STA 711: Probability & Measure Theory
Robert L. Wolpert

8 The Laws of Large Numbers

The traditional interpretation of the *probability* of an event E is its *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 *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 average of the RVs, $\bar{X}_n \equiv \frac{1}{n}\sum_{i=1}^n X_i$.

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

Laws of Large Numbers:

$$\frac{S_n - n\mu}{\sigma n} \longrightarrow 0 \qquad (pr. \text{ and } a.s.)$$

Central Limit Theorem:

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \Longrightarrow \mathsf{No}(0,1) \qquad (\text{in dist.})$$

Law of the Iterated Logarithm:

$$\limsup_{n\to\infty} \pm \frac{S_n - n\mu}{\sigma\sqrt{2n\log\log n}} = 1.0 \qquad (a.s.)$$

Together these three give a clear picture of how quickly and in what sense $\frac{1}{n}S_n$ tends to μ. We begin with the Law of Large Numbers (LLN), first in its "weak" form (asserting convergence *pr.*) and then in its "strong" form (convergence *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:

Proposition 1 (Markov's Inequality) *Let $\phi(x) \ge 0$ be non-decreasing on $\mathbb{R}_+$. For any random variable $X \ge 0$ and constant $a \in \mathbb{R}_+$,*

$$\mathsf{P}[X \ge a] \le \mathsf{P}[\phi(X) \ge \phi(a)] \le \mathsf{E}[\phi(X)]/\phi(a)$$

To see this, set $Y := \phi(a)\mathbf{1}_A$ for the event $A := \{\phi(X) \geq \phi(a)\}$ and note $Y \leq \phi(X)$ so $\mathsf{E}Y \leq \mathsf{E}\phi(X)$.

Theorem 1 **(L_2 WLLN)** *Let $\{X_n\}$ be independent random variables with the same mean $\mu = \mathsf{E}[X_n]$ and uniformly bounded variance $\mathsf{E}(X_n-\mu)^2 \leq B$ 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:*

$$(\forall \epsilon > 0) \quad \mathsf{P}[|\bar{X}_n - \mu| > \epsilon] \to 0. \tag{1}$$

Proof.

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

so for $\epsilon > 0$

$$\begin{aligned} \mathsf{P}[|\bar{X}_n - \mu| > \epsilon] &= \mathsf{P}[(S_n - n\mu)^2 > (n\epsilon)^2] \\ &\leq \mathsf{E}[(S_n-n\mu)^2]/n^2\epsilon^2 \\ &\leq B/n\,\epsilon^2 \to 0 \quad \text{as } n \to \infty. \end{aligned}$$

∎

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 $\mathsf{P}[|\bar{X}_n - \mu| > \epsilon]$ were *summable* then by B-C we could conclude almost-sure convergence; unfortunately we have only the bound $\mathsf{P}[|\bar{X}_n - \mu| > \epsilon] < c/n$ 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 $|S_k - S_{n^2}|$ isn't too big for any $n^2 \leq k < (n+1)^2$.

Theorem 2 **(L_2 SLLN)** *Under the same conditions,*

$$\mathsf{P}[\bar{X}_n \to \mu] = 1. \tag{2}$$

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

$$\begin{aligned} \mathsf{P}[|S_n| > n\epsilon] &\leq \mathsf{E}|S_n|^2/(n\epsilon)^2 \leq nB/n^2\epsilon^2 = B/n\epsilon^2 \\ \mathsf{P}[|S_{n^2}| > n^2\epsilon \ \text{ i.o.}] &= 0 \text{ by B-C } \Rightarrow S_{n^2}/n^2 \to 0 \text{ a.s. Set} \\ D_n &:= \max_{n^2 \leq k < (n+1)^2} |S_k - S_{n^2}| \\ \mathsf{E}D_n^2 &= \mathsf{E}\left[\max_{n^2 \leq k < (n+1)^2} |S_k - S_{n^2}|^2\right] \\ &\leq \mathsf{E}\sum_{n^2 \leq k < (n+1)^2} |S_k - S_{n^2}|^2 = \sum_{n^2 \leq k < (n+1)^2} \mathsf{E}|S_k - S_{n^2}|^2 \\ &\leq \sum_{n^2 \leq k < (n+1)^2} (k-n^2)B \leq 4n^2 B\ , \text{ so} \\ \mathsf{P}[D_n > n^2\epsilon] &\leq 4n^2B/n^4\,\epsilon^2 \Rightarrow D_n/n^2 \to 0 \text{ a.s.} \\ \left|\frac{S_k}{k}\right| &\leq \frac{|S_{n^2}| + D_n}{n^2} \to 0 \text{ a.s. as } k \to \infty, \text{ where } n = \lfloor \sqrt{k} \rfloor. \end{aligned}$$

