Uniform Laws of Large Numbers

Carlos C. Rodríguez
http://omega.albany.edu:8008/

September 30, 2004

What is a Law of Large Numbers?

I am glad you asked! The Laws of Large Numbers, or LLNs for short, come in three basic flavors: Weak, Strong and Uniform. They all state that the observed frequencies of events tend to approach the actual probabilities as the number of observations increases. Saying it in another way, the LLNs show that under certain conditions, we can assymptotically *learn* the probabilities of events from their observed frequencies. To add some drama we could say that if God is not cheating and S/he doesn't change the innitial standard probabilistic model too much then, in principle, we (or other machines, or even the universe as a whole) could eventually find out the Truth, the whole Truth, and nothing but the Truth.

Bull! The Devil, is in the details.

I suspect that for reasons not too different in spirit to the ones above, famous minds of the past took the slippery slope of defining probabilities as the limits of relative frequencies. They became known as "frequentists". They wrote the books and indoctrinated generations of confused students.

As we shall see below, all the LLNs follow from the addition and product rules of probability theory. So, no matter what interpretation is ascribed to the concept of probability, if the numerical values of the events under consideration follow the addition and product rules then the LLNs are just an inevitable logical consequence. In other words, you don't have to be a frequentist to enjoy the LLNs. In fact, due to the very existence of the LLNs, it is not possible to define probabilities with the limit frequencies in a consistent way. This is simply because all LLNs state only probabilistic convergence of frequencies to probabilities (the convergence is either in probability or with probability 1). The concept that we want to interpret (namely probability) is needed to define the very concept (namely the LLNs) that is suppose to explain it. The frequentist concept of probability eats its own tail!

The Weak Law

The Weak Law of Large Numbers (WLLN) goes back to the beginnings of probability theory. It was discovered for the case of random coin flips by James Bernoulli at around 1700 but only appeared in print posthumously in his *Ars Conjectandy* in 1713. Later on, in 1800, Poisson generalized the result for general independent coin flips. After that Tchebychev in 1866 discovered his inequality and generalized the law for arbitrary sequences of independent random variables with second moments. Finally, his student Markov extended it to some classes of dependent random variables. Markov's inequality is almost a triviality but it has found innumerable applications.

Theorem 1 (Markov's inequality) *If X is nonnegative and $t > 0$,*

$$P\{X \geq t\} \leq \frac{EX}{t}$$

Proof: for $t > 0$,

$$X \geq X1_{[X \geq t]} \geq t1_{[X \geq t]}$$

and by the monotonicity of expectations we find that,

$$EX \geq tP\{X \geq t\}\bullet$$

Two important consequences of Markov's inequality are:

Tchebychev's inequality If $V(X)$ denotes the variance of X then,

$$P\{|X - EX| \geq t\} = P\{|X - EX|^2 \geq t^2\} \leq \frac{V(X)}{t^2}$$

Chernoff's method For $t > 0$ find the best s in,

$$P\{X \geq t\} = P\{e^{sX} \geq e^{st}\} \leq \frac{Ee^{sX}}{e^{st}}$$

Thus, when $X_1, X_2, \ldots, X_n$ are independent and identically distributed (iid) as X the sample mean,

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

has mean EX and variance $V(X)/n$ so by Tchebychev, for any $\epsilon > 0$

$$P\{|\bar{X}_n - EX| \geq \epsilon\} \leq \frac{V(X)}{n\epsilon^2}$$

and it immediately follows that,

$$\lim_{n\to\infty} P\{|\bar{X}_n - EX| \geq \epsilon\} = 0$$

which is what is meant by the sentence "the sample mean converges in probability to the expected value". That's the WLLN. For the special case of coin flips, i.e. for binary r.v.'s Bin(p), with $P\{X = 1\} = 1 - P\{X = 0\} = p$ the Tchebychev bound gives,

$$P\{|\bar{X}_n - p| \geq \epsilon\} \leq \frac{p(1-p)}{n\epsilon^2}$$

showing that the observed frequency of ones converges in probability to the true probability p of observing a 1.

The Strong Law

The bounds above obtained from Tchebychev's inequality are very poor. By using Chernoff's method an exponential bound can be obtained. In fact we have,

Hoeffding's inequality

$$P\left\{|\bar{X}_n - p| \geq \epsilon\right\} \leq 2e^{-2n\epsilon^2}$$

and by the classic Borel-Cantelli lemma it follows that,

$$P\{\omega : \lim_{n\to\infty} \bar{X}_n(\omega) = p\} = 1$$

which is the definition that the observed frequency of ones converges with probability one (or a.s. for almost surely) to the true probability p of observing a 1.

