Chapter 27

Mixing

A stochastic process is mixing if its values at widely-separated times are asymptotically independent.

Section 27.1 defines mixing, and shows that it implies ergodicity.

Section 27.2 gives some examples of mixing processes, both deterministic and non-deterministic.

Section 27.3 looks at the weak convergence of distributions produced by mixing, and the resulting decay of correlations.

Section 27.4 defines *strong* mixing, and the "mixing coefficient" which measures it. It then states, but does not prove, a central limit theorem for strongly mixing sequences. (The proof would demand first working through the central limit theorem for martingales.)

For stochastic processes, "mixing" means "asymptotically independent": that is, the statistical dependence between $X(t_1)$ and $X(t_2)$ goes to zero as $|t_1 - t_2|$ increases. To make this precise, we need to specify how we measure the dependence between $X(t_1)$ and $X(t_2)$. The most common and natural choice (first used by Rosenblatt, 1956) is the total variation distance between their joint distribution and the product of their marginal distributions, but there are other ways of measuring such "decay of correlations"[1]. Under all reasonable choices, IID processes are, naturally enough, special cases of mixing processes. This suggests that many of the properties of IID processes, such as laws of large numbers and central limit theorems, should continue to hold for mixing processes, at least if the approach to independence is sufficiently rapid. This in turn means that many statistical methods originally developed for the IID case will continue to work when the data-generating process is mixing; this is true both of parametric methods, such as linear regression, ARMA models being mixing (Doukhan, 1995, sec. 2.4.1), and of nonparametric methods like kernel prediction (Bosq, 1998). Considerations of time will prevent us from going into

[1] The term is common, but slightly misleading: lack of correlation, in the ordinary covariance-normalized-by-standard-deviations sense, implies independence only in special cases, like Gaussian processes. Nonetheless, see Theorem 350.

the purely statistical aspects of mixing processes, but the central limit theorem at the end of this chapter will give some idea of the flavor of results in this area: much like IID results, only with the true sample size replaced by an effective sample size, with a smaller discount the faster the rate of decay of correlations.

27.1 Definition and Measurement of Mixing

Definition 338 (Mixing) *A dynamical system* $\Xi, \mathcal{X}, \mu, T$ *is* mixing *when, for any* $A, B \in \mathcal{X}$,

$$\lim_{t\rightarrow\infty} |\mu(A \cap T^{-t}B) - \mu(A)\mu(T^{-t}B)| = 0 \quad (27.1)$$

Lemma 339 *If* μ *is* T*-invariant, mixing is equivalent to*

$$\lim_{t\rightarrow\infty} \mu(A \cap T^{-t}B) = \mu(A)\mu(B) \quad (27.2)$$

PROOF: By stationarity, $\mu(T^{-t}B) = \mu(B)$, so $\mu(A)\mu(T^{-t}B) = \mu(A)\mu(B)$. The result follows. □

Theorem 340 *Mixing implies ergodicity.*

PROOF: Let A be any invariant set. By mixing, $\lim_t \mu(T^{-t}A \cap A) = \mu(T^{-t}A)\mu(A)$. But $T^{-t}A = A$ for every t, so we have $\lim \mu(A) = \mu^2(A)$, or $\mu(A) = \mu^2(A)$. This can only be true if $\mu(A) = 0$ or
$mu(A) = 1$, i.e., only if μ is T-ergodic. □

Everything we have established about ergodic processes, then, applies to mixing processes.

Definition 341 *A dynamical system is* asymptotically stationary, *with stationary limit* m, *when* $\lim_t \mu(T^{-t}A) = m(A)$ *for all* $A \in \mathcal{X}$.

Lemma 342 *An asymptotically stationary system is mixing iff*

$$\lim_{t\rightarrow\infty} \mu(A \cap T^{-t}B) = \mu(A)m(B) \quad (27.3)$$

for all $A, B \in \mathcal{X}$.

