1| \leq n\epsilon^3\}} n\epsilon^3 |X_1|) + nP(|X_1| > n\epsilon^3) \\
&\leq \epsilon E(|X_1|) + nP(|X_1| > n\epsilon^3).
\end{aligned}$$

[6] Gerald B. Folland, *Real Analysis: Modern Techniques and Their Applications*, second ed., p. 316, Proposition 10.2.

Now,

$$n\epsilon^3 P(|X_1| > n\epsilon^3) = n\epsilon^3 \int_{(n\epsilon^3,\infty)} d(X_{1*}P)(y) \leq \int_{(n\epsilon^3,\infty)} y d(X_{1*}P)(y),$$

which tends to 0 as $n \to \infty$ because

$$\int_{\mathbb{R}} |y| d(X_{1*}P)(y) = E(|X_1|) < \infty.$$

Therefore,

$$\limsup_n P(|S_n - E(T_n)| \geq n\epsilon) \leq \epsilon E(|X_1|),$$

that is,

$$\limsup_n P\left(\left|\frac{S_n - E(T_n)}{n}\right| \geq \epsilon\right) \leq \epsilon E(|X_1|),$$

which implies that $\frac{S_n - E(T_n)}{n}$ converges in probability to 0.

Because $E(X_1) = 0$,

$$E(1_{\{|X_1| \leq n\epsilon^3\}} X_1) + E(1_{\{|X_1| > n\epsilon^3\}} X_1) = 0,$$

and hence

$$|E(T_n)| = |nE(Y_{n,1})| = n|E(1_{\{|X_1| \leq n\epsilon^3\}} X_1)| = n|E(1_{\{|X_1| > n\epsilon^3\}} X_1)|,$$

thus

$$\frac{|E(T_n)|}{n} = |E(1_{\{|X_1| > n\epsilon^3\}} X_1)| \leq E(1_{\{|X_1| > n\epsilon^3\}} |X_1|),$$

which tends to 0 as $n \to \infty$, because $E(|X_1|) < \infty$. Thus $\frac{E(T_n)}{n}$ converges in probability to 0, and therefore $\frac{S_n}{n}$ converges in probability to 0, completing the proof. □

**Lemma 3** (Bienaymé's formula)**.** *If $X_n$, $n \geq 1$, are $L^2$ random variables that are pairwise uncorrelated, then*

$$\mathrm{Var}\left(\sum_{k=1}^n X_k\right) = \sum_{k=1}^n \mathrm{Var}(X_k).$$

*Proof.* Let $Y_k = X_k - E(X_k)$. Using that the $X_k$ are pairwise uncorrelated, for $j \neq k$ we have

$$\begin{aligned} E(Y_j Y_k) &= E(X_j X_k - X_j E(X_k) - X_k E(X_j) + E(X_j)E(X_k)) \\ &= E(X_j)E(X_k) - E(X_j)E(X_k) - E(X_k)E(X_j) + E(X_j)E(X_k) \\ &= 0, \end{aligned}$$

showing that the $Y_k$ are pairwise uncorrelated. Then, because $E(S_n) = \sum_{k=1}^n E(X_k)$,

$$\begin{aligned}
\mathrm{Var}(S_n) &= E\left(\left(S_n - \sum_{k=1}^n E(X_k)\right)^2\right) \\
&= E\left(\left(\sum_{k=1}^n Y_k\right)^2\right) \\
&= E\left(\sum_{k=1}^n Y_k^2 + \sum_{j\neq k} Y_j Y_k\right) \\
&= \sum_{k=1}^n E(Y_k^2) + \sum_{j\neq k} E(Y_j)E(Y_k) \\
&= \sum_{k=1}^n E(Y_k^2),
\end{aligned}$$

and as $E(Y_k^2) = \mathrm{Var}(X_k)$, we have

$$\mathrm{Var}(S_n) = \sum_{k=1}^n \mathrm{Var}(X_k).$$

□

We now prove a weak law of large numbers, that is sometimes attributed to Markov, that neither supposes that the random variables are independent nor supposes that they are identically distributed. We remind ourselves that if a sequence $Y_n$ of $L^2$ random variables converges to 0 in $L^2$ then it converges in probability to 0; this is proved using Chebyshev's inequality.

