8 The Laws of Large Numbers

The traditional interpretation of the \textit{probability} of an event \( E \) is its \textit{asymptotic frequency}: the limit as \( n \to \infty \) of the fraction of \( n \) repeated, similar, and independent trials in which \( E \) occurs. Similarly the “expectation” of a random variable \( X \) is taken to be its \textit{asymptotic average}, the limit as \( n \to \infty \) of the average of \( n \) repeated, similar, and independent replications of \( X \). For statisticians trying to make inference about the underlying probability distribution \( f(x \mid \theta) \) governing observed random variables \( X_i \), this suggests that we should be interested in the probability distribution for large \( n \) of quantities like the sample average of the RVs, \( \bar{X}_n := \frac{1}{n} \sum_{i=1}^{n} X_i \).

Three of the most celebrated theorems of probability theory concern this quantity. For independent random variables \( X_i \), all with the same probability distribution satisfying \( \mathbb{E}|X_i|^3 < \infty \), set \( \mu := \mathbb{E}X_i \), \( \sigma^2 := \mathbb{E}|X_i - \mu|^2 \), and \( S_n := \sum_{i=1}^{n} X_i \). The three main results are:

\textbf{Laws of Large Numbers (LLN):}
\[
\frac{S_n - n\mu}{\sigma n} \rightarrow 0 \quad \text{(pr. and a.s.)}
\]

\textbf{Central Limit Theorem (CLT):}
\[
\frac{S_n - n\mu}{\sigma \sqrt{n}} \Rightarrow \text{N}(0, 1) \quad \text{(in dist.)}
\]

\textbf{Law of the Iterated Logarithm (LIL):}
\[
\limsup_{n \to \infty} \frac{S_n - n\mu}{\sigma \sqrt{2n \log \log n}} = 1.0 \quad \text{(a.s.)}
\]

Together these three give a clear picture of how quickly and in what sense \( \frac{1}{n} S_n \) tends to \( \mu \). We begin with the Law of Large Numbers (LLN), first in its “weak” form (asserting convergence \( \text{pr.} \)) and then in its “strong” form (convergence \( \text{a.s.} \)). There are several versions of both theorems. The simplest requires the \( X_i \) to be IID and \( L_2 \); stronger results allow us to weaken (but not eliminate) the independence requirement, permit non-identical distributions, and consider what happens if we relax the \( L_3 \) requirement and allow the RVs to be only \( L_2 \) or \( L_1 \) (or worse!).

The text covers these things well; to complement it I am going to: (1) Prove the simplest version, and with it the Borel-Cantelli theorems; and (2) Show what happens with Cauchy random variables, which don’t satisfy the requirements (the LLN fails).

8.1 Proofs of the Weak and Strong Laws

Here are two simple versions (one Weak, one Strong) of the Law of Large Numbers; first we prove an elementary but very useful result:

\textbf{Proposition 1 (Markov’s Inequality)} Let \( \phi(x) \geq 0 \) be non-decreasing on \( \mathbb{R}_+ \). For any random variable \( X \geq 0 \) and constant \( a \in \mathbb{R}_+ \),
\[
P[X \geq a] \leq P[\phi(X) \geq \phi(a)] \leq \mathbb{E}[\phi(X)]/\phi(a)
\]
To see this, set $Y := \phi(a) 1_A$ for the event $A := \{ \phi(X) \geq \phi(a) \}$ and note $Y \leq \phi(X)$ so $EY \leq E\phi(X)$.

**Theorem 1 (L2 WLLN)** Let $\{X_n\}$ be independent random variables with the same mean $\mu = E[X_n]$ and uniformly bounded variance $E(X_n - \mu)^2 \leq B$ for some fixed bound $B < \infty$. Set $S_n := \sum_{j \leq n} X_j$ and $\bar{X}_n := S_n/n = \frac{1}{n} \sum_{j \leq n} X_j$. Then:

$$E(S_n - n\mu)^2 = \sum_{i=1}^{n} E(X_i - \mu)^2 \leq nB$$

so for $\epsilon > 0$

$$P[|\bar{X}_n - \mu| > \epsilon] = P[(S_n - n\mu)^2 > (n\epsilon)^2] \leq E[(S_n - n\mu)^2]/n^2 \epsilon^2 \leq B/n \epsilon^2 \to 0 \textup{ as } n \to \infty.$$

This Law of Large Numbers is called weak because its conclusion is only that $\bar{X}_n$ converges to zero in probability (Eqn (1)); the strong Law of Large Numbers asserts convergence of a stronger sort, called almost sure convergence (Eqn (2) below). If $P[|\bar{X}_n - \mu| > \epsilon]$ were summable then by B-C we could conclude almost-sure convergence; unfortunately we have only the bound $P[|\bar{X}_n - \mu| > \epsilon] < c/n$ which tends to zero but isn’t summable. It is summable along the subsequence $n^2$, however; our approach to proving a strong LLN is to show that $|S_k - S_n|/n$ isn’t too big for any $n^2 \leq k < (n+1)^2$.