PROOF: Directly from the fact that in this case $m(B) = \lim_t T^{-t}B$. □

Theorem 343 *Suppose* $\mathcal{G}$ *is a* π*-system, and* μ *is an asymptotically stationary measure. If*

$$\lim_{t} \left|\mu(A \cap T^{-t}B) - \mu(A)\mu(T^{-t}B)\right| = 0 \quad (27.4)$$

for all $A, B \in \mathcal{G}$*, then it holds for all pairs of sets in* $\sigma(\mathcal{G})$*. If* $\sigma(\mathcal{G}) = \mathcal{X}$*, then the process is mixing.*

PROOF(after Durrett, 1991, Lemma 6.4.3): Via the π-λ theorem, of course. Let Λ_A be the class of all B such that the equation holds, for a given $A \in \mathcal{G}$. We need to show that Λ_A really is a λ-system.

$\Xi \in \Lambda_A$ is obvious. $T^{-t}\Xi = \Xi$ so $\mu(A \cap \Xi) = \mu(A) = \mu(A)\mu(\Xi)$.

Closure under complements. Let B_1 and B_2 be two sets in Λ_A, and assume $B_1 \subset B_2$. Because set-theoretic operations commute with taking inverse images, $T^{-t}(B_2 \setminus B_1) = T^{-t}B_2 \setminus T^{-t}B_1$. Thus

$$\begin{aligned} 0 \leq |\mu\left(A \cap T^{-t}(B_2 \setminus B_1)\right) - \mu(A)\mu(T^{-t}(B_2 \setminus B_1))| & \quad (27.5)\\ = \quad |\mu(A \cap T^{-t}B_2) - \mu(A \cap T^{-t}B_1) - \mu(A)\mu(T^{-t}B_2) + \mu(A)\mu(T^{-t}B_1)| & \\ \leq \quad |\mu(A \cap T^{-t}B_2) - \mu(A)\mu(T^{-t}B_2)| & \quad (27.6)\\ + |\mu(A \cap T^{-t}B_1) - \mu(A)\mu(T^{-t}B_1)| & \end{aligned}$$

Taking limits of both sides, we get that $\lim|\mu(A \cap T^{-t}(B_2 \setminus B_1)) - \mu(A)\mu(T^{-t}(B_2 \setminus B_1))| = 0$, so that $B_2 \setminus B_1 \in \Lambda_A$.

Closure under monotone limits: Let B_n be any monotone increasing sequence in Λ_A, with limit B. Thus, $\mu(B_n) \uparrow \mu(B)$, and at the same time $m(B_n) \uparrow m(B)$, where m is the stationary limit of μ. Using Lemma 342, it is enough to show that

$$\lim_t \mu(A \cap T^{-t}B) = \mu(A)m(B) \quad (27.7)$$

Since $B_n \subset B$, we can always use the following trick:

$$\begin{aligned} \mu(A \cap T^{-t}B) &= \mu(A \cap T^{-t}B_n) + \mu(A \cap T^{-t}(B \setminus B_n)) \quad (27.8)\\ \lim_t \mu(A \cap T^{-t}B) &= \mu(A)m(B_n) + \lim_t \mu(A \cap T^{-t}(B \setminus B_n)) \quad (27.9) \end{aligned}$$

For any $\epsilon > 0$, $\mu(A)m(B_n)$ can be made to come within ϵ of $\mu(A)m(B)$ by taking n sufficiently large. Let us now turn our attention to the second term.

$$\begin{aligned} 0 \leq \lim_t \mu(A \cap T^{-t}(B \setminus B_n)) &= \lim_t \mu(T^{-t}(B \setminus B_n)) \quad (27.10)\\ &= \lim_t \mu(T^{-t}B \setminus T^{-t}B_n) \quad (27.11)\\ &= \lim_t \mu(T^{-t}B) - \lim_t \mu(T^{-t}B_n) \quad (27.12)\\ &= m(B) - m(B_n) \quad (27.13) \end{aligned}$$

which again can be made less than any positive ϵ by taking n large. So, for sufficiently large n, $\lim_t \mu(A \cap T^{-t}B)$ is always within 2ϵ of $\mu(A)m(B)$. Since ϵ can be made arbitrarily small, we conclude that $\lim_t \mu(A \cap T^{-t}B) = \mu(A)m(B)$. Hence, $B \in \Lambda_A$.