□

Each of these LLNs required only that $\mathsf{Cov}(X_n, X_m) \le 0$, 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 $|\mathsf{Cov}(X_n, X_m)| \le ar^n$ 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 surperfluous, or $X_n \equiv X_1$ with any nontrivial disribution 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 $0 < a_n \nearrow \infty$. 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 $\mu_n := \mathsf{E}X_n$ and variances $\sigma_n^2 := \mathsf{V}(X_n)$. 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*

$$\sum_k \mathsf{V}\Big(\frac{X_k}{b_k}\Big) < \infty.$$

Then

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

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* $\epsilon > 0$, $\sum_{n=1}^\infty \mathsf{P}[|X_1| > n\epsilon] < \infty$.

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 L_1 SLLN) *Let $\{X_n\}$ be iid and set $S_n := \sum_{i\le n} X_i$. There exists $c \in \mathbb{R}$ such that*

$$\bar{X}_n = S_n/n \to c \quad \text{a.s.}$$

if and only if $\mathsf{E}|X_1| < \infty$, in which case $c = \mathsf{E}X_1$.

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

Corollary 1

$$\{X_n\} \subset L_1 \Rightarrow \bar{X}_n \to \mu \quad \text{a.s.}$$
$$\{X_n\} \subset L_2 \Rightarrow \frac{1}{n}\sum_{i\le n}(X_n - \bar{X}_n)^2 \to \sigma^2 \quad \text{a.s.}$$

Here's a quick summary of some LLN facts:

I. Weak version, non-*iid*, L_2: $\mu_i = \mathsf{E}X_i$, $\sigma_{ij} = \mathsf{E}[X_i - \mu_i][X_j - \mu_j]$

   A. $Y_n = (S_n - \Sigma\mu_i)/n$ satisfies $\mathsf{E}Y_n = 0$, $\mathsf{E}Y_n^2 = \frac{1}{n^2}\Sigma_{i\le n}\sigma_{ii} + \frac{2}{n^2}\Sigma_{i<j\le n}\sigma_{ij}$;

      1. If $\sigma_{ii} \le M$ and $\sigma_{ij} \le 0$ *or* $|\sigma_{ij}| < Mr^{|i-j|}$, $r < 1$,
         Chebychev $\Longrightarrow Y_n \to 0$, *pr.*

      2. (pairwise) IID L_2 is OK

II. Strong version, non-*iid*, L_2: $\mathsf{E}X_i = 0$, $\mathsf{E}X_i^2 \le M$, $\mathsf{E}X_iX_j \le 0$.

   A. $\mathsf{P}[|S_n| > n\epsilon] < \frac{Mn}{n^2\epsilon^2} = \frac{M}{n\epsilon^2}$

      1. $\mathsf{P}[|S_{n^2}| > n^2\epsilon] < \frac{M}{n^2\epsilon^2}$, $\Sigma_n \mathsf{P}[|S_{n^2}| > n^2\epsilon] < \frac{M\pi^2}{6\epsilon^2} < \infty$

      2. Borel-Cantelli: $\mathsf{P}[|S_{n^2}| > n^2\epsilon \; i.o.] = 0$, $\therefore \frac{1}{n^2}S_{n^2} \to 0$ *a.s.*

      3. $D_n \equiv \max_{n^2\le k<(n+1)^2} |S_k - S_{n^2}|$, so
         $\mathsf{E}D_n^2 \le 2n\mathsf{E}|S_{(n+1)^2-1} - S_{n^2}| \le 4n^2M$

      4. Chebychev: $\mathsf{P}[D_n > n^2\epsilon] < \frac{4n^2M}{n^4\epsilon^2}$, $\therefore D_n/n^2 \to 0$ *a.s.*

   B. $|S_k/k| \le \frac{|S_{n^2}|+D_n}{n^2} \to 0$ *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 $\sum_n \mathsf{P}[X_n \ne Y_n] < \infty$, then:

      1. $\sum_n |X_n - Y_n| < \infty$ *a.s.*

      2. $\sum_{i=1}^n X_i$ converges iff $\sum_{i=1}^n Y_i$ converges

      3. If $(\exists a_n \nearrow)(\exists X) \ni \frac{1}{a_n}\sum_{i\le n} X_i \to X$ then $\frac{1}{a_n}\sum_{i\le n} Y_i \to X$ too.