**Theorem 4** (Markov's weak law of large numbers)**.** *If $X_n$, $n \geq 1$, are $L^2$ random variables that are pairwise uncorrelated and which satisfy*

$$\frac{1}{n^2}\sum_{n=k}^n \mathrm{Var}(X_k) \to 0,$$

*then $\frac{1}{n}\sum_{k=1}^n (X_k - E(X_k))$ converges to 0 in $L^2$ and thus in probability.*

*Proof.* Because $E\left(\sum_{k=1}^n X_k\right) = \sum_{k=1}^n E(X_k)$ and $\mathrm{Var}(X) = E((X - E(X))^2)$,

using Bienaymé's formula we get

$$E\left(\left(\frac{1}{n}\sum_{k=1}^{n}(X_k - E(X_k))\right)^2\right) = \frac{1}{n^2}E\left(\left(\sum_{k=1}^{n}X_k - \sum_{k=1}^{n}E(X_k)\right)^2\right)$$
$$= \frac{1}{n^2}\mathrm{Var}\left(\sum_{k=1}^{n}X_k\right)$$
$$= \frac{1}{n^2}\sum_{k=1}^{n}\mathrm{Var}(X_k).$$

Thus

$$E\left(\left(\frac{1}{n}\sum_{k=1}^{n}(X_k - E(X_k))\right)^2\right) = \frac{1}{n^2}\sum_{k=1}^{n}\mathrm{Var}(X_k) \to 0$$

as $n \to \infty$, namely, $\frac{1}{n}\sum_{k=1}^{n}(X_k - E(X_k))$ converges to 0 in $L^2$ as $n \to \infty$, proving the claim. □

# 3 Ergodic theory

We here assemble machinery and results that we will use to prove the strong law of large numbers. For a probability space $(\Omega, \mathscr{F}, P)$, a function $T : \Omega \to \Omega$ is said to be a **measure-preserving transformation** if (i) it is measurable, and (ii) $T_*P = P$. To say that $T_*P = P$ means that for any $A \in \mathscr{F}$, $(T_*P)(A) = P(A)$, i.e. $P(T^{-1}(A)) = P(A)$.

A collection $\mathscr{A}$ of subsets of $\Omega$ is called an **algebra of sets** if (i) $\emptyset \in \mathscr{A}$, (ii) $\Omega \in \mathscr{A}$, (iii) if $A, B \in \mathscr{A}$ then $A \cup B, A \setminus B \in \mathscr{A}$. If $\mathscr{G}$ is a nonempty collection of subsets of $\Omega$, it is a fact that there is a unique algebra of sets $A(\mathscr{G})$ that (i) contains $\mathscr{G}$ and (ii) is contained in any algebra of sets that contains $\mathscr{G}$. We call $A(\mathscr{G})$ the **algebra of sets generated by** $\mathscr{G}$.

A collection $\mathscr{S}$ of subsets of $\Omega$ is called a **semialgebra of sets** if (i) $\emptyset \in \mathscr{S}$, (ii) $\Omega \in \mathscr{S}$, (iii) if $A, B \in \mathscr{S}$ then $A \cap B \in \mathscr{S}$, (iv) if $A, B \in \mathscr{S}$ then there are pairwise disjoint $E_1, \ldots, E_n \in \mathscr{S}$ such that

$$A \setminus B = \bigcup_{k=1}^{n} E_k.$$

It is a fact that $A(\mathscr{S})$ is equal to the collection of all unions of finitely many pairwise disjoint elements of $\mathscr{S}$.[7]

A nonempty collection $\mathscr{M}$ of subsets of $\Omega$ is called a **monotone class** if whenever $A_n \in \mathscr{M}$, $A_n \subset A_{n+1}$ (an **increasing sequence of sets**), implies that $\bigcup_n A_n \in \mathscr{M}$ and $A_n \in \mathscr{M}$, $A_{n+1} \subset A_n$ (a **decreasing sequence of sets**), implies that $\bigcap_n A_n \in \mathscr{M}$. In other words, a monotone class is a nonempty