**Theorem 2 (L2 SLLN)** Under the same conditions,

$$P[\bar{X}_n \to \mu] = 1.$$  

**Proof.** Without loss of generality take $\mu = 0$ (otherwise subtract $\mu$ from each $X_n$), and fix $\epsilon > 0$. Set $S_n := \sum_{j \leq n} X_j$. Then

$$P[|S_n| > n\epsilon] \leq E|S_n|^2/(n\epsilon)^2 \leq nB/n^2 \epsilon^2 = B/n^2$$

$P[|S_n| > n^2 \epsilon \ i.o.] = 0$ by B-C $\Rightarrow S_n/n \to 0$ a.s. Set

$$D_n := \max_{n^2 \leq k < (n+1)^2} |S_k - S_{n^2}|$$

$$E D_n^2 = E \left[ \max_{n^2 \leq k < (n+1)^2} |S_k - S_{n^2}|^2 \right] \leq E \sum_{n^2 \leq k < (n+1)^2} |S_k - S_{n^2}|^2 = \sum_{n^2 \leq k < (n+1)^2} E|S_k - S_{n^2}|^2 \leq \sum_{n^2 \leq k < (n+1)^2} (k - n^2)B \leq 4n^2 B,$$

so

$$P[D_n > n^2 \epsilon] \leq 4n^2 B/n^4 \epsilon^2 \Rightarrow D_n/n^2 \to 0 \textup{ a.s.}$$

$$\frac{|S_k|}{k} \leq \frac{|S_{n^2}| + D_n}{n^2} \to 0 \textup{ a.s. \ as } k \to \infty, \textup{ where } n = \lfloor \sqrt{k} \rfloor.$$
Each of these LLNs required only that $\text{Cov}(X_n, X_m) \leq 0$, not pairwise (let alone full) independence. We’ll see below in Section (8.3) that even positive correlations are okay if they fall off fast enough — e.g., if $|\text{Cov}(X_n, X_m)| \leq ar^n$ for some $a > 0$, $0 < r < 1$ — with a similar proof. The uniform $L_2$ bound isn’t necessary either. There are a variety of LLNs with either or both of the $L_2$ bound and independence weakened in some way, but they can’t be dispensed with altogether — consider iid Cauchy random variables, for example, to show $L_2$ isn’t entirely superfluous, or $X_n \equiv X_1$ with any nontrivial distribution to show the need for at least a modicum of independence.

8.2 Other Strong Laws

**Lemma 1 (Lévy)** If $\{X_n\}$ is an independent sequence then $\sum_{n=1}^{\infty} X_n$ converges pr. if and only if it converges a.s.

**Lemma 2 (Kronecker)** Suppose $\{x_n\} \subset \mathbb{R}$ and $0 < a_n \not\to \infty$. Then

$$\sum_{k=1}^{\infty} \frac{x_k}{a_k} \text{ converges } \Rightarrow \frac{1}{a_n} \sum_{k=1}^{n} x_k \to 0.$$

**Theorem 3 (Kolmogorov Convergence Criterion)** Let $\{X_n\} \subset L_2$ be an independent sequence with means $\mu_n := \mathbb{E}X_n$ and variances $\sigma_n^2 := \mathbb{V}(X_n)$. Then

$$\sum_{n=1}^{\infty} \sigma_n^2 < \infty \Rightarrow \sum_{n=1}^{\infty} (X_n - \mu_n) \text{ converges a.s.}$$

**Theorem 4 (Kronecker $L_2$ SLLN)** Let $\{X_n\} \subset L_2$ be an independent sequence, and let $\{b_n\} \subset \mathbb{R}_+$ be a monotone sequence such that

$$\sum_{k} \mathbb{V} \left( \frac{X_k}{b_k} \right) < \infty.$$

Then

$$\frac{S_n - \mathbb{E}S_n}{b_n} \to 0 \text{ a.s.}$$

In particular, for iid $\{X_n\}$ we may take $b_n = n$.

