Lecture 9

The Strong Law of Large Numbers

Reading: Grimmett-Stirzaker 7.2; David Williams "Probability with Martingales" 7.2
Further reading: Grimmett-Stirzaker 7.1, 7.3-7.5

With the Convergence Theorem (Theorem 54) and the Ergodic Theorem (Theorem 55) we have two very different statements of convergence of something to a stationary distribution. We are looking at a recurrent Markov chain $(X_t)_{t\geq 0}$, i.e. one that visits *every* state at arbitrarily large times, so clearly X_t itself does not converge, as $t \to \infty$. In this lecture, we look more closely at the different types of convergence and develop methods to show the so-called almost sure convergence, of which the statement of the Ergodic Theorem is an example.

9.1 Modes of convergence

Definition 59 Let X_n, $n \geq 1$, and X be random variables. Then we define

1. $X_n \to X$ *in probability*, if for all $\varepsilon > 0$, $\mathbb{P}(|X_n - X| > \varepsilon) \to 0$ as $n \to \infty$.
2. $X_n \to X$ *in distribution*, if $\mathbb{P}(X_n \leq x) \to \mathbb{P}(X \leq x)$ as $n \to \infty$, for all $x \in \mathbb{R}$ at which $x \mapsto \mathbb{P}(X \leq x)$ is continuous.
3. $X_n \to X$ *in* L^1, if $\mathbb{E}(|X_n|) < \infty$ for all $n \geq 1$ and $\mathbb{E}(|X_n - X|) \to 0$ as $n \to \infty$.
4. $X_n \to X$ *almost surely* (a.s.), if $\mathbb{P}(X_n \to X \text{ as } n \to \infty) = 1$.

Almost sure convergence is the notion that we will study in more detail here. It helps to consider random variables as functions $X_n : \Omega \to \mathbb{R}$ on a sample space Ω, or at least as functions of a common, typically infinite, family of independent random variables. What is different here from previous parts of the course (except for the Ergodic Theorem, which we yet have to inspect more thoroughly), is that we want to calculate probabilities that fundamentally depend on an infinite number of random variables. So far, we have been able to revert to events depending on only finitely many random variables by conditioning. This will not work here.

Let us start by recalling the definition of convergence of sequences, as $n \to \infty$,

$$x_n \to x \qquad \iff \qquad \forall_{m\geq 1}\exists_{n_m\geq 1}\forall_{n\geq n_m}|x_n - x| < 1/m.$$

If we want to consider all sequences $(x_n)_{n\geq 1}$ of possible values of the random variables $(X_n)_{n\geq 1}$, then

$$n_m = \inf\{k \geq 1 : \forall_{n\geq k}|x_n - x| < 1/m\} \in \mathbb{N} \cup \{\infty\}$$

will vary as a function of the sequence $(x_n)_{n\geq 1}$, and so it will become a random variable

$$N_m = \inf\{k \geq 1 : \forall_{n\geq k}|X_n - X| < 1/m\} \in \mathbb{N} \cup \{\infty\}$$

as a function of $(X_n)_{n\geq 1}$. This definition of N_m permits us to write

$$\mathbb{P}(X_n \to X) = \mathbb{P}(\forall_{m\geq 1}N_m < \infty).$$

This will occasionally help, when we are given almost sure convergence, but is not much use when we want to prove almost sure convergence. To prove almost sure convergence, we can transform as follows

$$\begin{aligned} & \mathbb{P}(X_n \to X) = \mathbb{P}(\forall_{m\geq 1}\exists_{N\geq 1}\forall_{n\geq N}|X_n - X| < 1/m) = 1 \\ \iff \quad & \mathbb{P}(\exists_{m\geq 1}\forall_{N\geq 1}\exists_{n\geq N}|X_n - X| \geq 1/m) = 0. \end{aligned}$$

We are used to events such as $A_{m,n} = \{|X_n - X| \geq 1/m\}$, and we understand events as subsets of Ω, or loosely identify this event as set of all $((x_k)_{k\geq 1}, x)$ for which $|x_n - x| \geq 1/m$. This is useful, because we can now translate $\exists_{m\geq 1}\forall_{N\geq 1}\exists n \geq N$ into set operations and write

$$\mathbb{P}(\cup_{m\geq 1} \cap_{N\geq 1} \cup_{n\geq N} A_{m,n}) = 0.$$

This event can only have zero probability if all events $\cap_{N\geq 1} \cup_{n\geq N} A_{m,n}$, $m \geq 1$, have zero probability (formally, this follows from the sigma-additivity of the measure $\mathbb{P}$). The Borel-Cantelli lemma will give a criterion for this.

