Chapter 26

Decomposition of Stationary Processes into Ergodic Components

This chapter is concerned with the decomposition of asymptotically-mean-stationary processes into ergodic components.

Section 26.1 shows how to write the stationary distribution as a mixture of distributions, each of which is stationary and ergodic, and each of which is supported on a distinct part of the state space. This is connected to ideas in nonlinear dynamics, each ergodic component being a different basin of attraction.

Section 26.2 lays out some connections to statistical inference: ergodic components can be seen as minimal sufficient statistics, and lead to powerful tests.

26.1 Construction of the Ergodic Decomposition

In the last lecture, we saw that the stationary distributions of a given dynamical system form a convex set, with the ergodic distributions as the extremal points. A standard result in convex analysis is that any point in a convex set can be represented as a convex combination of the extremal points. Thus, any stationary distribution can be represented as a mixture of stationary and ergodic distributions. We would like to be able to determine the weights used in the mixture, and even more to give them some meaningful stochastic interpretation.

Let's begin by thinking about the effective distribution we get from taking time-averages starting from a given point. For every measurable set B, and every finite t, $A_t \mathbf{1}_B(x)$ is a well-defined measurable function. As B ranges over the σ-field $\mathcal{X}$, holding x and t fixed, we get a set function, and one which,

moreover, meets the requirements for being a probability measure. Suppose we go further and pass to the limit.

Definition 316 (Long-Run Distribution) *The* long-run distribution starting from the point x *is the set function* $\lambda(x)$*, defined through* $\lambda(x, B) = \lim_t A_t \mathbf{1}_B(x)$*, when the limit exists for all* $B \in \mathcal{X}$*. If* $\lambda(x)$ *exists,* x *is an* ergodic point*. The set of all ergodic points is* E*.*

Notice that whether or not $\lambda(x)$ exists depends *only* on x (and T and $\mathcal{X}$); the initial distribution has nothing to do with it. Let's look at some properties of the long-run distributions. (The name "ergodic point" is justified by one of them, Proposition 318.)

Proposition 317 *If* $x \in E$*, then* $\lambda(x)$ *is a probability distribution.*

PROOF: For every t, the set function given by $A_t \mathbf{1}_B(x)$ is clearly a probability measure. Since $\lambda(x)$ is defined by passage to the limit, the Vitali-Hahn Theorem (285) says $\lambda(x)$ must be as well. □

Proposition 318 *If* $x \in E$*, then* $\lambda(x)$ *is ergodic.*

PROOF: For every invariant set I, $\mathbf{1}_I(T^n x) = \mathbf{1}_I(x)$ for all n. Hence $A\mathbf{1}_I(x)$ exists and is either 0 or 1. This means $\lambda(x)$ assigns every invariant set either probability 0 or probability 1, so by Definition 300 it is ergodic. □

Proposition 319 *If* $x \in E$*, then* $\lambda(x)$ *is an invariant function of* x*, i.e.,* $\lambda(x) = \lambda(Tx)$*.*

PROOF: By Lemma 275, $A\mathbf{1}_B(x) = A\mathbf{1}_B(Tx)$, when the appropriate limit exists. Since, by assumption, it does in this case, for every measurable set $\lambda(x, B) = \lambda(Tx, B)$, and the set functions are thus equal. □

Proposition 320 *If* $x \in E$*, then* $\lambda(x)$ *is a stationary distribution.*

PROOF: For all B and x, $\mathbf{1}_{T^{-1}B}(x) = \mathbf{1}_B(Tx)$. So $\lambda(x, T^{-1}B) = \lambda(Tx, B)$. Since, by Proposition 319, $\lambda(Tx, B) = \lambda(x, B)$, it finally follows that $\lambda(x, B) = \lambda(x, T^{-1}B)$, which proves that $\lambda(x)$ is an invariant distribution. □

Proposition 321 *If* $x \in E$ *and* $f \in L_1(\lambda(x))$*, then* $\lim_t A_t f(x)$ *exists, and is equal to* $\mathbf{E}_{\lambda(x)}[f]$*.*

PROOF: This is true, by the definition of $\lambda(x)$, for the indicator functions of all measurable sets. Thus, by linearity of A_t and of expectation, it is true for all simple functions. Standard arguments then let us pass to all the functions integrable with respect to the long-run distribution. □

At this point, you should be tempted to argue as follows. If μ is an AMS distribution with stationary mean m, then $Af(x) = \mathbf{E}_m[f|\mathcal{I}]$ for almost all x.