The proof of Hoeffding's inequality uses the following result for bounded r.v.'s with zero mean,

Lemma 1 *If EX=0 and $a \leq X \leq b$, then for any $s > 0$,*

$$Ee^{sX} \leq e^{s^2(b-a)^2/8}$$

Proof: Let $a \leq x \leq b$ and define $\lambda \in [0, 1]$ by

$$\lambda = \frac{b-x}{b-a}$$

Notice that for any $s > 0$ we have,

$$sx = \lambda sa + (1-\lambda)sb$$

Thus, exp(.) convex implies,

$$e^{sx} \leq \frac{b-x}{b-a}e^{sa} + \frac{x-a}{b-a}e^{sb}$$

Replacing x with the r.v. X, taking expectations and letting $p = -a/(b-a)$ (notice that $EX = 0$ implies $p \in [0,1]$) we can write,

$$\begin{aligned} E\{e^{sX}\} &\leq \frac{b}{b-a}e^{sa} - \frac{a}{b-a}e^{sb} \\ &= (1-p)e^{sa} + pe^{sb} \\ &= (1-p+pe^{s(b-a)})e^{-ps(b-a)} \\ &\stackrel{\text{def}}{=} e^{\phi(u)}, \end{aligned}$$

where $u = s(b-a)$ and,

$$\phi(u) = -pu + \log(1-p+pe^u).$$

The lemma will follow from the last inequality above by showing that,

$$\phi(u) \leq \frac{u^2}{8} = \frac{s^2(b-a)^2}{8}.$$

To see that this is true just expand $\phi(u)$ about zero,

$$\phi(u) = \phi(0) + u\phi'(0) + \frac{1}{2}u^2\phi''(\theta)$$

where $\theta \in [0,u]$ exists by Taylor's theorem, and notice that $\phi(0) = \phi'(0) = 0$ and

$$\phi''(u) = \frac{p(1-p)e^{-u}}{(p+(1-p)e^{-u})^2} \leq \frac{1}{4}$$

this is just a special case of $z(1-z) \leq 1/4$ for $z = p/(p+(1-p)e^{-u})$. Alternatively, just take derivative equal 0 to find that the max (1/4) is achieved when $e^{-u} = p/(1-p)$.•

Notice that for the special case of $X \in \{1,-1\}$ with equal probability 1/2 for each value the result follows at once from,

$$Ee^{sX} = \cosh\ s \leq e^{s^2/2}$$

by comparing the two series term by term. It is just this case that is needed in the main VC-theorem below.

We are now ready to show

Proof: [Hoeffding's inequality]

We actually show a more general version for $X_1, \ldots, X_n$ independent with $a_i \leq X_i \leq b_i$. Let $Z_i = X_i - EX_i$ we have,

$$
\begin{aligned}
P\{|\sum_{i=1}^{n} Z_i| \geq t\} &\leq P\{\sum_{i=1}^{n} Z_i \geq t\} + P\{\sum_{i=1}^{n} -Z_i \geq t\} \\
&\leq 2e^{-st} \prod_{i=1}^{n} e^{s^2(b_i-a_i)^2/8} \\
&= 2\exp\{\frac{s^2}{8}\sum_{i=1}^{n}(b_i-a_i)^2 - st\}
\end{aligned}
$$

where we are using Chernoff's method and the previous lemma. The upper bound is optimized for when $s = 4t/\sum(b_i - a_i)^2$ producing,

$$
P\{|\sum_{i=1}^{n} Z_i| \geq t\} \leq 2e^{-2t^2/\sum(b_i-a_i)^2}
$$

which implies the claimed bound for the special case of coin flips. Just replace $t = n\epsilon$ and notice that for binary variables $\sum(b_i - a_i)^2 = n.\bullet$

The Modern Strong Uniform Laws

The historical evolution of laws of large numbers have been coincidental with important paradigm shifts in the theory of probability. The weak law of Bernoulli and Poisson with the later refinements of Tchebychev and Markov are characteristic of the early era of probability. Then came the strong laws of Borel, Cantelli, Kolmogorov and others. These characterized the time of the axiomatic formalization of probability as part of measure theory during the first part of the twentieth century. The latest addition, to this saga is what we'll concentrate on here. These are the so called strong uniform laws that have a combinatorial flavor and were discovered by Vapnik and Chervonenkis in the 1970's in connection with statistical learning.

