# Laws of large numbers and Birkhoff's ergodic theorem

Vaughn Climenhaga

March 9, 2013

In preparation for the next post on the central limit theorem, it's worth recalling the fundamental results on convergence of the average of a sequence of random variables: the law of large numbers (both weak and strong), and its strengthening to non-IID sequences, the Birkhoff ergodic theorem.

## 1 Convergence of random variables

First we need to recall the different ways in which a sequence of random variables may converge. Let $Y_n$ be a sequence of real-valued random variables and $Y$ a single random variable to which we want the sequence $Y_n$ to "converge". There are various ways of formalising this.

### 1.1 Almost sure convergence

The strongest notion of convergence is "almost sure" convergence: we write $Y_n \xrightarrow{a.s.} Y$ if

$$\mathbb{P}(Y_n \to Y) = 1. \tag{1}$$

If $\Omega$ is the probability space on which the random variables are defined and $\nu$ is the probability measure defining $\mathbb{P}$, then this condition can be rewritten as

$$\nu\{\omega \in \Omega \mid Y_n(\omega) \to Y(\omega)\} = 1. \tag{2}$$

### 1.2 Convergence in probability

A weaker notion of convergence is convergence "in probability": we write $Y_n \xrightarrow{p} Y$ if

$$\mathbb{P}(|Y_n - Y| \geq \epsilon) \to 0 \text{ for any } \epsilon > 0. \tag{3}$$

In terms of $\Omega$ and $\nu$, this condition is

$$\nu\{\omega \in \Omega \mid |Y_n(\omega) - Y(\omega)| \geq \epsilon\} \to 0 \tag{4}$$

Almost sure convergence implies convergence in probability (by Egorov's theorem, but not vice versa. For example, let $I_n \subset [0,1]$ be any sequence of intervals such that for every $x \in [0,1]$ the sets

$$\{n \mid x \in I_n\}, \qquad \{n \mid x \notin I_n\}$$

are both infinite. Let $\Omega = [0,1]$ and let $Y_n = \mathbf{1}_{I_n}$ be the characteristic function of the interval $I_n$. Then $Y_n \xrightarrow{p} 0$ but $Y_n \stackrel{a.s.}{\not\longrightarrow} 0$.

### 1.3 Convergence in distribution

A still weaker notion of convergence is convergence "in distribution": we write $Y_n \xrightarrow{d} Y$ if, writing $F_n, F \colon \mathbb{R} \to [0,1]$ for the cumulative distribution functions of $Y_n$ and $Y$, we have $F_n(t) \to F(t)$ at all $t$ where $F(t)$ is continuous.

Convergence in probability implies convergence in distribution, but the converse fails if $Y$ is not a.s.-constant. Here is one broad class of examples showing this: suppose $Y \colon \Omega \to \mathbb{R}$ has $\mathbb{P}(Y \in A) = \mathbb{P}(Y \in -A)$ for every interval $A \subset \mathbb{R}$ (for example, this is true if $Y$ is normal with zero mean). Then $-Y$ and $Y$ have the same CDF, and so any sequence which converges in distribution to one of the two will also converge in distribution to the other; on the other hand, $Y_n$ cannot converge in probability to both $Y$ and $-Y$ unless $Y = 0$ a.s.

## 2 Weak law of large numbers

Given a sequence of real-valued random variables $X_n$, we consider the sums

$$S_n = X_1 + X_2 + \cdots + X_n.$$

Then $\frac{1}{n} S_n$ is the average of the first $n$ observations.

Suppose that the sequence $X_n$ is independent and identically distributed (IID) and that $X_n$ is integrable – that is, $\mathbb{E}(|X_n|) < \infty$. Then in particular the mean $\mu = \mathbb{E}(X_n)$ is finite. The weak law of large numbers says that $\frac{1}{n}S_n$ converges in probability to the constant function $\mu$. Because the limiting distribution here is a constant, it is enough to show convergence in distribution. This fact leads to a well-known proof of the weak law of large numbers using characteristic functions.

If a random variable $Y$ is absolutely continuous – that is, if it has a probability density function $f$ – then its characteristic function $\varphi_Y$ is the Fourier transform of $f$. More generally, the characteristic function of $Y$ is

$$\varphi_Y(t) = \mathbb{E}(e^{itY}). \tag{5}$$

Characteristic functions are related to convergence in distribution by Lévy's continuity theorem, which says (among other things) that $Y_n \xrightarrow{d} Y$ if and only if $\varphi_{Y_n}(t) \to \varphi_Y(t)$ for all $t \in \mathbb{R}$. In particular, to prove the weak law of large numbers it suffices to show that the characteristic functions of $\frac{1}{n}S_n$ converge pointwise to the function $e^{it\mu}$.

Let $\varphi$ be the characteristic function of $X_n$. (Note that each $X_n$ has the same characteristic function because they are identically distributed.) Let $\varphi_n$ be the characteristic function of $\frac{1}{n}S_n$ – then

$$\varphi_n(t) = \mathbb{E}(e^{\frac{it}{n}(X_1+\cdots+X_n)}).$$

Because the variables $X_n$ are independent, we have

$$\varphi_n(t) = \prod_{j=1}^{n} \mathbb{E}(e^{\frac{it}{n}X_j}) = \varphi\left(\frac{t}{n}\right)^n. \tag{6}$$

By Taylor's theorem and by linearity of expectation, we have for $t \approx 0$ that

$$\varphi(t) = \mathbb{E}(e^{itX_j}) = \mathbb{E}(1 + itX_j + o(t^2)) = 1 + it\mu + o(t),$$

and together with (6) this gives

$$\varphi_n(t) = \left(1 + \frac{it\mu}{n} + o(t/n)\right)^n \to e^{it\mu},$$

which completes the proof.

