Chapter 24

The Almost-Sure Ergodic Theorem

> This chapter proves Birkhoff's ergodic theorem, on the almost-sure convergence of time averages to expectations, under the assumption that the dynamics are asymptotically mean stationary.

This is not the usual proof of the ergodic theorem, as you will find in e.g. Kallenberg. Rather, it uses the AMS machinery developed in the last lecture, following Gray (1988, sec. 7.2), in turn following Katznelson and Weiss (1982). The central idea is that of "blocking": break the infinite sequence up into non-overlapping blocks, show that each block is well-behaved, and conclude that the whole sequence is too. This is a very common technique in modern ergodic theory, especially among information theorists. In pure probability theory, the usual proof of the ergodic theorem uses a result called the "maximal ergodic lemma", which is clever but somewhat obscure, and doesn't seem to generalize well to non-stationary processes: see Kallenberg, ch. 10.

We saw, at the end of the last chapter, that if time-averages converge in the long run, they converge on conditional expectations. Our work here is showing that they (almost always) converge. We'll do this by showing that their lim infs and lim sups are (almost always) equal. This calls for some preliminary results about the upper and lower limits of time-averages.

Definition 294 *For any observable f, define the* lower and upper limits of its time averages *as, respectively,*

$$\underline{A}f(x) \equiv \liminf_{t\rightarrow\infty} A_t f(x) \quad (24.1)$$

$$\overline{A}f(x) \equiv \limsup_{t\rightarrow\infty} A_t f(x) \quad (24.2)$$

Define L_f as the set of x where the limits coincide:

$$L_f \equiv \left\{x \,\middle|\, \underline{A}f(x) = \overline{A}f(x)\right\} \quad (24.3)$$

Lemma 295 *$\underline{A}f$ and $\overline{A}f$ are invariant functions.*

PROOF: Use our favorite trick, and write $A_t f(Tx) = \frac{t+1}{t} A_{t+1} f(x) - f(x)/t$. Clearly, the lim sup and lim inf of this expression will equal the lim sup and lim inf of $A_{t+1}f(x)$, which is the same as that of $A_t f(x)$. □

Lemma 296 *The set of L_f is invariant.*

PROOF: Since $\underline{A}f$ and $\overline{A}f$ are both invariant, they are both measurable with respect to $\mathcal{I}$ (Lemma 262), so the set of x such that $\underline{A}f(x) = \overline{A}f(x)$ is in $\mathcal{I}$, therefore it is invariant (Definition 261). □

Lemma 297 *An observable f has the ergodic property with respect to an AMS measure μ if and only if it has it with respect to the stationary limit m.*

PROOF: By Lemma 296, L_f is an invariant set. But then, by Lemma 287, $m(L_f) = \mu(L_f)$. (Take $f = \mathbf{1}_{L_f}$ in the lemma.) f has the ergodic property with respect to μ iff $\mu(L_f) = 1$, so f has the ergodic property with respect to μ iff it has it with respect to m. □

Theorem 298 (Almost-Sure Ergodic Theorem (Birkhoff)) *If a dynamical system is AMS with stationary mean m, then all bounded observables have the ergodic property, and with probability 1 (under both μ and m),*

$$Af = \mathbf{E}_m [f|\mathcal{I}] \tag{24.4}$$

for all $f \in L_1(m)$.

PROOF: From Theorem 291 and its corollaries, it is enough to prove that all $L_1(m)$ observables have ergodic properties to get Eq. 24.4. From Lemma 297, it is enough to show that the observables have ergodic properties in the stationary system $\Xi, \mathcal{X}, m, T$. (Accordingly, all expectations in the rest of this proof will be with respect to m.) Since any observable can be decomposed into its positive and negative parts, $f = f^+ - f^-$, assume, without loss of generality, that f is positive. Since $\overline{A}f \geq \underline{A}f$ everywhere, it suffices to show that $\mathbf{E}\left[\overline{A}f - \underline{A}f\right] \leq 0$. This in turn will follow from $\mathbf{E}\left[\overline{A}f\right] \leq \mathbf{E}[f] \leq \mathbf{E}[\underline{A}f]$. (Since f is bounded, the integrals exist.)

We'll prove that $\mathbf{E}\left[\overline{A}f\right] \leq \mathbf{E}[f]$, by showing that the time average comes close to its lim sup, but *from above* (in the mean). Proving that $\mathbf{E}[\underline{A}f] \geq \mathbf{E}[f]$ will be entirely parallel.

Since f is bounded, we may assume that $\overline{f} \leq M$ everywhere.

For every $\epsilon > 0$, for every x there exists a *finite* t such that

$$A_t f(x) \geq \overline{f}(x) - \epsilon \tag{24.5}$$

This is because $\overline{f}$ is the limit of the *least* upper bounds. (You can see where this is going already — the time-average has to be close to its lim sup, but close *from above*.)