We start with a powerful generalization of Hoeffding's inequality for general functions of independent r.v.'s satisfying the bounded difference assumption. Let $S \subset R^n$ and denote by $e_i \in R^n$ the ith cannonical vector with all zeros except for a 1 in the ith position. We say that a function $h : S \to R$ has bounded differences in S if for all $1 \leq i \leq n$,

$$
|h(x) - h(x + te_i)| \leq c_i
$$

for all $x \in S$ and all $t \in R$ so that $(x + te_i) \in S$. This means that the function does not change by more than c_i along the ith direction. We have,

McDiarmid's inequality Let h have bounded differences. For all $t > 0$,

$$
P\{|h(X_1, \ldots, X_n) - Eh| \geq t\} \leq 2e^{-2t^2/\sum c_i^2}
$$

Notice that when $h = \sum X_i$ we recover Hoeffding's inequality.
Proof: [McDiarmid's inequality] The idea is to write,

$$h - Eh = \sum_{i=1}^{n} Z_i$$

by using,

$$Z_i(X_1, \ldots, X_i) = E\{h|X_1, \ldots, X_i\} - E\{h|X_1, \ldots, X_{i-1}\}$$

these Z_i have zero mean and are bounded a.s. within the interval $[L_i, U_i]$ with the lower and upper limits given by the inf and sup over $X_i = u$ of Z_i. Thus, L_i and U_i depend only on $X_1, \ldots, X_{i-1}$ and $U_i - L_i \leq c_i$ is inherited from the bounded difference assumption about h. Therefore, using Chernoff's method and the previous lemma we have that for all $s > 0$,

$$\begin{aligned} P\{h - Eh \geq t\} &\leq e^{-st} E\left\{ e^{s\sum_{i=1}^{n-1} Z_i} E\{e^{sZ_n}|X_1, \ldots, X_{n-1}\}\right\} \\ &\leq e^{-st} e^{s^2 \sum_{i=1}^{n} c_i^2/8} \end{aligned}$$

where the lemma was used n times. Now optimize s and copy the steps used for the proof of Hoeffding's to obtain the result. •

Corollary Let ν_n be the empirical probability measure based on the iid sample $X_1, X_2, \ldots, X_n$. The function,

$$h_n = h_n(X_1, \ldots, X_n) = \sup_{A \in \mathcal{A}} |\nu_n\{A\} - \nu\{A\}|$$

has bounded differences for any class of sets $\mathcal{A}$.

Proof: By changing only one of the X_i the function h_n changes by at most $c_i = 1/n$.•

It then follows inmediately from McDiarmid's inequality that,

$$P\{|h_n - Eh_n| \geq t\} \leq 2e^{-2nt^2}$$

Thus, if we can show that $Eh_n \to 0$ as $n \to \infty$ we can deduce from the above inequality that, for any $t > 0$ and for any n sufficiently large,

$$P\{\sup_{A \in \mathcal{A}} |\nu_n\{A\} - \nu\{A\}| \geq t\} \leq 2e^{-2nt^2}$$

and by the Borel-Cantelli lemma we would have obtained that,

$$\sup_{A \in \mathcal{A}} |\nu_n\{A\} - \nu\{A\}| \to 0 \text{ a.s.}$$

as $n \to \infty$, i.e. we'll have a uniform strong law of large numbers over the class $\mathcal{A}$.

Enter Combinatorics

If $\mathcal{A}$ is a colletion of subsets of R^d we define the *shatter coefficients* associated to the class $\mathcal{A}$ as,

$$S(n,\mathcal{A}) = \max_{x_1,\ldots,x_n \in R^d} |\{A \cap \{x_1,\ldots,x_n\} : A \in \mathcal{A}\}|.$$

The integer $S(n,\mathcal{A})$ is the maximum number of subsets of a set of n points that appear in elements of $\mathcal{A}$. Here is a post-modern version of the Vapnik-Chervonenkis inequality due to Devroye and Lugosi.

Theorem: [VC inequality]

$$E\{\sup_{A\in\mathcal{A}} |\nu_n\{A\} - \nu\{A\}|\} \leq 2\left(\frac{\log 2S(n,\mathcal{A})}{n}\right)^{1/2}.$$

Before proving this, notice that classes $\mathcal{A}$ for which the rhs of the above inequality goes to zero allow strong uniform laws of large numbers. In other words, the class $\mathcal{A}$ must not be too populated in such a way that the logarithm of its shatter coefficients must increase at a rate slower than n. The proof uses the following Lemma which also has independent interest.