So, it's reasonable to hope that m is a combination of the $\lambda(x)$, and yet further that

$$Af(x) = \mathbf{E}_{\lambda(x)}[f]$$

for μ-almost-all x. This is basically true, but will take some extra assumptions to get it to work.

Definition 322 (Ergodic Component) *Two ergodic points $x, y \in E$ belong to the same* ergodic component *when $\lambda(x) = \lambda(y)$. We will write the ergodic components as C_i, and the function mapping x to its ergodic component as $\phi(x)$. $\phi(x)$ is not defined if x is not an ergodic point. By a slight abuse of notation, we will write $\lambda(C_i, B)$ for the common long-run distribution of all points in C_i.*

Obviously, the ergodic components partition the set of ergodic points. (The partition is not necessarily countable, and in some important cases, such as that of Hamiltonian dynamical systems in statistical mechanics, it must be uncountable (Khinchin, 1949).) Intuitively, they form the coarsest partition which is still fully informative about the long-run distribution. It's also pretty clear that the partition is left alone with the dynamics.

Proposition 323 *For all ergodic points x, $\phi(x) = \phi(Tx)$.*

PROOF: By Lemma 319, $\lambda(x) = \lambda(Tx)$, and the result follows. □

Notice that I have been careful not to say that the ergodic components are invariant sets, because we've been using that to mean sets which are both left along by the dynamics *and* are measurable, i.e. members of the σ-field $\mathcal{X}$, and we have not established that any ergodic component is measurable, which in turn is because we have not established that $\lambda(x)$ is a measurable function.

Let's look a little more closely at the difficulty. If B is a measurable set, then $A_t \mathbf{1}_B(x)$ is a measurable function. If the limit exists, then $A\mathbf{1}_B(x)$ is also a measurable function, and consequently the set $\{y : A\mathbf{1}_B(y) = A\mathbf{1}_B(x)\}$ is a measurable set. Then

$$\phi(x) = \bigcap_{B \in \mathcal{X}} \{y : A\mathbf{1}_B(x) = A\mathbf{1}_B(y)\} \tag{26.1}$$

gives the ergodic component to which x belongs. The difficulty is that the intersection is over *all* measurable sets B, and there are generally an uncountable number of them (even if Ξ is countable!), so we have no guarantee that the intersection of uncountably many measurable sets is measurable. Consequently, we can't say that any of the ergodic components is measurable.

The way out, as so often in mathematics, is to cheat; or, more politely, to make an assumption which is strong enough to force open an exit, but not so strong that we can't support it or verify it[1] What we will assume is that

[1] For instance, we *could* just assume that uncountable intersections of measurable sets are measurable, but you will find it instructive to try to work out the consequences of this assumption, and to examine whether it holds for the Borel σ-field $\mathcal{B}$ — say on the unit interval, to keep things easy.

there is a *countable* collection of sets $\mathcal{G}$ such that $\lambda(x) = \lambda(y)$ if and only if $\lambda(x, G) = \lambda(y, G)$ for every $G \in \mathcal{G}$. Then the intersection in Eq. 26.1 need only run over the countable class $\mathcal{G}$, rather than all of $\mathcal{X}$, which will be enough to reassure us that $\phi(x)$ is a measurable set.

Definition 324 (Countable Extension Space) *A measurable space $\Omega, \mathcal{F}$ is a* countable extension space *when there is a countable field $\mathcal{G}$ of sets in Ω such that $\mathcal{F} = \sigma(\mathcal{G})$, i.e., $\mathcal{G}$ is the* generating field *of the σ-field, and any normalized, non-negative, finitely-additive set function on $\mathcal{G}$ has a unique extension to a probability measure on $\mathcal{F}$.*

The reason the countable extension property is important is that it lets us get away with just checking properties of measures on a countable class (the generating field $\mathcal{G}$). Here are a few important facts about countable extension spaces; proofs, along with a much more detailed treatment of the general theory, are given by Gray (1988, chs. 2 and 3), who however calls them "standard" spaces.