[7] V. I. Bogachev, *Measure Theory*, volume I, p. 8, Lemma 1.2.14.

collection of subsets of $\Omega$ such that if $A_n$ is a monotone sequence in $\mathscr{M}$ then $\lim_{n\to\infty} A_n \in \mathscr{M}$. If $\mathscr{G}$ is a nonempty collection of subsets of $\Omega$, it is a fact that there is a unique monotone class $M(\mathscr{G})$ that (i) contains $\mathscr{G}$ and (ii) is contained in any monotone class that contains $\mathscr{G}$. We call $M(\mathscr{G})$ the **monotone class generated by** $\mathscr{G}$. The **monotone class theorem** states that if $\mathscr{A}$ is an algebra of sets then $\sigma(\mathscr{A}) = M(\mathscr{A})$.[8]

The following lemma gives conditions under which we can establish that a function is measure-preserving.[9]

**Lemma 5.** *Let $T : \Omega \to \Omega$ be a function and suppose that $\mathscr{S}$ is a semialgebra of sets for which $\mathscr{F}$ is the $\sigma$-algebra generated by $\mathscr{S}$. If (i) $T^{-1}(A) \in \mathscr{F}$ for each $A \in \mathscr{S}$ and (ii) $P(T^{-1}(A)) = P(A)$ for each $A \in \mathscr{S}$, then $T$ is measure-preserving.*

*Proof.* Let

$$\mathscr{C} = \{A \in \mathscr{F} : T^{-1}(A) \in \mathscr{F}, P(T^{-1}(A)) = P(A)\};$$

we wish to prove that $\mathscr{C} = \mathscr{F}$.

If $A_n \in \mathscr{C}$ is an increasing sequence, let $A = \bigcup_{n=1}^{\infty} A_n$. Then, as $T^{-1}(A_n) \in \mathscr{F}$ for each $n$,

$$T^{-1}(A) = \bigcup_{n=1}^{\infty} T^{-1}(A_n) \in \mathscr{F}$$

and as (i) $T^{-1}(A_n)$ is an increasing sequence, (ii) $P$ is continuous from below,[10] and (iii) $P(T^{-1}(A_n)) = P(A_n)$,

$$\begin{aligned} P(T^{-1}(A)) &= P\left(\bigcup_{n=1}^{\infty} T^{-1}(A_n)\right) \\ &= \lim_{n\to\infty} P(T^{-1}(A_n)) \\ &= \lim_{n\to\infty} P(A_n) \\ &= P\left(\bigcup_{n=1}^{\infty} A_n\right) \\ &= P(A), \end{aligned}$$

and hence $A \in \mathscr{C}$. If $A_n \in \mathscr{C}$ is a decreasing sequence, let $A = \bigcap_{n=1}^{\infty} A_n$. Because $T^{-1}(A_n) \in \mathscr{F}$,

$$T^{-1}(A) = \bigcap_{n=1}^{\infty} T^{-1}(A_n) \in \mathscr{F},$$

[8] V. I. Bogachev, *Measure Theory*, volume I, p. 33, Theorem 1.9.3.
[9] Peter Walters, *An Introduction to Ergodic Theory*, p. 20, Theorem 1.1.
[10] V. I. Bogachev, *Measure Theory*, volume I, p. 9, Proposition 1.3.3.

and as (i) $T^{-1}(A_n)$ is a decreasing sequence, (ii) $P$ is continuous from above, and (iii) $P(T^{-1}(A_n)) = P(A_n)$,

$$\begin{aligned} P(T^{-1}(A)) &= P\left(\bigcap_{n=1}^{\infty} T^{-1}(A_n)\right) \\ &= \lim_{n\to\infty} P(T^{-1}(A_n)) \\ &= \lim_{n\to\infty} P(A_n) \\ &= P\left(\bigcap_{n=1}^{\infty} A_n\right) \\ &= P(A), \end{aligned}$$

and hence $A \in \mathscr{C}$. Therefore, $\mathscr{C}$ is a monotone class.