Proposition 60 *The following implications hold*

$$\begin{array}{ccc} X_n \to X \text{ almost surely} & & \\ \Downarrow & & \\ X_n \to X \text{ in probability} & \Rightarrow & X_n \to X \text{ in distribution} \\ \Uparrow & & \\ X_n \to X \text{ in } L^1 & \Rightarrow & \mathbb{E}(X_n) \to \mathbb{E}(X) \end{array}$$

No other implications hold in general.

Proof: Most of this is Part A material. Some counterexamples are on Assignment 5. It remains to prove that almost sure convergence implies convergence in probability. Suppose, $X_n \to X$ almost surely, then the above considerations yield $\mathbb{P}(\forall_{m\geq 1}N_m < \infty) = 1$, i.e. $\mathbb{P}(N_k < \infty) \geq \mathbb{P}(\forall_{m\geq 1}N_m < \infty) = 1$ for all $k \geq 1$.

Now fix $\varepsilon > 0$. Choose $m \geq 1$ such that $1/m < \varepsilon$. Then clearly $|X_n - X| > \varepsilon > 1/m$ implies $N_m > n$ so that

$$\mathbb{P}(|X_n - X| > \varepsilon) \leq \mathbb{P}(N_m > n) \to \mathbb{P}(N_m = \infty) = 0,$$

as $n \to \infty$, for any $\varepsilon > 0$. Therefore, $X_n \to X$ in probability. □

9.2 The first Borel-Cantelli lemma

Let us now work on a sample space Ω. It is safe to think of $\Omega = \mathbb{R}^{\mathbb{N}} \times \mathbb{R}$ and $\omega \in \Omega$ as $\omega = ((x_n)_{n\geq 1}, x)$ as the set of possible outcomes for an infinite family of random variables (and a limiting variable).

The Borel-Cantelli lemmas are useful to prove almost sure results. Particularly limiting results often require certain events to happen infinitely often (i.o.) or only a finite number of times. Logically, this can be expressed as follows. Consider events $A_n \subset \Omega$, $n \geq 1$. Then

$$\omega \in A_n \text{ i.o.} \quad \iff \quad \forall_{n\geq 1} \exists_{m\geq n} \quad \omega \in A_m \quad \iff \quad \omega \in \bigcap_{n\geq 1} \bigcup_{m\geq n} A_m$$

Lemma 61 (Borel-Cantelli (first lemma)) *Let $A = \bigcap_{n\geq 1} \bigcup_{m\geq n} A_m$ be the event that infinitely many of the events A_n occur. Then*

$$\sum_{n\geq 1} \mathbb{P}(A_n) < \infty \quad \Rightarrow \quad \mathbb{P}(A) = 0$$

Proof: We have that $A \subset \bigcup_{m\geq n} A_m$ for all $n \geq 1$, and so

$$\mathbb{P}(A) \leq \mathbb{P}\left(\bigcup_{m\geq n} A_m\right) \leq \sum_{m\geq n} \mathbb{P}(A_m) \to 0 \qquad \text{as } n \to \infty$$

whenever $\sum_{n\geq 1} \mathbb{P}(A_n) < \infty$. □

9.3 The Strong Law of Large Numbers

Theorem 62 *Let $(X_n)_{n\geq 1}$ be a sequence of independent and identically distributed (iid) random variables with $\mathbb{E}(X_1^4) < \infty$ and $\mathbb{E}(X_1) = \mu$. Then*

$$\frac{S_n}{n} := \frac{1}{n} \sum_{i=1}^{n} X_i \to \mu \qquad \textit{almost surely.}$$

Fact 63 *Theorem 62 remains valid without the assumption $\mathbb{E}(X_1^4) < \infty$, just assuming $\mathbb{E}(|X_1|) < \infty$.*

The proof for the general result is hard, but under the extra moment condition $\mathbb{E}(X_1^4) < \infty$ there is a nice proof.