Proposition 325 *Every countable space is a countable extension space.*

Proposition 326 *Every Borel space is a countable extension space.*

Remember that finite-dimensional Euclidean spaces are Borel spaces.

Proposition 327 *A countable product of countable extension spaces is a countable extension space.*

The last proposition is important for us: if Σ is a countable extension space, it means that $\Xi \equiv \Sigma^{\mathbb{N}}$ is too. So if we have a discrete- or Euclidean- valued random sequence, we can switch to the sequence space, and still appeal to generating-class arguments based on countable fields. Without further ado, then, let's assume that Ξ, the state space of our dynamical system, is a countable extension space, with countable generating field $\mathcal{G}$.

Lemma 328 *$x \in E$ iff $\lim_t A_t \mathbf{1}_G(x)$ converges for every $G \in \mathcal{G}$.*

Proof: "If": A direct consequence of Definition 324, since the set function $A\mathbf{1}_G(x)$ extends to a unique measure. "Only if": a direct consequence of Definition 316, since every member of the generating field is a measurable set. □

Lemma 329 *The set of ergodic points is measurable: $E \in \mathcal{X}$.*

Proof: For each $G \in \mathcal{G}$, the set of x where $A_t \mathbf{1}_G(x)$ converges is measurable, because G is a measurable set. The set where those relative frequencies converge for all $G \in \mathcal{G}$ is the intersection of countably many measurable sets, hence itself measurable. This set is, exactly, the set of ergodic points (Lemma 328). □

Lemma 330 *All the ergodic components are measurable sets, and $\phi(x)$ is a measurable function. Thus, all $C_i \in \mathcal{I}$.*

PROOF: For each G, the set $\{y: \lambda(y,G) = \lambda(x,G)\}$ is measurable. So their intersection over all $G \in \mathcal{G}$ is also measurable. But, by the countable extension property, this intersection is precisely the set $\{y: \lambda(y) = \lambda(x)\}$. So the ergodic components are measurable sets, and, since $\phi^{-1}(C_i) = C_i$, ϕ is measurable. Since we have already seen that $T^{-1}C_i = C_i$, and now that $C_i \in \mathcal{X}$, we may say that $C_i \in \mathcal{I}$. □

Remark: Because C_i is a (measurable) invariant set, $\lambda(x, C_i) = 1$ for every $x \in C_i$. However, it does not follow that there might not be a smaller set, also with long-run measure 1, i.e., there might be a $B \subset C_i$ such that $\lambda(x, B) = 1$. For an extreme example, consider the uniform contraction on $\mathbb{R}$, with $Tx = ax$ for some $0 \leq a \leq 1$. Every trajectory converges on the origin. The only ergodic invariant measure the the Dirac delta function. Every point belongs to a single ergodic component.

More generally, if a little roughly[2], the ergodic components correspond to the dynamical systems idea of *basins of attraction*, while the support of the long-run distributions corresponds to the actual *attractors*. Basins of attraction typically contain points which are not actually parts of the attractor.

Theorem 331 (Ergodic Decomposition of AMS Processes) *Suppose $\Xi, \mathcal{X}$ is a countable extension space. If μ is an asymptotically mean stationary measure on Ξ, with stationary mean m, then $\mu(E) = m(E) = 1$, and, for any $f \in L_1(m)$, and μ- and m- almost all x,*

$$Af(x) = \mathbf{E}_{\lambda(x)}[f] = \mathbf{E}_m[f|\mathcal{I}] \tag{26.2}$$

so that

$$m(B) = \int \lambda(x, B) d\mu(x) \tag{26.3}$$

PROOF: For every set $G \in \mathcal{G}$, $A_t \mathbf{1}_G(x)$ converges for μ- and m- almost all x (Theorem 298). Since there are only countably many G, the set on which they all converge also has probability 1; this set is E. Since (Proposition 321) $Af(x) = \mathbf{E}_{\lambda(x)}[f]$, and (Theorem 298 again) $Af(x) = \mathbf{E}_m[f|\mathcal{I}]$ a.s., we have that $\mathbf{E}_{\lambda(x)}[f] = \mathbf{E}_m[f|\mathcal{I}]$ a.s.