$\mathscr{S} \subset \mathscr{C}$. If $A \in A(\mathscr{S})$, then there are pairwise disjoint $A_1, \ldots, A_n \in \mathscr{S}$ with $A = \bigcup_{k=1}^{n} A_k$. As $T^{-1}(A_k) \in \mathscr{F}$,

$$T^{-1}(A) = \bigcup_{k=1}^{n} T^{-1}(A_k) \in \mathscr{F}.$$

As the $A_k$ are pairwise disjoint, so are the sets $T^{-1}(A_k)$, so, because $P(T^{-1}(A_k)) = P(A_k)$,

$$\begin{aligned} P(T^{-1}(A)) &= P\left(\bigcup_{k=1}^{n} T^{-1}(A_k)\right) \\ &= \sum_{k=1}^{n} P(T^{-1}(A_k)) \\ &= \sum_{k=1}^{n} P(A_k) \\ &= P\left(\bigcup_{k=1}^{n} A_k\right) \\ &= P(A). \end{aligned}$$

Therefore $A \in \mathscr{C}$. This shows that $A(\mathscr{S}) \subset \mathscr{C}$.

The monotone class theorem tells us $\sigma(A(\mathscr{S})) = M(A(\mathscr{S}))$. On the one hand, $\mathscr{F} = \sigma(\mathscr{S})$ and hence $\mathscr{F} = \sigma(A(\mathscr{S}))$. On the other hand, $A(\mathscr{S}) \subset \mathscr{C}$ and the fact that $\mathscr{C}$ is a monotone class yield

$$M(A(\mathscr{S})) \subset M(\mathscr{C}) = \mathscr{C}.$$

Therefore

$$\mathscr{F} \subset \mathscr{C}.$$

Of course $\mathscr{C} \subset \mathscr{F}$, so $\mathscr{C} = \mathscr{F}$, proving the claim. □

Let $(\Omega, \mathscr{F}, P)$ be a probability space. A measure-preserving transformation $T : \Omega \to \Omega$ is called **ergodic** if $A \in \mathscr{F}$ and $T^{-1}(A) = A$ implies that $P(A) = 0$ or $P(A) = 1$. It is proved[11] using the **Birkhoff ergodic theorem** that for a measure-preserving transformation $T : \Omega \to \Omega$, $T$ is ergodic if and only if for all $A, B \in \mathscr{F}$,

$$\frac{1}{n} \sum_{k=0}^{n-1} P(T^{-1}(A) \cap B) \to P(A)P(B).$$

A measure-preserving transformation $T : \Omega \to \Omega$ is called **weak-mixing** if for all $A, B \in \mathscr{F}$,

$$\frac{1}{n} \sum_{k=0}^{n-1} |P(T^{-k}(A) \cap B) - P(A)P(B)| \to 0.$$

It is immediate that a weak-mixing transformation is ergodic.

A measure-preserving transformation $T : \Omega \to \Omega$ is called **strong-mixing** if for all $A, B \in \mathscr{F}$,

$$P(T^{-n}(A) \cap B) \to P(A)P(B).$$

If a sequence of real numbers $a_n$ tends to 0, then

$$\frac{1}{n} \sum_{k=0}^{n-1} |a_k| \to 0,$$

and using this we check that a strong-mixing transformation is weak-mixing.

The following statement gives conditions under which a measure-preserving transformation is ergodic, weak-mixing, or strong-mixing.[12]

**Theorem 6.** *Let $(\Omega, \mathscr{F}, P)$ be a probability space, let $\mathscr{S}$ be a semialgebra that generates $\mathscr{F}$, and let $T : \Omega \to \Omega$ be a measure-preserving transformation.*

1. *$T$ is ergodic if and only if for all $A, B \in \mathscr{S}$,*

$$\frac{1}{n} \sum_{k=0}^{n-1} P(T^{-k}(A) \cap B) \to P(A)P(B).$$

2. *$T$ is weak-mixing if and only if for all $A, B \in \mathscr{S}$,*

$$\frac{1}{n} \sum_{k=0}^{n-1} |P(T^{-k}(A) \cap B) - P(A)P(B)| \to 0.$$

3. *$T$ is strong-mixing if and only if for all $A, B \in \mathscr{S}$,*

$$P(T^{-n}(A) \cap B) \to P(A)P(B).$$

[11] Peter Walters, *An Introduction to Ergodic Theory*, p. 37, Corollary 1.14.2.
[12] Peter Walters, *An Introduction to Ergodic Theory*, p. 41, Theorem 1.17.