We conclude, from the $\pi - \lambda$ theorem, that Eq. 27.4 holds for all $A \in \mathcal{G}$ and all $B \in \sigma(\mathcal{G})$. The same argument can be turned around for A, to show that Eq. 27.4 holds for all pairs $A, B \in \sigma(\mathcal{G})$. If $\mathcal{G}$ generates the whole σ-field $\mathcal{X}$, then clearly Definition 338 is satisfied and the process is mixing. □

27.2 Examples of Mixing Processes

Example 344 (IID Sequences) *IID sequences are mixing from Theorem 343, applied to finite-dimensional cylinder sets.*

Example 345 (Ergodic Markov Chains) *Another application of Theorem 343 shows that ergodic Markov chains are mixing.*

Example 346 (Irrational Rotations of the Circle are Not Mixing) *Irrational rotations of the circle, $Tx = x + \phi \bmod 1$, ϕ irrational, are ergodic (Example 310), and stationary under the Lebesgue measure. They are not, however, mixing. Recall that $T^t x$ is dense in the unit interval, for arbitrary initial x. Because it is dense, there is a sequence t_n such that $t_n \phi \bmod 1$ goes to $1/2$. Now let $A = [0, 1/4]$. Because T maps intervals to intervals (of equal length), it follows that $T^{-t_n} A$ becomes an interval disjoint from A, i.e., $\mu(A \cap T^{-t_n} A) = 0$. But mixing would imply that $\mu(A \cap T^{-t_n} A) \rightarrow 1/16 > 0$, so the process is not mixing.*

Example 347 (Deterministic, Reversible Mixing: The Cat Map) *Here $\Xi = [0, 1)^2$, $\mathcal{X}$ are the appropriate Borel sets, μ is Lebesgue measure on the square, and $Tx = (x_1 + x_2, x_1 + 2x_2) \bmod 1$. This is known as the* cat map. *It is a deterministic, invertible transformation, but it can be shown that it is actually mixing. (For a proof, which uses Theorem 349, the Fibonacci numbers and a clever trick with Fourier transforms, see Lasota and Mackey (1994, example 4.4.3, pp. 77–78).) The origins of the name lie with a figure in Arnol'd and Avez (1968), illustrating the mixing action of the map by successively distorting an image of a cat.*

27.3 Convergence of Distributions Under Mixing

To show how distributions converge (weakly) under mixing, we need to recall some properties of Markov operators. Remember that, for a Markov process, the time-evolution operator for observables, K, was defined through $Kf(x) = \mathbf{E}[f(X_1)|X_0 = x]$. Remember also that it induces an adjoint operator for the evolution of distributions, taking signed measures to signed measures, through the intermediary of the transition kernel. We can view the measure-updating operator U as a linear operator on $L_1(\mu)$, which takes non-negative μ-integrable functions to non-negative μ-integrable functions, and probability densities to probability densities. Since dynamical systems are Markov processes, all of this remains valid; we have K defined through $Kf(x) = f(Tx)$, and U through the adjoint relationship, $\mathbf{E}_\mu[f(X)Kg(X)] = \mathbf{E}[Uf(X)g(X)]\mu$, where $g \in L_\infty$ and $f \in L_1(\mu)$. These relations continue to remain valid for powers of the operators.