Now let $f = \mathbf{1}_B$. As we know (Lemma 289), $\mathbf{E}_\mu[A\mathbf{1}_B(X)] = \mathbf{E}_m[\mathbf{1}_B(X)] = m(B)$. But, for each x, $A\mathbf{1}_B(x) = \lambda(x, B)$, so $m(B) = \mathbf{E}_\mu[\lambda(X, B)]$. □

In words, we have decomposed the stationary mean m into the long-run distributions of the ergodic components, with weights given by the fraction of the initial measure μ falling into each component. Because of Propositions 313 and 315, we may be sure that by mixing stationary ergodic measures, we obtain an ergodic measure, and that our decomposition is unique.

[2] I don't want to get into subtleties arising from the dynamicists tendency to define things topologically, rather than measure-theoretically.

26.2 Statistical Aspects

26.2.1 Ergodic Components as Minimal Sufficient Statistics

The connection between sufficient statistics and ergodic decompositions is a very pretty one. First, recall the idea of parametric statistical sufficiency.[3]

Definition 332 (Sufficiency, Necessity) *Let $\mathcal{P}$ be a class of probability measures on a common measurable space $\Omega, \mathcal{F}$, indexed by a parameter θ. A σ-field $\mathcal{S} \subseteq \mathcal{F}$ is* parametrically sufficient for θ, *or just* sufficient, *when $\mathbb{P}_\theta(A|\mathcal{S}) = \mathbb{P}_{\theta'}(A|\mathcal{S})$ for all θ, θ'. That is, all the distributions in $\mathcal{P}$ have the same distribution, conditional on $\mathcal{S}$. A random variable such that $\mathcal{S} = \sigma(S)$ is called a* sufficient statistic. *A σ-field is* necessary *(for the parameter θ) if it is a sub-σ-field of every sufficient σ-field; a* necessary statistic *is defined similarly. A σ-field which is both necessary and sufficient is* minimal sufficient.

Remark: The idea of sufficiency originates with Fisher; that of necessity, so far as I can work out, with Dynkin. This definition (after Dynkin (1978)) is based on what ordinary theoretical statistics texts call the "Neyman factorization criterion" for sufficiency. We will see all these concepts again when we do information theory.

Lemma 333 *$\mathcal{S}$ is sufficient for θ if and only if there exists an $\mathcal{F}$-measurable function $\lambda(\omega, A)$ such that*

$$\mathbb{P}_\theta(A|\mathcal{S}) = \lambda(\omega, A) \tag{26.4}$$

almost surely, for all θ.

Proof: Nearly obvious. "Only if": since the conditional probability exists, there must be some such function (it's a version of the conditional probability), and since all the conditional probabilities are versions of one another, the function cannot depend on θ. "If": In this case, we have a single function which is a version of all the conditional probabilities, so it must be true that $\mathbb{P}_\theta(A|\mathcal{S}) = \mathbb{P}_{\theta'}(A|\mathcal{S})$. □

Theorem 334 *If a process on a countable extension space is asymptotically mean stationary, then ϕ is a minimal sufficient statistic for its long-run distribution.*

Proof: The set of distributions $\mathcal{P}$ is now the set of all long-run distributions generated by the dynamics, and θ is an index which tracks them all unambiguously. We need to show both sufficiency and necessity. *Sufficiency:* The σ-field

[3] There is also a related idea of predictive statistical sufficiency, which we unfortunately will not be able to get to. Also, note that most textbooks on theoretical statistics state things in terms of random variables and measurable functions thereof, rather than σ-fields, but this is the more general case (Blackwell and Girshick, 1954).

generated by ϕ is the one generated by the ergodic components, $\sigma(\{C_i\})$. (Because the C_i are mutually exclusive, this is a particularly simple σ-field.) Clearly, $\mathbb{P}_\theta(A|\sigma(\{C_i\})) = \lambda(\phi(x), A)$ for all x and θ, so (Lemma 333), ϕ is a sufficient statistic. *Necessity*: Follows from the fact that a given ergodic component contains *all* the points with a given long-run distribution. Coarser σ-fields will not, therefore, preserve conditional probabilities. □

This theorem may not seem particularly exciting, because there isn't, necessarily, anything whose distribution matches the long-run distribution. However, it has deeper meaning under two circumstances when $\lambda(x)$ really is the asymptotic distribution of random variables.