# 4 The strong law of large numbers

Let $\mu$ be a Borel probability measure on $\mathbb{R}$ with **finite first moment**:

$$\int_{\mathbb{R}} |x| dm(x) < \infty.$$

We shall specify when we use the hypothesis that $\mu$ has finite first moment; until we say so, what we say merely supposes that it is a Borel probability measure on $\mathbb{R}$.

For $n \geq 0$, let $\Omega_n = \mathbb{R}$, let $\mathscr{B}_n = \mathscr{B}_{\mathbb{R}}$, the Borel $\sigma$-algebra of $\mathbb{R}$, and let $\mu_n = \mu$, for which $(\Omega_n, \mathscr{B}_n, \mu_n)$ is a probability space. Let $\Omega = \prod_{n=0}^{\infty} \Omega_n$. A **cylinder set** is a subset of $\Omega$ of the form

$$\prod_{n=0}^{\infty} A_n,$$

where $A_n \in \mathscr{B}_n$ for each $n$ and where $\{n \geq 0 : A_n \neq \Omega_n\}$ is finite. We denote by $\mathscr{C}$ the collection of all cylinder sets. It is a fact that $\mathscr{C}$ is a semialgebra of sets.[13]

The **product $\sigma$-algebra** is $\mathscr{F} = \sigma(\mathscr{C})$, the $\sigma$-algebra generated by the collection of all cylinder sets. The **product measure**,[14] which we denote by $P$, is the unique probability measure on $\mathscr{F}$ such that for any cylinder set $\prod_{n=0}^{\infty} A_n$,

$$P(\prod_{n=0}^{\infty} A_n) = \prod_{n=0}^{\infty} \mu_n(A_n).$$

Let $\pi_n : \Omega \to \Omega_n$, $n \geq 0$, be the projection map. Define $\tau : \Omega \to \Omega$ by

$$\tau(\omega_0, \omega_1, \ldots) = (\omega_1, \omega_2, \ldots),$$

the **left shift map**. In other words, for $n \geq 0$, $\tau$ satisfies $\pi_n \circ \tau = \pi_{n+1}$, and so $\pi_n = \pi_0 \circ (\tau^n)$.

**Lemma 7.** *$\tau$ is measure-preserving.*

*Proof.* For $A = \prod_{n=0}^{\infty} A_n \in \mathscr{C}$,

$$\tau^{-1}(A) = \prod_{n=0}^{\infty} B_n = B,$$

where $B_0 = \Omega_0$ and for $n \geq 1$, $B_n = A_{n-1}$. $B$ is a cylinder set so a fortiori belongs to $\mathscr{F}$, and

$$P(B) = \prod_{n=0}^{\infty} \mu_n(B_n) = \mu_0(\Omega_0) \cdot \prod_{n=1}^{\infty} \mu_n(B_n) = \prod_{n=1}^{\infty} \mu_n(A_{n-1}) = \prod_{n=1}^{\infty} \mu_{n-1}(A_{n-1}),$$

[13] S. J. Taylor, *Introduction to Measure Theory and Integration*, p. 136, §6.1; `http://individual.utoronto.ca/jordanbell/notes/productmeasure.pdf`

[14] `http://individual.utoronto.ca/jordanbell/notes/productmeasure.pdf`

so

$$P(\tau^{-1}(A)) = \prod_{n=0}^{\infty} \mu_n(A_n) = P(A).$$

Therefore by Lemma 5, because $\mathscr{C}$ is a semialgebra that generates $\mathscr{F}$, it follows that $\tau$ is measure-preserving. □