**Lemma 3** The following are equivalent for any iid sequence of random variables:

1. $X_1 \in L_1$.
2. $\lim_{n \to \infty} |X_n|/n = 0$ a.s.
3. For each $\epsilon > 0$, $\sum_{n=1}^{\infty} P(|X_1| > n\epsilon) < \infty$. 

Page 3
The most-cited and most-used version of the Strong Law for iid sequences is that due to Kolmogorov, with no moment assumptions:

**Theorem 5 (Kolmogorov’s L1 SLLN)** Let \( \{X_n\} \) be iid and set \( S_n := \sum_{i \leq n} X_i \). There exists \( c \in \mathbb{R} \) such that

\[
\bar{X}_n = \frac{S_n}{n} \to c \quad \text{a.s.}
\]

if and only if \( \mathbb{E}|X_1| < \infty \), in which case \( c = \mathbb{E}X_1 \).

This has the pleasant consequence that the usual estimators for the mean and variance of iid sequences are consistent:

**Corollary 1**

\[
\{X_n\} \subset L_1 \implies \bar{X}_n \to \mu \quad \text{a.s.}
\]

\[
\{X_n\} \subset L_2 \implies \frac{1}{n} \sum_{i \leq n} (X_n - \bar{X}_n)^2 \to \sigma^2 \quad \text{a.s.}
\]
Here’s a quick summary of some LLN facts:

I. Weak version, non-iid, $L_2$: $\mu_i = \text{EX}_i$, $\sigma_{ij} = \text{E}[X_i - \mu_i][X_j - \mu_j]$
   A. $Y_n = (S_n - \Sigma \mu_i)/n$ satisfies $\text{E}Y_n = 0$, $\text{E}Y_n^2 = \frac{1}{n} \Sigma \sigma_{ii} + \frac{1}{n^2} \Sigma_{i<j\leq n} \sigma_{ij}$;
      1. If $\sigma_{ii} \leq M$ and $\sigma_{ij} \leq 0$ or $|\sigma_{ij}| < Mr^{1/2 - 1}$, $r < 1,$
         Chebychev $\implies Y_n \to 0, \text{ pr.}$
      2. (pairwise) IID $L_2$ is OK

II. Strong version, non-iid, $L_2$: $\text{EX}_i = 0$, $\text{EX}_i^2 \leq M$, $\text{EX}_i X_j \leq 0$.
   A. $\text{P}[|S_n| > ae] < \frac{Mn}{n^2 a^2} = \frac{M}{na^2}$
      1. $\text{P}[|S_{n^2}| > n^2 e] < \frac{M}{n^2 a^2}$, $\Sigma_n \text{P}[|S_{n^2}| > n^2 e] < \frac{4\pi^2 M}{6a^2} < \infty$
      2. Borel-Cantelli: $\text{P}[|S_{n^2}| > n^2 e \text{ i. a.}] = 0$, $\therefore \frac{1}{n^2} S_{n^2} \to 0 \text{ a.s.}$
      3. $D_n \equiv \max_{a < (n+1)^2} |S_k - S_{n^2}|$, so
         $\text{E}D_n^2 \leq 2n\text{E}[S_{(n+1)^2 - 1} - S_{n^2}] \leq 4n^2 M$
      4. Chebychev: $\text{P}[D_n > n^2 e] < \frac{4n^2 M}{n^2 a^2}$, $\therefore D_n/n^2 \to 0 \text{ a.s}$
   B. $|S_k/k| \leq \frac{\sqrt{2\log n}}{n^2} \to 0 \text{ a.s as } k \to \infty$, QED
      1. Bernoulli RVs, normal number theorem, Monte Carlo integration.