Lemma 64 *In the situation of Theorem 62, there is a constant $K < \infty$ such that for all $n \geq 0$*

$$\mathbb{E}((S_n - n\mu)^4) \leq Kn^2.$$

Proof: Let $Z_k = X_k - \mu$ and $T_n = Z_1 + \ldots + Z_n = S_n - n\mu$. Then

$$\mathbb{E}(T_n^4) = \mathbb{E}\left(\left(\sum_{i=1}^n Z_i\right)^4\right) = n\mathbb{E}(Z_1^4) + 3n(n-1)\mathbb{E}(Z_1^2 Z_2^2) \le Kn^2$$

by expanding the fourth power and noting that most terms vanish such as

$$\mathbb{E}(Z_1 Z_2^3) = \mathbb{E}(Z_1)\mathbb{E}(Z_2^3) = 0.$$

K was chosen appropriately, say $K = 4\max\{\mathbb{E}(Z_1^4), (\mathbb{E}(Z_1^2))^2\}$. □

Proof of Theorem 62: By the lemma,

$$\mathbb{E}\left(\left(\frac{S_n}{n} - \mu\right)^4\right) \le Kn^{-2}$$

Now, by Tonelli's theorem,

$$\mathbb{E}\left(\sum_{n\ge 1}\left(\frac{S_n}{n} - \mu\right)^4\right) = \sum_{n\ge 0}\mathbb{E}\left(\left(\frac{S_n}{n} - \mu\right)^4\right) < \infty \quad \Rightarrow \quad \sum_{n\ge 1}\left(\frac{S_n}{n} - \mu\right)^4 < \infty \quad \text{a.s.}$$

But if a series converges, the underlying sequence converges to zero, and so

$$\left(\frac{S_n}{n} - \mu\right)^4 \to 0 \quad \text{almost surely} \qquad \Rightarrow \qquad \frac{S_n}{n} \to \mu \quad \text{almost surely.}$$

□

This proof did not use the Borel-Cantelli lemma, but we can also conclude by the Borel-Cantelli lemma:

Proof of Theorem 62: We know by Markov's inequality that

$$\mathbb{P}\left(\frac{1}{n}|S_n - n\mu| \ge n^{-\gamma}\right) \le \frac{\mathbb{E}((S_n/n - \mu)^4)}{n^{-4\gamma}} = Kn^{-2+4\gamma}.$$

Define for $\gamma \in (0, 1/4)$

$$A_n = \left\{\frac{1}{n}|S_n - n\mu| \ge n^{-\gamma}\right\} \quad \Rightarrow \quad \sum_{n\ge 1}\mathbb{P}(A_n) < \infty \quad \Rightarrow \quad \mathbb{P}(A) = 0$$

by the first Borel-Cantelli lemma, where $A = \bigcap_{n\ge 1}\bigcup_{m\ge n} A_n$. But now, event A^c happens if and only if

$$\exists_N \forall_{n\ge N} \left|\frac{S_n}{n} - \mu\right| < n^{-\gamma} \qquad \Rightarrow \qquad \frac{S_n}{n} \to \mu.$$

□

9.4 The second Borel-Cantelli lemma

We won't need the second Borel-Cantelli lemma in this course, but include it for completeness.

Lemma 65 (Borel-Cantelli (second lemma)) *Let $A = \bigcap_{n\geq 1}\bigcup_{m\geq n} A_n$ be the event that infinitely many of the events A_n occur. Then*

$$\sum_{n\geq 1}\mathbb{P}(A_n) = \infty \text{ and } (A_n)_{n\geq 1} \text{ independent} \quad \Rightarrow \quad \mathbb{P}(A) = 1.$$

Proof: The conclusion is equivalent to $\mathbb{P}(A^c) = 0$. By de Morgan's laws

$$A^c = \bigcup_{n\geq 1}\bigcap_{m\geq n} A_m^c.$$

However,

$$\begin{aligned}\mathbb{P}\left(\bigcap_{m\geq n} A_m^c\right) &= \lim_{r\to\infty}\mathbb{P}\left(\bigcap_{m=n}^{r} A_m^c\right)\\ &= \prod_{m\geq n}(1-\mathbb{P}(A_m)) \leq \prod_{m\geq n}\exp\left(-\mathbb{P}(A_m)\right) = \exp\left(-\sum_{m\geq n}\mathbb{P}(A_m)\right) = 0\end{aligned}$$

whenever $\sum_{n\geq 1}\mathbb{P}(A_n) = \infty$. Thus

$$\mathbb{P}(A^c) = \lim_{n\to\infty}\mathbb{P}\left(\bigcap_{m\geq n} A_m^c\right) = 0.$$