1. If Ξ is really a sequence space, so that $X = S_1, S_2, S_3, \ldots$, then $\lambda(x)$ really *is* the asymptotic marginal distribution of the S_t, conditional on the starting point.

2. Even if Ξ is not a sequence space, if stronger conditions than ergodicity known as "mixing", "asymptotic stability", etc., hold, there are reasonable senses in which $\mathcal{L}(X_t)$ does converge, and converges on the long-run distribution.[4]

In both these cases, knowing the ergodic component thus turns out to be necessary and sufficient for knowing the asymptotic distribution of the observables. (Cf. Corollary 337 below.)

26.2.2 Testing Ergodic Hypotheses

Finally, let's close with an application to hypothesis testing, inspired by Badino (2004).

Theorem 335 *Let $\Xi, \mathcal{X}$ be a measurable space, and let μ_0 and μ_1 be two infinite-dimensional distributions of one-sided, discrete-parameter strictly-stationary Σ-valued stochastic processes, i.e., μ_0 and μ_1 are distributions on $\Xi^{\mathbb{N}}, \mathcal{X}^{\mathbb{N}}$, and they are invariant under the shift operator. If they are also ergodic under the shift, then there exists a sequence of sets $R_t \in \mathcal{X}^t$ such that $\mu_0(R_t) \to 0$ while $\mu_1(R_t) \to 1$.*

PROOF: By Proposition 314, there exists a set $R \in \mathcal{X}^{\mathbb{N}}$ such that $\mu_0(R) = 0$, $\mu_1(R) = 1$. So we just need to approximate B by sets which are defined on the first t observations in such a way that $\mu_i(R_t) \to \mu_i(R)$. If $R_t \downarrow R$, then monotone convergence will give us the necessary convergence of probabilities. Here is a construction with cylinder sets[5] that gives us the necessary sequence

[4] Lemma 305 already gave us a *kind* of distributional convergence, but it is of a very weak sort, known as "convergence in Cesàro mean", which was specially invented to handle sequences which are not convergent in normal senses! We will see that there is a direct correspondence between levels of distributional convergence and levels of decay of correlations.

[5] Introduced in Chapters 2 and 3. It's possible to give an alternative construction using the Hilbert space of all square-integrable random variables, and then projecting onto the subspace of those which are $\mathcal{X}^t$ measurable.

of approximations. Let

$$R_t \equiv R \cup \prod_{n=t+1}^{\infty} \Xi_t \tag{26.5}$$

Clearly, R_t forms a non-increasing sequence, so it converges to a limit, which equally clearly must be R. Hence $\mu_i(R_t) \to \mu_i(R) = i$. □

Remark: "R" is for "rejection". Notice that the regions R_t will in general depend on the actual sequence $X_1, X_2, \ldots X_t \equiv X_1^t$, and not necessarily be permutation-invariant. When we come to the asymptotic equipartition theorem in information theory, we will see a more explicit way of constructing such tests.

Corollary 336 *Let H_0 be "X_i are IID with distribution p_0" and H_1 be "X_i are IID with distribution p_1". Then, as $t \to \infty$, there exists a sequence of tests of H_0 against H_1 whose size goes to 0 while their power goes to 1.*

PROOF: Let μ_0 be the product measure induced by p_0, and μ_1 the product measure induced p_1, and apply the previous theorem. □

Corollary 337 *If X is a strictly stationary (one-sided) random sequence whose shift representation has countably-many ergodic components, then there exists a sequence of functions ϕ_t, each $\mathcal{X}_t$-measurable, such that $\phi_t(X_1^t)$ converges on the ergodic component with probability 1.*

PROOF: From Theorem 52, we can write $X_1^t = \pi_{1:t}U$, for a sequence-valued random variable U, using the projection operators of Chapter 2. For each ergodic component, by Theorem 335, there exists a sequence of sets $R_{t,i}$ such that $\mathbb{P}(X_1^t \in R_{t,i}) \to 1$ if $U \in C_i$, and goes to zero otherwise. Let $\phi(X_1^t)$ be the set of all C_i for which $X_1^t \in R_{t,i}$. By Theorem 331, U is in some component with probability 1, and, since there are only countably many ergodic components, with probability 1 X_1^t will eventually leave all but one of the $R_{t,i}$. The remaining one is the ergodic component. □