III. Weak version, pairwise-iid, $L_1$
   A. Equivalent sequences: If $\sum_n \text{P}[X_n \neq Y_n] < \infty$, then:
      1. $\sum_n |X_n - Y_n| < \infty \text{ a.s}$
      2. $\sum_{i=1}^n X_i$ converges iff $\sum_{i=1}^n Y_i$ converges
      3. If $(\exists a_n \nearrow) (3X) \ni \frac{1}{a_n} \sum_{i\leq n} X_i \to X$ then $\frac{1}{a_n} \sum_{i\leq n} Y_i \to X$ too.
   B. If $\{X_n\}$ are iid $L_1$ then $X_n, Y_n := X_n 1_{|X_n| \leq n}$ are equivalent

IV. Strong version, iid, $L_1$
   A. Kolmogorov: For $\{X_n\}$ IID, $(\exists c \ni \tilde{X}_n \to c \text{ a.s.}) \Leftrightarrow (X_n \in L_1)$.
   V. Counterexamples: Cauchy,
   A. $X_i \sim \frac{dx}{\pi [1 + x^2]} \implies \text{P}[|S_n|/n \leq \epsilon] \equiv \frac{2}{\pi} \tan^{-1}(\epsilon) \not\to 1$, WLLN fails.
   B. $\text{P}[X_i = \pm n] = \frac{\epsilon}{\pi n^2}, n \geq 1$; $X_i \notin L_1$, and $S_n/n \not\to 0 \text{ pr. or a.s.}$
   C. $\text{P}[X_i = \pm n] = \frac{\epsilon}{\pi n^2 \log n}, n > 1$; $X_i \notin L_1$, but $S_n/n \to 0 \text{ pr. and not a.s.}$
   D. Medians: for ANY RVs $X_n \to X_\infty \text{ pr.}$, then $m_n \to m_\infty$ if $m_\infty$ is unique.
Let $X_i$ be iid standard Cauchy RVs, with

$$P[X_1 \leq t] = \int_{-\infty}^{t} \frac{dx}{\pi [1 + x^2]} = \frac{1}{2} + \frac{1}{\pi} \arctan(t)$$

and characteristic function (we’ll learn more about these next week)

$$E e^{i\omega X_1} = \int_{-\infty}^{\infty} e^{i\omega x} \frac{dx}{\pi [1 + x^2]} = e^{-|\omega|},$$

so $S_n/n$ has characteristic function

$$E e^{i\omega S_n/n} = E e^{i\hat{\omega}[X_1+\cdots+X_n]} = \left(E e^{i\hat{\omega}X_1}\right)^n = (e^{-|\hat{\omega}|})^n = e^{-|\omega|}.$$

Thus $S_n/n$ also has the standard Cauchy distribution with $P[S_n/n \leq t] = \frac{1}{2} + \frac{1}{\pi} \arctan(t)$; in particular, $S_n/n$ does not converge almost surely, or even in probability.

### 8.3 An LLN for Correlated Sequences

In many applications we would like a Law of Large Numbers for sequences of random variables that are not independent. For example, in Markov Chain Monte Carlo integration, we have a stationary Markov chain $\{X_t\}$ (this means that the distribution of $X_t$ is the same for all $t$ and that the conditional distribution of $X_u$ for $u > t$, given $\{X_s \mid s \leq t\}$, depends only on $X_t$) and want to estimate the population mean $E[\phi(X_t)]$ for some function $\phi(\cdot)$ by the sample mean

$$E[\phi(X_t)] \approx \frac{1}{T} \sum_{t=0}^{T-1} \phi(X_t).$$

Even though they are identically distributed, the random variables $Y_t := \phi(X_t)$ won’t be independent if the $X_t$ aren’t independent, so the LLN we already have doesn’t quite apply.

**Theorem 6** If an $L_2$ sequence $\{X_t\}$ has a constant mean $\mu = E X_t$ and a summable autocovariance $\gamma_{st} := E(X_s - \mu)(X_t - \mu)$ that satisfies $\sum_{t=-\infty}^{\infty} |\gamma_{st}| \leq c < \infty$ uniformly in $s$, then they obey a LLN:

$$E[Y_t] = \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} Y_t.$$

**Proof.** Let $S_T$ be the sum of the first $T$ $Y_t$s and (as usual) set $\bar{Y}_T := S_T/T$. The variance of $S_T$ is

$$E[(S_T - T\mu)^2] = \sum_{s=0}^{T-1} \sum_{t=0}^{T-1} E[(X_s - \mu)(X_t - \mu)]$$

$$\leq \sum_{s=0}^{T-1} \sum_{t=-\infty}^{\infty} |\gamma_{st}| \leq Tc,$$
so \( \hat{Y}_T \) has variance \( \text{Var}[\hat{Y}_T] \leq c/T \) and by Chebychev’s inequality

\[
P[|\hat{Y}_T - \mu| > \epsilon] \leq \frac{\text{E}[(\hat{Y}_T - \mu)^2]}{\epsilon^2} = \frac{\text{E}[(S_T - T\mu)^2]}{T^2\epsilon^2} \leq \frac{Tc}{T^2\epsilon^2} = \frac{c}{T\epsilon^2} \to 0 \quad \text{as } T \to \infty.
\]

As in the iid case, this WLLN can be extended to a SLLN by applying it along the sequence \( \{T^2 : T \in \mathbb{N}\} \), applying B/C, then filling in the gaps.