Define $t(x, \epsilon)$ to be the smallest t such that $\overline{f}(x) \leq \epsilon + A_t f(x)$. Then, since $\overline{f}$ is invariant, we can add from from time 0 to time $t(x, \epsilon) - 1$ and get:

$$\sum_{n=0}^{t(x,\epsilon)-1} K^n f(x) + \epsilon t(x, \epsilon) \geq \sum_{n=0}^{t(x,\epsilon)-1} K^n \overline{f}(x) \tag{24.6}$$

Define $B_N \equiv \{x | t(x, \epsilon) \geq N\}$, the set of "bad" x, where the sample average fails to reach a reasonable (ϵ) distance of the lim sup before time N. Because $t(x, \epsilon)$ is finite, $m(B_N)$ goes to zero as $N \to \infty$. Chose a N such that $m(B_N) \leq \epsilon/M$, and, for the corresponding bad set, drop the subscript. (We'll see why this level is important presently.)

We'll find it convenient to not deal directly with f, but with a related function which is better-behaved on the bad set B. Set $\tilde{f}(x) = M$ when $x \in B$, and $= f(x)$ elsewhere. Similarly, define $\tilde{t}(x, \epsilon)$ to be 1 if $x \in B$, and $t(x, \epsilon)$ elsewhere. Notice that $\tilde{t}(x, \epsilon) \leq N$ for all x. Something like Eq. 24.6 still holds for the nice-ified function $\tilde{f}$, specifically,

$$\sum_{n=0}^{\tilde{t}(x,\epsilon)-1} K^n \overline{f}(x) \leq \sum_{n=0}^{\tilde{t}(x,\epsilon)-1} K^n \tilde{f}(x) + \epsilon \tilde{t}(x, \epsilon) \tag{24.7}$$

If $x \in B$, this reduces to $\overline{f}(x) \leq M + \epsilon$, which is certainly true because $\overline{f}(x) \leq M$. If $x \notin B$, it will follow from Eq. 24.6, provided that $T^n x \notin B$, for all $n \leq \tilde{t}(x, \epsilon) - 1$. To see that this, in turn, must be true, suppose that $T^n x \in B$ for some such n. Because (we're assuming) $n < t(x, \epsilon)$, it must be the case that

$$A_n f(x) < \overline{f}(x) - \epsilon \tag{24.8}$$

Otherwise, $t(x, \epsilon)$ would not be the *first* time at which Eq. 24.5 held true. Similarly, because $T^n x \in B$, while $x \notin B$, $t(T^n x, \epsilon) > N \geq t(x, \epsilon)$, and so

$$A_{t(x,\epsilon)-n} f(T^n x) < \overline{f}(x) - \epsilon \tag{24.9}$$

Combining the last two displayed equations,

$$A_{t(x,\epsilon)} f(x) < \overline{f}(x) - \epsilon \tag{24.10}$$

contradicting the definition of $t(x, \epsilon)$. Consequently, there can be no $n < t(x, \epsilon)$ such that $T^n x \in B$.

We are now ready to consider the time average $A_L f$ over a stretch of time of some considerable length L. We'll break the time indices over which we're averaging into blocks, each block ending when $T^t x$ hits B again. We need to make sure that L is sufficiently large, and it will turn out that $L \geq N/(\epsilon/M)$ suffices, so that $NM/L \leq \epsilon$. The end-points of the blocks are defined recursively, starting with $b_0 = 0$, $b_{k+1} = b_k + \tilde{t}(T^{b_k} x, \epsilon)$. (Of course the b_k are implicitly dependent on x and ϵ and N, but suppress that for now, since these are constant through the argument.) The number of completed blocks, C, is the large k such that $L-1 \geq b_k$. Notice that $L-b_C \leq N$, because $\tilde{t}(x,\epsilon) \leq N$, so if $L-b_C > N$, we could squeeze in another block after b_C, contradicting its definition.

Now let's examine the sum of the lim sup over the trajectory of length L.

$$\sum_{n=0}^{L-1} K^n \overline{f}(x) = \sum_{k=1}^{C} \sum_{n=b_{k-1}}^{b_k} K^n \overline{f}(x) + \sum_{n=b_C}^{L-1} K^n \overline{f}(x) \tag{24.11}$$

For each term in the inner sum, we may assert that

$$\sum_{n=0}^{\tilde{t}(T^{b_k}x,\epsilon)-1} K^n \overline{f}(T^{b_k}x) \leq \sum_{n=0}^{\tilde{t}(T^{b_k}x,\epsilon)-1} K^n \tilde{f}(T^{b_k}x) + \epsilon \tilde{t}(T^{b_k}x,\epsilon) \tag{24.12}$$

on the strength of Equation 24.7, so, returning to the over-all sum,

$$\begin{aligned}
\sum_{n=0}^{L-1} K^n \overline{f}(x) &\leq \sum_{k=1}^{C} \sum_{n=b_{k-1}}^{b_k-1} K^n \tilde{f}(x) + \epsilon(b_k - b_{k-1}) + \sum_{n=b_C}^{L-1} K^n \overline{f}(x) && (24.13)\\
&= \epsilon b_C + \sum_{n=0}^{b_C-1} K^n \tilde{f}(x) + \sum_{n=b_C}^{L-1} K^n \overline{f}(x) && (24.14)\\
&\leq \epsilon b_C + \sum_{n=0}^{b_C-1} K^n \tilde{f}(x) + \sum_{n=b_C}^{L-1} M && (24.15)\\
&\leq \epsilon b_C + M(L-1-b_C) + \sum_{n=0}^{b_C-1} K^n \tilde{f}(x) && (24.16)\\
&\leq \epsilon b_C + M(N-1) + \sum_{n=0}^{b_C-1} K^n \tilde{f}(x) && (24.17)\\
&\leq \epsilon L + M(N-1) + \sum_{n=0}^{L-1} K^n \tilde{f}(x) && (24.18)
\end{aligned}$$