Lemma 348 *In any Markov process, $U^n d$ converges weakly to 1, for all initial probability densities d, if and only if $U^n f$ converges weakly to $\mathbf{E}_\mu[f]$, for all initial L_1 functions f, i.e. $\mathbf{E}_\mu[U^n f(X)g(X)] \rightarrow \mathbf{E}_\mu[f(X)]\mathbf{E}_\mu[g(X)]$ for all bounded, measurable g.*

PROOF: "If": If d is a probability density with respect to μ, then $\mathbf{E}_{\mu}[d] = 1$. "Only if": Re-write an arbitrary $f \in L_1(\mu)$ as the difference of its positive and negative parts, $f = f^+ - f^-$. A positive f is a re-scaling of some density, $f = cd$ for constant $c = \mathbf{E}_{\mu}[f]$ and a density d. Through the linearity of U and its powers,

$$\begin{aligned} \lim U^t f &= \lim U^t f^+ - \lim U^t f^- && (27.14)\\ &= \mathbf{E}_{\mu}\left[f^+\right] \lim U^t d^+ - \mathbf{E}_{\mu}\left[f^-\right] \lim U^t d^- && (27.15)\\ &= \mathbf{E}_{\mu}\left[f^+\right] - \mathbf{E}_{\mu}\left[f^-\right] && (27.16)\\ &= \mathbf{E}_{\mu}\left[f^+ - f^-\right] = \mathbf{E}_{\mu}[f] && (27.17) \end{aligned}$$

using the linearity of expectations at the last step. □

Theorem 349 *A T-invariant probability measure μ is T-mixing if and only if any initial probability measure $\nu << \mu$ converges weakly to μ under the action of T, i.e., iff, for all bounded, measurable f,*

$$\mathbf{E}_{U^t \nu}[f(X)] \rightarrow \mathbf{E}_{\mu}[f(X)] \quad (27.18)$$

PROOF: Exercise. The way to go is to use the previous lemma, of course. With that tool, one can prove that the convergence holds for indicator functions, and then for simple functions, and finally, through the usual arguments, for all L_1 densities.

Theorem 350 (Decay of Correlations) *A stationary system is mixing if and only if*

$$\lim_{t \rightarrow \infty} \mathrm{cov}\left(f(X_0), g(X_t)\right) = 0 \quad (27.19)$$

for all bounded observables f, g.

PROOF: Exercise, from the fact that convergence in distribution implies convergence of expectations of all bounded measurable functions. □

It is natural to ask what happens if $U^t \nu \rightarrow \mu$ not weakly but strongly. This is known as *asymptotic stability* or (especially in the nonlinear dynamics literature) *exactness*. Remarkably enough, it is equivalent to the requirement that $\mu(T^t A) \rightarrow 1$ whenever $\mu(A) > 0$. (Notice that for once the expression involves *images* rather than pre-images.) There is a kind of hierarchy here, where different levels of convergence of distribution (Cesáro, weak, strong) match different sorts of ergodicity (metric transitivity, mixing, exactness). For more details, see Lasota and Mackey (1994).

27.4 A Central Limit Theorem for Mixing Sequences

Notice that I say "*a* central limit theorem", rather than "*the* central limit theorem". In the IID case, the necessary and sufficient condition for the CLT is well-known (you saw it in 36-752) and reasonably comprehensible. In the mixing case, a necessary and sufficient condition is known[2], but not commonly used, because quite opaque and hard to check. Rather, the common practice is to rely upon a large set of distinct sufficient conditons. Some of these, it must be said, are pretty ugly, but they are more susceptible of verification.

Recall the notation that X_t^- consists of the entire past of the process, including X_t, and X_t^+ its entire future.

Definition 351 (Mixing Coefficients) *For a stochastic process X_t, define the* strong-, Rosenblatt- *or* α- mixing coefficients *as*

$$\alpha(t_1, t_2) = \sup\left\{|\mathbb{P}(A \cap B) - \mathbb{P}(A)\,\mathbb{P}(B)| : \ A \in \sigma(X_{t_1}^-),\ B \in \sigma(X_{t_2}^+)\right\} \quad (27.20)$$