□

As a technical detail: to justify some of the limiting probabilities, we use "continuity of $\mathbb{P}$" along increasing and decreasing sequences of events, that follows from the sigma-additivity of $\mathbb{P}$, cf. Grimmett-Stirzaker, Lemma 1.3.(5).

9.5 Examples

Example 66 (Arrival times in Poisson process) A Poisson process has independent and identically distributed inter-arrival times $(Z_n)_{n\geq 0}$ with $Z_n \sim Exp(\lambda)$. We denoted the partial sums (arrival times) by $T_n = Z_0 + \ldots + Z_{n-1}$. The Strong Law of Large Numbers yields

$$\frac{T_n}{n} \to \frac{1}{\lambda} \qquad \text{almost surely, as } n\to\infty.$$

Example 67 (Return times of Markov chains) For a positive-recurrent discrete-time Markov chain we denoted by

$$N_i = N_i^{(1)} = \inf\{n > 0 : M_n = i\}, \quad N_i^{(m+1)} = \inf\{n > N_i^{(m)} : M_n = i\}, m \in \mathbb{N},$$

the successive return times to 0. By the strong Markov property, the random variables $N_i^{(m+1)} - N_i^{(m)}$, $m \geq 1$ are independent and identically distributed. If we define $N_i^{(0)} = 0$ and start from i, then this holds for $m \geq 0$. The Strong Law of Large Number yields

$$\frac{N_i^{(m)}}{m} \to \mathbb{E}_i(N_i) \qquad \text{almost surely, as } m \to \infty.$$

Similarly, in continuous time, for

$$H_i = H_i^{(1)} = \inf\{t \geq T_1 : X_t = i\}, \quad H_i^{(m)} = T_{N_i^{(m)}}, m \in \mathbb{N},$$

we get

$$\frac{H_i^{(m)}}{m} \to \mathbb{E}_i(H_i) = m_i \qquad \text{almost surely, as } m \to \infty.$$

Example 68 (Empirical distributions) If $(Y_n)_{n\geq 1}$ is an infinite sample (independent and identically distributed random variables) from a discrete distribution ν on $\mathbb{S}$, then the random variables $B_n^{(i)} = 1_{\{Y_n = i\}}$, $n \geq 1$, are also independent and identically distributed for each fixed $i \in \mathbb{S}$, as functions of independent variables. The Strong Law of Large Numbers yields

$$\nu_i^{(n)} = \frac{\#\{k = 1, \ldots, n : Y_k = i\}}{n} = \frac{B_1^{(i)} + \ldots + B_n^{(i)}}{n} \to \mathbb{E}(B_1^{(i)}) = \mathbb{P}(Y_1 = i) = \nu_i$$

almost surely, as $n \to \infty$. The probability mass function $\nu^{(n)}$ is called empirical distribution. It lists relative frequencies in the sample and, for a specific realisation, can serve as an approximation of the true distribution. In applications of statistics, it is the sample distribution associated with a population distribution. The result that empirical distributions converge to the true distribution, is true uniformly in i and in higher generality, it is usually referred to as the Glivenko-Cantelli theorem.

Remark 69 (Discrete ergodic theorem) If $(M_n)_{n\geq 0}$ is a positive-recurrent discrete-time Markov chain, the Ergodic Theorem is a statement very similar to the example of empirical distributions

$$\frac{\#\{k = 0, \ldots, n-1 : M_k = i\}}{n} \to \mathbb{P}_\eta(M_0 = i) = \eta_i \qquad \text{almost surely, as } n \to \infty,$$

for a stationary distribution η, but of course, the M_n, $n \geq 0$, are not independent (in general). Therefore, we need to work a bit harder to deduce the Ergodic Theorem from the Strong Law of Large Numbers.