   B. If $\{X_n\}$ are iid L_1 then X_n, $Y_n := X_n 1_{[|X_n|\le n]}$ are equivalent

IV. Strong version, *iid*, L_1

   A. Kolmogorov: For $\{X_n\}$ IID, $(\exists c \ni \bar{X}_n \to c \; a.s.) \Leftrightarrow (X_n \in L_1)$.

V. Counterexamples: Cauchy,

   A. $X_i \sim \frac{dx}{\pi[1+x^2]} \Longrightarrow \mathsf{P}[|S_n|/n \le \epsilon] \equiv \frac{2}{\pi}\tan^{-1}(\epsilon) \not\to 1$, WLLN fails.

   B. $\mathsf{P}[X_i = \pm n] = \frac{c}{n^2}$, $n \ge 1$; $X_i \notin L_1$, and $S_n/n \not\to 0$ *pr.* or *a.s.*

   C. $\mathsf{P}[X_i = \pm n] = \frac{c}{n^2 \log n}$, $n > 1$; $X_i \notin L_1$, but $S_n/n \to 0$ *pr.* and not *a.s.*

   D. Medians: for ANY RVs $X_n \to X_\infty$ *pr.*, then $m_n \to m_\infty$ if m_∞ is unique.

Let X_i be iid standard Cauchy RVs, with

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

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

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

so S_n/n has characteristic function

$$\mathsf{E}\, e^{i\omega S_n/n} = \mathsf{E}\, e^{i\frac{\omega}{n}[X_1+\cdots+X_n]} = \Big(\mathsf{E}\, e^{i\frac{\omega}{n}X_1}\Big)^n = (e^{-|\frac{\omega}{n}|})^n = e^{-|\omega|}.$$

Thus S_n/n also has the standard Cauchy distribution with $\mathsf{P}[S_n/n \le t] = \frac{1}{2} + \frac{1}{\pi}\arctan(t)$; 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 $\{X_s \mid s \le t\}$, depends only on X_t) and want to estimate the population mean $\mathsf{E}[\phi(X_t)]$ for some function $\phi(\cdot)$ by the sample mean

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

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

A sequence of random variables Y_t is called *stationary* if each Y_t has the same probability distribution and, moreover, each finite set $(Y_{t_1+h}, Y_{t_2+h}, ..., Y_{t_k+h})$ has a joint distribution that doesn't depend on h. The sequence is called "L_2" if each Y_t has a finite variance σ^2 (and hence also a well-defined mean μ); by stationarity it also follows that the *covariance*

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

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

Theorem 6 *If a stationary L_2 sequence has a summable covariance,* i.e., *satisfies* $\sum_{t=-\infty}^{\infty} |\gamma(t)| \le c < \infty$, *then*

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

Proof. Let S_T be the sum of the first T Y_ts and set (as usual) $\bar{Y}_T \equiv S_T/T$. The variance of S_T is

$$\begin{aligned}
\mathsf{E}[(S_T - T\mu)^2] &= \sum_{s=0}^{T-1}\sum_{t=0}^{T-1} \mathsf{E}[(X_s-\mu)(X_t-\mu)] \\
&\le \sum_{s=0}^{T-1}\sum_{t=-\infty}^{\infty} |\gamma(s-t)| \\
&\le T\,c,
\end{aligned}$$

so $\bar{Y}_T$ has variance $\mathsf{V}[\bar{Y}_T] \le c/T$ and by Chebychev's inequality

$$\begin{aligned}
\mathsf{P}[|\bar{Y}_T - \mu| > \epsilon] &\le \frac{\mathsf{E}[(\bar{Y}_T-\mu)^2]}{\epsilon^2} \\
&= \frac{\mathsf{E}[(S_T - T\mu)^2]}{T^2\epsilon^2} \\
&\le \frac{T\,c}{T^2\epsilon^2} \\
&= \frac{c}{T\epsilon^2} \to 0 \quad \text{as } T\to\infty.
\end{aligned}$$

A strong LLN follows with a bit more work, just as for *iid* random variables. Note we didn't need full stationarity. It would be good enough for $\{Y_t\}$ to be "2^{nd}-order stationary," *i.e.*, to have a common mean $\mu = \mathsf{E}Y_t$ and a covariance function $\gamma_{st} = \mathsf{E}(Y_s-\mu)(Y_t-\mu) = \gamma(s-t)$ 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 σ^2, then Y_t has a well-defined finite mean μ and $\bar{Y}_T \to \mu$.
   Here $\gamma_{st} = \begin{cases} \sigma^2 & \text{if } s=t \\ 0 & \text{if } s\ne t \end{cases}$, so $c = \sigma^2 < \infty$.