where the last step, going from b_C to L, uses the fact that both ϵ and $\tilde{f}$ are non-negative. Taking expectations of both sides,

$$\begin{aligned}
\mathbf{E}\left[\sum_{n=0}^{L-1} K^n \overline{f}(X)\right] &\leq \mathbf{E}\left[\epsilon L + M(N-1) + \sum_{n=0}^{L-1} K^n \tilde{f}(X)\right] && (24.19)\\
\sum_{n=0}^{L-1} \mathbf{E}\left[K^n \overline{f}(X)\right] &\leq \epsilon L + M(N-1) + \sum_{n=0}^{L-1} \mathbf{E}\left[K^n \tilde{f}(X)\right] && (24.20)\\
L\mathbf{E}\left[\overline{f}(x)\right] &\leq \epsilon L + M(N-1) + L\mathbf{E}\left[\tilde{f}(X)\right] && (24.21)
\end{aligned}$$

using the fact that $\overline{f}$ is invariant on the left-hand side, and that m is stationary on the other. Now divide both sides by L.

$$\begin{aligned} \mathbf{E}\left[\overline{f}(x)\right] &\leq \epsilon + \frac{M(N-1)}{L} + \mathbf{E}\left[\tilde{f}(X)\right] & (24.22) \\ &\leq 2\epsilon + \mathbf{E}\left[\tilde{f}(X)\right] & (24.23) \end{aligned}$$

since $MN/L \leq \epsilon$. Now let's bound $\mathbf{E}\left[\tilde{f}(X)\right]$ in terms of $\mathbf{E}[f]$:

$$\begin{aligned} \mathbf{E}\left[\tilde{f}\right] &= \int \tilde{f}(x)dm & (24.24) \\ &= \int_{B^c} \tilde{f}(x)dm + \int_B \tilde{f}(x)dm & (24.25) \\ &= \int_{B^c} f(x)dm + \int_B Mdm & (24.26) \\ &\leq \mathbf{E}[f] + \int_B Mdm & (24.27) \\ &= \mathbf{E}[f] + Mm(B) & (24.28) \\ &\leq \mathbf{E}[f] + M\frac{\epsilon}{M} & (24.29) \\ &= \mathbf{E}[f] + \epsilon & (24.30) \end{aligned}$$

using the definition of $\tilde{f}$ in Eq. 24.26, the non-negativity of f in Eq. 24.27, and the bound on $m(B)$ in Eq. 24.29. Substituting into Eq. 24.23,

$$\mathbf{E}\left[\overline{f}\right] \leq \mathbf{E}[f] + 3\epsilon \qquad (24.31)$$

Since ϵ can be made arbitrarily small, we conclude that

$$\mathbf{E}\left[\overline{f}\right] \leq \mathbf{E}[f] \qquad (24.32)$$

as was to be shown.

The proof of $\mathbf{E}\left[\underline{f}\right] \geq \mathbf{E}[f]$ proceeds in parallel, only the nice-ified function $\tilde{f}$ is set equal to 0 on the bad set.

Since $\mathbf{E}\left[\underline{f}\right] \geq \mathbf{E}[f] \geq \mathbf{E}\left[\overline{f}\right]$, we have that $\mathbf{E}\left[\underline{f} - \overline{f}\right] \geq 0$. Since however it is always true that $\overline{f} - \underline{f} \geq 0$, we may conclude that $\overline{f} - \underline{f} = 0$ m-almost everywhere. Thus $m(L_f) = 1$, i.e., the time average converges m-almost everywhere. Since this is an invariant event, it has the same measure under μ and its stationary limit m, and so the time average converges μ-almost-everywhere as well. By Corollary 292, $Af = \mathbf{E}_m[f|\mathcal{I}]$, as promised. □

Corollary 299 *Under the assumptions of Theorem 298, all $L_1(m)$ functions have ergodic properties, and Eq. 24.4 holds a.e. m and μ.*

Proof: We need merely show that the ergodic properties hold, and then the equation follows. To do so, define $\overline{f}_M(x) \equiv \overline{f}(x) \wedge M$, an upper-limited version of the lim sup. Reasoning entirely parallel to the proof of Theorem 298 leads to the conclusion that $\mathbf{E}\left[\overline{f}_M\right] \leq \mathbf{E}[f]$. Then let $M \to \infty$, and apply the monotone convergence theorem to conclude that $\mathbf{E}\left[\overline{f}\right] \leq \mathbf{E}[f]$; the rest of the proof goes through as before. □