If the system is conditionally stationary, then $\alpha(t_1, t_2) = \alpha(t_2, t_1) = \alpha(|t_1 - t_2|) \equiv \alpha(\tau)$. If $\alpha(\tau) \to 0$, then the process is strong-mixing *or* α-mixing*. If $\alpha(\tau) = O(e^{-b\tau})$ for some $b > 0$, the process is* exponentially mixing*, b is the* mixing rate*, and $1/b$ is the* mixing time*. If $\alpha(\tau) = O(\tau^{-k})$ for some $k > 0$, then the process is* polynomially mixing*.*

Notice that $\alpha(t_1, t_2)$ is just the total variation distance between the joint distribution, $\mathcal{L}\left(X_{t_1}^-, X_{t_2}^+\right)$, and the product of the marginal distributions, $\mathcal{L}\left(X_{t_1}^-\right) \times \mathcal{L}\left(X_{t_2}^+\right)$. Thus, it is a natural measure of the degree to which the future of the system depends on its past. However, there are at least four other mixing coefficients (β, ϕ, ψ and ρ) regularly used in the literature. Since any of these others going to zero implies that α goes to zero, we will stick with α-mixing, as in Rosenblatt (1956).

Also notice that if X_t is a Markov process (e.g., a dynamical system) then the Markov property tells us that we only need to let the supremum run over measurable sets in $\sigma(X_{t_1})$ and $\sigma(X_{t_2})$.

Lemma 352 *If a dynamical system is α-mixing, then it is mixing.*

Proof: α is the supremum of the quantity appearing in the definition of mixing. □

Notation: For the remainder of this section,

$$S_n \equiv \sum_{k=1}^{n} X_n \quad (27.21)$$

$$\sigma_n^2 \equiv \mathbf{Var}\left[S_n\right] \quad (27.22)$$

$$Y_n(t) \equiv \frac{S_{[nt]}}{\sigma_n} \quad (27.23)$$

where n is any positive integer, and $t \in [0, 1]$.

[2] Doukhan (1995, p. 47) cites Jakubowski and Szewczak (1990) as the source, but I have not verified the reference.

Definition 353 X_t obeys the central limit theorem *when*

$$\frac{S_n}{\sigma\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1) \tag{27.24}$$

for some positive σ.

Definition 354 X_t obeys the functional central limit theorem *or* the invariance principle *when*

$$Y_n \xrightarrow{d} W \tag{27.25}$$

where W *is a standard Wiener process on* $[0,1]$*, and the convergence is in the Skorokhod topology of Sec. 15.1.*

Theorem 355 (Central Limit Theorem for α-Mixing Sequences) *Let* X_t *be a stationary sequence with* $\mathbf{E}[X_t] = 0$. *Suppose* X *is* α*-mixing, and that for some* $\delta > 0$

$$\mathbf{E}\left[|X_t|^{2+\delta}\right] \leq \infty \tag{27.26}$$

$$\sum_{n=0}^{\infty} \alpha^{\frac{\delta}{2+\delta}}(n) \leq \infty \tag{27.27}$$

Then

$$\lim_{n\to\infty} \frac{\sigma_n^2}{n} = \mathbf{E}\left[|X_1|^2\right] + 2\sum_{k=1}^{\infty} \mathbf{E}[X_1 X_k] \equiv \sigma^2 \tag{27.28}$$

If $\sigma^2 > 0$*, moreover,* X_t *obeys both the central limit theorem with variance* σ^2*, and the functional central limit theorem.*

PROOF: Complicated, and based on a rather technical central limit theorem for martingale difference arrays. See Doukhan (1995, sec. 1.5), or, for a simplified presentation, Durrett (1991, sec. 7.7). □

For the rate of convergence of of $\mathcal{L}(S_n/\sqrt{n})$ to a Gaussian distribution, in the total variation metric, see Doukhan (1995, sec. 1.5.2), summarizing several works. Polynomially-mixing sequences converge polynomially in n, and exponentially-mixing sequences converge exponentially.

There are a number of results on central limit theorems and functional central limit theorems for deterministic dynamical systems. A particularly strong one was recently proved by Tyran-Kamińska (2005), in a friendly paper which should be accessible to anyone who's followed along this far, but it's too long for us to do more than note its existence.