**Lemma 8.** *$\tau$ is strong-mixing.*

*Proof.* Let $A = \prod_{n=0}^{\infty} A_n$ and $B = \prod_{n=0}^{\infty} B_n$ be cylinder sets. For $n \geq 0$,

$$\tau^{-n}(A) = \prod_{m=0}^{\infty} C_m,$$

where $C_m = \Omega_m$ for $0 \leq m \leq n-1$ and $C_m = A_{m-n}$ for $m \geq n$. Because $A$ and $B$ are cylinder sets, there is some $N$ such that when $m \geq N$, $A_m = \Omega_m$ and $B_m = \Omega_m$. Thus for $n \geq N$,

$$\begin{aligned}
\tau^{-n}(A) \cap B &= \prod_{m=0}^{\infty} C_m \cap \prod_{m=0}^{\infty} B_m \\
&= \prod_{m=0}^{\infty} (C_m \cap B_m) \\
&= \prod_{m=0}^{n-1} (C_m \cap B_m) \times \prod_{m=n}^{\infty} (C_m \cap B_m) \\
&= \prod_{m=0}^{n-1} (\Omega_m \cap B_m) \times \prod_{m=n}^{\infty} (A_{m-n} \cap \Omega_m) \\
&= \prod_{m=0}^{n-1} B_m \times \prod_{m=n}^{\infty} A_{m-n}.
\end{aligned}$$

Hence

$$\begin{aligned}
P(\tau^{-n}(A) \cap B) &= \prod_{m=0}^{n-1} \mu_m(B_m) \cdot \prod_{m=n}^{\infty} \mu_m(A_{m-n}) \\
&= \prod_{m=0}^{n-1} \mu_m(B_m) \cdot \prod_{m=0}^{\infty} \mu_{m+n}(A_m) \\
&= \prod_{m=0}^{n-1} \mu_m(B_m) \cdot \prod_{m=0}^{\infty} \mu_m(A_m) \\
&= P(B) \cdot P(A).
\end{aligned}$$

That is, there is some $N$ such that when $n \geq N$,

$$P(\tau^{-n}(A) \cap B) = P(A)P(B),$$

and so a fortiori,

$$\lim_{n\to\infty} P(\tau^{-n}(A)\cap B) = P(A)P(B).$$

Therefore, because the cylinder sets generate the $\sigma$-algebra $\mathscr{F}$, by Theorem 3 we get that $\tau$ is strong-mixing. □

We now use the hypothesis that $\mu$ has finite first moment.[15]

**Lemma 9.** $\pi_0 \in L^1(P)$.

*Proof.* $T = \pi_0 : \Omega \to \Omega_0$ is measurable, and $T_*P = \mu_0$. The statement that $\mu = \mu_0$ has finite first moment means that $f : \Omega_0 \to \mathbb{R}$ defined by $f(x) = |x|$ belongs to $L^1(\mu_0)$. Therefore by the change of variables theorem (2), we have $f \circ T \in L^1(P)$. □

**Lemma 10.** *For almost all $\omega \in \Omega$,*

$$\frac{1}{n}\sum_{k=0}^{n-1} \pi_k(\omega) \to \int_{\mathbb{R}} t d\mu(t).$$

*Proof.* Because $\tau : \Omega \to \Omega$ is ergodic and $\pi_0 \in L^1(P)$, the **Birkhoff ergodic theorem**[16] tells us that for almost all $\omega \in \Omega$,

$$\frac{1}{n}\sum_{k=0}^{n-1} \pi_0(\tau^k(\omega)) \to \int_{\Omega} \pi_0(\omega) dP(\omega),$$

i.e.,

$$\frac{1}{n}\sum_{k=0}^{n-1} \pi_k(\omega) \to \int_{\Omega_0} \omega_0 d\mu_0(\omega_0),$$

proving the claim. □

We will use Lemma 10 to prove the strong law of large numbers. First we prove two lemmas about **joint distributions**.[17]

**Lemma 11.** *If $X_1, \ldots, X_n : \Omega \to \mathbb{R}$ and $Y_1, \ldots, Y_n : \Omega' \to \mathbb{R}$ are random variables with the same joint distribution and for each $1 \le k \le n$, $\Phi_k : \mathbb{R}^n \to \mathbb{R}$ is a continuous function, then for $U_k = \Phi_k(X_1, \ldots, X_n)$ and $V_k = \Phi_k(Y_1, \ldots, Y_n)$, the random variables $U_1, \ldots, U_n$ and $V_1, \ldots, V_n$ have the same joint distribution.*

[15] Elias M. Stein and Rami Shakarchi, *Functional Analysis*, Princeton Lectures in Analysis, volume IV, p. 208, chapter 5, §2.1.