# 3 Strong law of large numbers and ergodic theorem

The strong law of large numbers states that not only does $\frac{1}{n}S_n$ converge to $\mu$ in probability, it also converges almost surely. This takes a little more work to prove. Rather than describe a proof here (a nice discussion of both laws, including a different proof of the weak law than the one above, can be found on Terry Tao's blog), we observe that the strong law of large numbers can be viewed as a special case of the Birkhoff ergodic theorem, and then give a proof of this result. First we state the ergodic theorem (or at least, the version of it that is most relevant for us).

**Theorem 1** *Let $(X, \mathcal{F}, \mu)$ be a probability space and $f\colon X \to X$ a measurable transformation. Suppose that $\mu$ is $f$-invariant and ergodic. Then for any $\varphi \in L^1(\mu)$, we have*

$$\frac{1}{n} S_n\varphi(x) \to \int \varphi \, d\mu \tag{7}$$

*for $\mu$-a.e. $x \in X$, where $S_n\varphi(x) = \varphi(x) + \varphi(fx) + \cdots + \varphi(f^{n-1}x)$.*

Before giving a proof, we describe how the strong law of large numbers is a special case of Theorem 1. Let $X_n$ be a sequence of IID random variables $\Omega \to \mathbb{R}$, and define a map $\pi\colon \Omega \to X := \mathbb{R}^{\mathbb{N}}$ by

$$\pi(\omega) = (X_1(\omega), X_2(\omega), \dots).$$

Let $\nu$ be the probability measure on $\Omega$ that determines $\mathbb{P}$, and let $\mu = \pi_*\nu = \nu \circ \pi^{-1}$ be the corresponding probability measure on $X$.

Because the variables $X_n$ are independent, $\mu$ has the form $\mu = \nu_1 \times \nu_2 \times \cdots$, and because they are identically distributed, all the marginal distributions $\nu_j$ are the same, so in fact $\mu = \nu^{\mathbb{N}}$ for some probability distribution $\nu$ on $\mathbb{R}$.

The measure $\mu$ is invariant and ergodic with respect to the dynamics on $X$ given by the shift map $f(x_1, x_2, x_3, \dots) = (x_2, x_3, x_4, \dots)$ (this is an example of a Bernoulli measure). Writing $x = (x_1, x_2, x_3, \dots) \in X$ and putting $\varphi(x) = x_1$, we see that for $x = \pi(\omega)$ we have

$$X_1(\omega) + \cdots + X_n(\omega) = S_n\varphi(x).$$

In particular, the convergence in (7) implies the strong law of large numbers.

## 4 Proving the ergodic theorem

To prove the ergodic theorem, it suffices to consider a function $\varphi \in L^1(\mu)$ with $\int \varphi \, d\mu = 0$ and show that the set

$$X_\varepsilon = \left\{ x \in X \mid \varlimsup_{n\to\infty} \frac{1}{n} S_n\varphi(x) > \varepsilon \right\}$$

has $\mu(X_\varepsilon) = 0$ for every $\varepsilon > 0$. Indeed, the set of points where (7) fails is the (countable) union of the sets $X_{1/k}$ for the functions $\pm(\varphi - \int \varphi \, d\mu)$, and thus has $\mu$-measure zero if this result holds.

Note that $X_\varepsilon$ is $f$-invariant, and so by ergodicity we either have $\mu(X_\varepsilon) = 0$ or $\mu(X_\varepsilon) = 1$. We assume that $\mu(X_\varepsilon) = 1$ and derive a contradiction by showing that this implies $\int \varphi \, d\mu > 0$.

The assumption on $\mu(X_\varepsilon)$ implies that $\varlimsup_{n\to\infty} S_n(\varphi - \varepsilon)(x) = \infty$ for $\mu$-a.e. $x$. The key step now is to use this fact to show that

$$\int \varphi \, d\mu \geq \varepsilon; \tag{8}$$

this is the content of the maximal ergodic theorem.

Proving the maximal ergodic theorem requires a small trick. Let $\psi = \varphi - \varepsilon$ and let $\psi_n(x) = \max\{S_k\psi(x) \mid 0 \leq k \leq n\}$. Then

$$\psi_{n+1} = \psi + \max\{0, \psi_n \circ f\}, \tag{9}$$

and because $\psi_n(x) \to \infty$ for $\mu$-a.e. $x$, this implies that $\psi_{n+1} - \psi_n \circ f$ converges $\mu$-a.e. to $\psi$. Now we want to argue that

$$\int \psi \, d\mu = \lim_{n\to\infty} \int (\psi_{n+1} - \psi_n \circ f) \, d\mu, \tag{10}$$

because the integral on the right is equal to $\int (\psi_{n+1} - \psi_n) \, d\mu$ by $f$-invariance of $\mu$, and this integral in turn is non-negative because $\psi_n$ is non-decreasing. So if (10) holds, then we have $\int \psi \, d\mu \geq 0$, which implies (8).

Pointwise convergence does not always yield convergence of integrals, so to verify (10) we need the Lebesgue dominated convergence theorem. Using (9) we have

$$\begin{aligned} \psi_{n+1} - \psi_n \circ f &= \psi + \max\{0, -\psi_n \circ f\} \\ &\leq \psi + \max\{0, -\psi \circ f\}, \end{aligned}$$

which is integrable, and so the argument just given shows that (10) holds and in particular $\int \varphi \, d\mu \geq \varepsilon$, contradicting the assumption on $\varphi$. This proves that $\mu(X_\varepsilon) = 0$, which as described above is enough to prove that (7) holds $\mu$-a.e.