Lemma $Ee^{sZ_i} \leq e^{s^2c^2/2}$ implies that,

$$E\{\max_{i\leq n} Z_i\} \leq c(2\log\ n)^{1/2}.$$

Proof:

$$\begin{aligned} e^{sE\{\max_{i\leq n} Z_i\}} &\leq E\{e^{s(\max_{i\leq n} Z_i)}\} \\ &= E\{\max_{i\leq n} e^{sZ_i}\} \\ &\leq \sum_{i\leq n} Ee^{sZ_i} \\ &\leq ne^{s^2c^2/2} \end{aligned}$$

where we have used Jensen's inequality and the hypothesis. Hence,

$$E\{\max_{i\leq n} Z_i\} \leq \frac{log\ n}{s} + \frac{sc^2}{2}$$

is valid for any $s > 0$. The best bound, claimed by the theorem, is obtained at $s = c^{-1}(2\log\ n)^{1/2}$.•

Proof: [VC inequality]

We divide the proof into three simple parts. First we show,

First symmetrization

$$E\{\sup_{A\in\mathcal{A}}|\nu_n\{A\}-\nu\{A\}|\}\leq E\{\sup_{A\in\mathcal{A}}|\nu_n\{A\}-\nu'_n\{A\}|\}$$

where ν'_n denotes the empirical measure associated to an independent copy $X'_1,\ldots,X'_n$ of the original sample $X_1,\ldots,X_n$. This is just a simple fact that follows from two applications of Jensen's inequality and the fact that the unconditional expectation is the expectation of the expectation conditional on the original sample,

$$\begin{aligned}
E\{\sup_{A\in\mathcal{A}}|\nu_n\{A\}-\nu\{A\}|\} &= E\{\sup_{A\in\mathcal{A}}|E\{\nu_n\{A\}-\nu'_n\{A\}|X_1\ldots,X_n\}|\}\\
&\leq E\{\sup_{A\in\mathcal{A}}E\{|\nu_n\{A\}-\nu'_n\{A\}|\,|X_1\ldots,X_n\}\}\\
&\leq E\{E\{\sup_{A\in\mathcal{A}}|\nu_n\{A\}-\nu'_n\{A\}|\,|X_1\ldots,X_n\}\}\\
&= E\{\sup_{A\in\mathcal{A}}|\nu_n\{A\}-\nu'_n\{A\}|\}.
\end{aligned}$$

The second step is,

Second symmetrization

Introduce independently of the two samples, n independent random signs $\epsilon_1,\ldots,\epsilon_n$ i.e., $P\{\epsilon_i=1\}=P\{\epsilon_i=-1\}=1/2$ and notice that if Z_i are any independent r.v.s symmetric about 0 then the joint distribution of $\epsilon_1Z_1,\ldots,\epsilon_nZ_n$ is the same as the joint distribution of $Z_1,\ldots,Z_n$. Hence,

$$E\{\sup_{A\in\mathcal{A}}|\nu_n\{A\}-\nu\{A\}|\}\leq\frac{1}{n}E\{\sup_{A\in\mathcal{A}}|\sum_{i=1}^{n}\epsilon_i(1[X_i\in A]-1[X'_i\in A])|\}$$

where we used $Z_i=1[X_i\in A]-1[X'_i\in A]$. Finally the third step,

Counting and bounding

Here is where combinatorics gets into the picture. To compute the sup over the class $\mathcal{A}$ we only need to check a finite number of sets $A\in\mathcal{A}$, namely those that pick different subsets of the $2n$ values $\{x_1,x'_1,\ldots,x_n,x'_n\}$. Thus, we only need to check at most $m=S(2n,\mathcal{A})$ sets in $\mathcal{A}$ to find the sup. Let's denote these sets by $A_1,A_2,\ldots,A_m$ and let,

$$Y_j=\sum_{i=1}^{n}\epsilon_i(1[X_i\in A_j]-1[X'_i\in A_j])$$

we can then write,

$$
\begin{aligned}
E\{\sup_{A\in\mathcal{A}}|\nu_n\{A\}-\nu\{A\}|\} &\leq \frac{1}{n}E\{\max_{j\leq m}|Y_j|\} \\
&= \frac{1}{n}E\{\max\{Y_1,-Y_1,\ldots,Y_m,-Y_m\}\}
\end{aligned}
$$

Now we apply the previous Lemma by noticing that,

$$
Ee^{sY_j} = Ee^{-sY_j} \leq \prod_{i=1}^{n} e^{s^2/2} = e^{ns^2/2}
$$

and obtain,

$$
E\{\sup_{A\in\mathcal{A}}|\nu_n\{A\}-\nu\{A\}|\} \leq \frac{\sqrt{n}}{n}(2\log\ 2m)^{1/2}
$$

the result follows by noticing that $m = S(2n,\mathcal{A}) \leq S(n,\mathcal{A})^2$.•