[16] Peter Walters, *An Introduction to Ergodic Theory*, p. 34, Theorem 1.14.

[17] Elias M. Stein and Rami Shakarchi, *Functional Analysis*, Princeton Lectures in Analysis, volume IV, p. 208, Lemma 2.2, Lemma 2.3.

*Proof.* Write $X = (X_1, \ldots, X_n)$, $Y = (Y_1, \ldots, Y_n)$, $U = (U_1, \ldots, U_n)$, and $V = (V_1, \ldots, V_n)$, which are Borel measurable. Let $\Phi = (\Phi_1, \ldots, \Phi_n)$, which is continuous $\mathbb{R}^n \to \mathbb{R}^n$, and for which

$$U = \Phi(X), \qquad V = \Phi(Y).$$

To show that $U_1, \ldots, U_n$ and $V_1, \ldots, V_n$ have the same joint distribution means to show that $U_*P = V_*P'$. Let $A \in \mathscr{B}_{\mathbb{R}^n}$, for which $\Phi^{-1}(A) \in \mathscr{B}_{\mathbb{R}^n}$ and

$$\begin{aligned}(U_*P)(A) &= P(U^{-1}(A)) \\ &= P(X^{-1}(\Phi^{-1}(A))) \\ &= P'(Y^{-1}(\Phi^{-1}(A))) \\ &= P'(V^{-1}(A)) \\ &= (V_*P')(A),\end{aligned}$$

where, because $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ have the same joint distribution,

$$P(X^{-1}(\Phi^{-1}(A))) = (X_*P)(\Phi^{-1}(A)) = (Y_*P')(\Phi^{-1}(A)) = P'(Y^{-1}(\Phi^{-1}(A))).$$

□

**Lemma 12.** *If sequences of random variables $X_n : (\Omega, \mathscr{F}, P) \to \mathbb{R}$ and $Y_n : (\Omega', \mathscr{F}', P') \to \mathbb{R}$, $n \geq 1$, have the same **finite-dimensional distributions** and there is some $a \in \mathbb{R}$ such that $X_n$ converges to $a$ almost surely, then $Y_n$ converges to $a$ almost surely.*

*Proof.* That the sequences $X_n$ and $Y_n$ have the same finite-dimensional distributions means that for each $n \geq 1$,

$$(X_1, \ldots, X_n)_*P = (Y_1, \ldots, Y_n)_*P',$$

i.e., for each $n \geq 1$ and for each $A \in \mathscr{B}_{\mathbb{R}^n}$,

$$P((X_1, \ldots, X_n) \in A) = P'((Y_1, \ldots, Y_n) \in A).$$

Define

$$\begin{aligned}E_{k,N,n} &= \left\{\omega \in \Omega : |X_n(\omega) - a| \leq \frac{1}{k}\right\} \\ F_{k,N,n} &= \left\{\omega \in \Omega' : |Y_n(\omega) - a| \leq \frac{1}{k}\right\}.\end{aligned}$$

Then define

$$\begin{aligned}E &= \bigcap_{k=1}^{\infty} \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} E_{k,N,n} = \bigcap_{k=1}^{\infty} \bigcup_{N=1}^{\infty} E_{k,N} = \bigcap_{k=1}^{\infty} E_k \\ F &= \bigcap_{k=1}^{\infty} \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} F_{k,N,n} = \bigcap_{k=1}^{\infty} \bigcup_{N=1}^{\infty} F_{k,N} = \bigcap_{k=1}^{\infty} F_k,\end{aligned}$$

and

$$G_{k,N,n} = \bigcap_{m=N}^{n} E_{k,N,m}$$
$$H_{k,N,n} = \bigcap_{m=N}^{n} F_{k,N,m}.$$