A sequence of random variables \( Y_t \) is called stationary if each \( Y_t \) has the same probability distribution and, moreover, each finite set \( (Y_{t_1+h}, Y_{t_2+h}, \cdots, Y_{t_k+h}) \) has a joint distribution that doesn’t depend on \( h \). The sequence is called “L2” if each \( Y_t \) has a finite variance \( \sigma^2 \) (and hence also a well-defined mean \( \mu \)); by stationarity it also follows that the covariance

\[
\gamma_{st} = \text{E}[(Y_s - \mu)(Y_t - \mu)]
\]

is finite and depends only on the absolute difference \( |t - s| \) (write: \( \gamma_{st} = \gamma(s - t) = \gamma(t - s) \)).

**Corollary 2** If a stationary L2 sequence \( \{Y_t\} \) has a summable covariance function, i.e., satisfies \( \sum_{t=-\infty}^{\infty} |\gamma(t)| \leq c < \infty \), then

\[
\text{E}[Y_t] = \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} Y_t
\]

Note we didn’t need full stationarity. It would be good enough for \( \{Y_t\} \) to be “2nd-order stationary,” i.e., to have a common mean \( \mu = \text{E}Y_t \) and a covariance function \( \gamma_{st} = \text{E}(Y_s - \mu)(Y_t - \mu) = \gamma(s - t) \) that depends only on the absolute difference \( |s - t| \).

### 8.3.1 Examples

1. **IID**: If \( X_t \) are independent and identically distributed, and if \( Y_t = \phi(X_t) \) has finite variance \( \sigma^2 \), then \( Y_t \) has a well-defined finite mean \( \mu \) and \( \hat{Y}_T \to \mu \).

   Here \( \gamma_{st} = \begin{cases} \sigma^2 & \text{if } s = t \\ 0 & \text{if } s \neq t \end{cases} \), so \( c = \sigma^2 < \infty \).

2. **AR**: If \( Z_t \) are iid \( \text{No}(0, 1) \) for \(-\infty < t < \infty, \mu \in \mathbb{R}, \sigma > 0, -1 < \rho < 1, \) and

\[
X_t := \mu + \sigma \sum_{s=0}^{\infty} \rho^s Z_{t-s} = \rho X_{t-1} + \alpha + \sigma Z_t,
\]

where \( \alpha = (1 - \rho)\mu \), then the \( X_t \) are identically distributed (all with the \( \text{No}(\mu, \frac{\sigma^2}{1-\rho^2}) \) distribution) but not independent (since \( \gamma_{st} = \frac{\sigma^2}{1-\rho^2} \rho^{|s-t|} \neq 0 \) for \( s \neq t \)). This is called an “autoregressive process” (because of equation \((*)\)), expressing \( X_t \) as a regression of previous
\(X_{\text{st}}\) of order one (because only one earlier \(X_i\) appears in (*)), and is about the simplest non-\(iid\) sequence occurring in applications. Since the covariance is summable,
\[
\sum_{t=-\infty}^{\infty} |\gamma(t)| = \frac{\sigma^2}{1 - \rho^2} < \infty,
\]
we again have \(\hat{X}_T \rightarrow \mu \) a.s. as \(T \rightarrow \infty\).

3. Geometric Ergodicity: If for some \(0 < \rho < 1\) and \(c > 0\) we have \(\gamma_{st} \leq c|t-s|\) for a Markov chain \(Y_t\) the chain is called Geometrically Ergodic (because \(c\rho^t\) is a geometric sequence), and the same argument as for AR(1) shows that \(\hat{Y}_t\) converges; Meyn & Tweedie (1993), Tierney (1994), and others have given conditions for MCMC chains to be Geometric Ergodic, and hence for the almost sure convergence of sample averages to population means. The mean squared error \(\text{MSE} := E[\hat{Y}_t - \mu]^2\) for a geometrically ergodic sequence is bounded by \(\text{MSE} \leq \frac{c^2}{1 - \rho^2}(\sigma^2/t) \asymp 1/t\), but the constant grows without bound as \(\rho \rightarrow 1\). Irreducible aperiodic finite-state Markov chains are geometrically ergodic, with \(\rho\) the second-largest eigenvalue of the one-step transition matrix.