2. **AR$_1$:** If Z_t are *iid* $\mathsf{No}(0,1)$ for $-\infty < t < \infty$, $\mu\in\mathbb{R}$, $\sigma>0$, $-1<\rho<1$, and

$$\begin{aligned}
X_t &\equiv \mu + \sigma\sum_{s=0}^{\infty} \rho^s Z_{t-s} \\
&= \rho X_{t-1} + \alpha + \sigma Z_t, \qquad (*)
\end{aligned}$$

   where $\alpha = (1-\rho)\mu$, then the X_t are identically distributed (all with the $\mathsf{No}(\mu, \frac{\sigma^2}{1-\rho^2})$ distribution) but not independent (since $\gamma_{st} = \frac{\sigma^2}{1-\rho^2}\rho^{|s-t|} \ne 0$ for $s\ne t$). This is called an "autoregressive process" (because of equation (*), expressing X_t as a regression of previous X_ss) of order one (because only one earlier X_s appears in (*)), and is about the simplest non-*iid* sequence occuring in applications. Since the covariance is summable,

$$\sum_{t=-\infty}^{\infty} |\gamma(t)| = \frac{\sigma^2}{1-\rho^2}\frac{1+|\rho|}{1-|\rho|} = \frac{\sigma^2}{(1-|\rho|)^2} < \infty,$$

   we again have $\bar{X}_T \to \mu$ *a.s.* as $T\to\infty$.

3. **Geometric Ergodicity:** If for some $0 < \rho < 1$ and $c > 0$ we have $\gamma_{st} \le c\rho^{|s-t|}$ 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 $\bar{Y}_t$ 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.

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

   (a) $X_{n+1} \equiv 2\,X_n \pmod 1$;

   (b) $X_{n+1} \equiv X_n + \alpha \pmod 1$ (Does it matter if α is rational?);

   (c) $X_{n+1} \equiv 4X_n(1 - X_n)$.

   For each, find a probability measure P (equivalently find a distribution for the random variable X_0) such that the X_n are all identically distributed; the sequence is called *ergodic* if each $E \in \mathcal{F}$ left invariant by the transformation T that takes X_n to X_{n+1}, $E = T^{-1}(E)$, is "almost trivial" in the sense that $\mathsf{P}[E] = 0$ or $\mathsf{P}[E] = 1$. Birkhoff's *Ergodic Theorem* asserts that $\bar{X}_n$ converges almost-surely to a T-invariant limit X_∞ as $n \to \infty$. Since only constants are T-invariant for ergodic sequences, it follows that $\bar{X}_n \to \mu = \mathsf{E}X_n$. The conditions here are weaker than those for the usual LLN; in all three cases above, for example, each X_n is completely determined by X_0 so there is complete dependence!

   For any L_1 distribution $\mu(dx)$ on $(\mathbb{R}, \mathcal{B})$, we can construct iid random variables X_n on the product probability space $\big(\Omega = \mathbb{R}^\infty, \mathcal{F} = \mathcal{B}^\infty, \mathsf{P} = \bigotimes \mu_n\big)$ and a measure-preserving transformation $T : \Omega \to \Omega$ by

   $$T(\omega_1, \omega_2, \omega_3, \dots) = (\omega_2, \omega_3, \omega_4, \dots),$$

   the left-shift. The σ-algebra $\mathcal{T} = \{A : \ A = T^{-1}(A)\}$ of T-invariant sets is just the tail σ-algebra for the independent random variables $\{X_n : \ X_n(\omega) = \omega_n\}$, 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

   $$\bar{X}_n = \frac{1}{n} \sum_{j=1}^{n} X_j$$

   converges almost-surely to a T-invariant and hence almost-surely constant random variable whose value must be μ, proving a strong LLN for iid random variables that assumes only L_1:

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

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

Last edited: September 17, 2015