Because $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_n)$ have the same distribution,

$$P(G_{k,N,n}) = P'(H_{k,N,n}).$$

But

$$E_{k,N} = \bigcap_{n=N}^{\infty} E_{k,N,n} = \bigcap_{n=N}^{\infty} G_{k,N,n} = \lim_{n\to\infty} G_{k,N,n}$$

and

$$F_{k,N} = \lim_{n\to\infty} H_{k,N,n},$$

so

$$P(E_{k,N}) = P'(F_{k,N}).$$

Then

$$P(E_k) = \lim_{N\to\infty} P(E_{k,N}) = \lim_{N\to\infty} P'(F_{k,N}) = P'(F_k).$$

Then

$$P(E) = \lim_{k\to\infty} P(E_k) = \lim_{k\to\infty} P'(F_k) = P'(F).$$

That the sequence $X_n$ converges almost surely means that $P(E) = 1$, and therefore $P'(F) = 1$, i.e. the sequence $Y_n$ converges almost surely. □

We now use what we have established to prove the **strong law of large numbers**.[18] (We write $(\Omega', \mathscr{F}', P')$ for a probability space because $(\Omega, \mathscr{F}, P)$ denotes the product probability space constructed already.)

**Theorem 13.** *If $X_n : (\Omega', \mathscr{F}', P') \to \mathbb{R}$, $n \geq 1$, are independent identically distributed $L^1$ random variables, then for almost all $\omega \in \Omega'$,*

$$\frac{1}{n}\sum_{k=1}^{n} X_k(\omega) \to E(X_1).$$

*Proof.* For $n \geq 0$, let $Y_n = X_{n+1}$; we do this to make the index set the same as for $\Omega_n$. Let $\mu_n = Y_{n*}P'$ and set $\mu = \mu_0$. Because the $Y_n$ are identically distributed, $\mu_n = \mu_0$ for each $n$.

Because $Y_0$ is $L^1$, $\mu$ has finite first moment. As $\mu_0 = Y_{0*}P'$, applying the change of variables theorem (2),

$$\int_{\mathbb{R}} t d\mu(t) = \int_{\mathbb{R}} t d(Y_{0*}P')(t) = \int_{\Omega'} Y_0(\omega) dP'(\omega) = E(Y_0).$$

[18] Elias M. Stein and Rami Shakarchi, *Functional Analysis*, Princeton Lectures in Analysis, volume IV, p. 206, Theorem 2.1.

With this, Lemma 10 says that for almost all $\omega \in \Omega$,

$$\frac{1}{n}\sum_{k=0}^{n-1}\pi_k(\omega) \to E(Y_0). \tag{3}$$

For each $n$,

$$(\pi_0, \ldots, \pi_n)_* P = \mu_0 \times \cdots \times \mu_n,$$

and because the $Y_n$ are independent,

$$(Y_0, \ldots, Y_n)_* P' = Y_{0*}P' \times \cdots \times Y_{n*}P' = \mu_0 \times \cdots \times \mu_n,$$

so the sequences $\pi_n$ and $Y_n$ have the same finite-dimensional distributions. For $n \geq 1$, define $\Phi_n : \mathbb{R}^n \to \mathbb{R}$ by

$$\Phi_n(t_1, \ldots, t_n) = \frac{1}{n}\sum_{k=1}^{n} t_k.$$

Lemma 11 then tells us that the sequences

$$U_n = \Phi_n(Y_0, \ldots, Y_{n-1}) = \frac{1}{n}\sum_{k=0}^{n-1} Y_k$$

and

$$V_n = \Phi_n(\pi_0, \ldots, \pi_{n-1}) = \frac{1}{n}\sum_{k=0}^{n-1} \pi_k,$$

$n \geq 1$, have the same finite-dimensional distributions.

Now, (3) says that $V_n$ converges to $E(Y_0)$ almost surely, and thus applying Lemma 12 we get that $U_n$ converges to $E(Y_0)$ almost surely. That is,

$$\frac{1}{n}\sum_{k=0}^{n-1} Y_k \to E(Y_0)$$

almost surely, and because $Y_k = X_{k+1}$,

$$\frac{1}{n}\sum_{k=1}^{n} X_k \to E(X_1)$$

almost surely, completing the proof. □