4. General Ergodicity: Consider the three sequences of random variables on \((\Omega, \mathcal{F}, P)\) with \(\Omega = (0, 1]\) and \(\mathcal{F} = \mathcal{B}(\Omega)\), each with \(X_0(\omega) = \omega\):

(a) \(X_{n+1} := 2X_n \) (mod 1);
(b) \(X_{n+1} := X_n + \alpha \) (mod 1) (Does it matter if \(\alpha\) is rational?);
(c) \(X_{n+1} := 4X_n(1 - X_n)\).

For each there exists a probability measure \(P\) (a distribution for \(X_0\)) such that the \(X_n\) are identically distributed— the \(\text{Un}(0,1)\) distribution for (a) and (b), and the “arcsin” distribution \(\text{Be}(\frac{1}{2}, \frac{1}{2})\) for (c), and each is of the form \(X_n = X_0(T^n(\omega))\) for a measure-preserving transformation \(T : \Omega \rightarrow \Omega\), a measurable mapping \(T : \Omega \rightarrow \Omega\) for which \(P(E) = P(T^{-1}(E))\) for each \(E \in \mathcal{F}\). Such a transformation is called ergodic if the only \(T\)-invariant events (those that satisfy \(E = T^{-1}(E)\)) are “almost trivial” in the sense that \(P[E] = 0\) or \(P[E] = 1\). A sequence \(X_n := X_0 \circ T^n\) is called ergodic if \(T\) is. Each of the sequences (a)–(c) is ergodic (can you show that?).

Birkhoff’s Ergodic Theorem asserts that, if \(X_n \in L_1(\Omega, \mathcal{F}, \mathbb{P})\) is an integrable ergodic sequence, then \(\bar{X}_n\) converges almost surely to a \(T\)-invariant limit \(X_\infty\) as \(n \rightarrow \infty\). Since only constants are \(T\)-invariant for ergodic sequences, it follows that \(\bar{X}_n \rightarrow \mu := \mathbb{E}X_n\) a.s. for ergodic sequences. The conditions here are weaker than those for the usual LLN; in all three cases above, for example, all ergodic, each \(X_n\) is completely determined by \(X_0\) so there is complete dependence, with \(\sigma(X_n) \subset \sigma(X_m)\) for all \(m \leq n\).

For any \(L_1\) distribution \(\mu(dx)\) on \((\mathbb{R}, \mathcal{B})\), we can construct iid random variables \(\{X_n\} \overset{iid}{\sim} \mu(dx)\) on the product probability space \((\Omega = \mathbb{R}^\infty, \mathcal{F} = \mathcal{B}^\infty, \mathbb{P} = \otimes \mu)\) and a measure-preserving transformation \(T : \Omega \rightarrow \Omega\) called the left-shift by
\[
X_n(\omega_1, \omega_2, \omega_3, \cdots) := \omega_n, \quad T(\omega_1, \omega_2, \omega_3, \cdots) := (\omega_2, \omega_3, \omega_4, \cdots).
\]
The \(\sigma\)-algebra \(\mathcal{T} = \{A : A = T^{-1}(A)\}\) of \(T\)-invariant sets is just the tail \(\sigma\)-algebra for the independent random variables \(\{X_n\}\), so by Kolmogorov’s zero-one law \(\mathcal{T}\) is almost-trivial.
and so $T$ is ergodic. It follows from Birkhoff’s Ergodic Theorem that the sample mean $\bar{X}_n := \frac{1}{n} \sum_{j=1}^{n} X_j$ converges almost-surely to a $T$-invariant and hence almost-surely constant random variable whose value must be $\mu$, proving a strong LLN for iid random variables that assumes only $L_1$:

**Theorem 7 (L₁ iid SLLN)** Let $\{X_n\}$ be iid $L_1(\Omega, \mathcal{F}, P)$ random variables with mean $\mu = E[X_n]$. Set $S_n := \sum_{j \leq n} X_j$ and $\bar{X}_n := S_n/n = \frac{1}{n} \sum_{j \leq n} X_j$. Then:

$$P[\bar{X}_n \to \mu] = 1.$$ 

Thus the “space average” $\int_{\mathbb{R}} X_n \, d\mu$ and the limiting “time average” $\lim_{n \to \infty} \bar{X}_n$ coincide for ergodic sequences.