36-754, Advanced Probability II
or
Almost None of Stochastic Processes

Cosma Shalizi

Spring 2006

Contents

Preface: Stochastic Processes in Measure-Theoretic Probability

This is intended to be a *second* course in stochastic processes (at least!); I am going to assume you have all had a first course on stochastic processes, using elementary probability theory, say our 36-703. You might then ask what the added benefit is of taking this course, of re-studying stochastic processes within the framework of measure-theoretic probability. There are a number of reasons to do this.

First, the measure-theoretic framework allows us to greatly generalize the range of processes we can consider. Topics like empirical process theory and stochastic calculus are basically incomprehensible without the measure-theoretic framework. Much of the impetus for developing measure-theoretic probability in the first place came from the impossibility of properly handling continuous random motion, especially the Wiener process, with only the tools of elementary probability.

Second, even topics like Markov processes and ergodic theory, which can be discussed without it, greatly benefit from measure-theoretic probability, because it lets us establish important results which are beyond the reach of elementary methods.

Third, many of the greatest names in twentieth century mathematics have worked in this area, and the theories they have developed are profound, useful and beautiful. Knowing them will make you a better person.

Definitions, lemmas, theorems, corollaries, examples, etc., are all numbered together, consecutively across lectures. Exercises are separately numbered within lectures.

Chapter 1

Basic Definitions: Indexed Collections and Random Functions

Section 1.1 introduces stochastic processes as indexed collections of random variables.

Section 1.2 builds the necessary machinery to consider random functions, especially the product σ-field and the notion of sample paths, and then re-defines stochastic processes as random functions whose sample paths lie in nice sets.

You will have seen, briefly, the definition of a stochastic process in 36-752, but it'll be useful to review it here.

We will flip back and forth between two ways of thinking about stochastic processes: as indexed collections of random variables, and as random functions.

As always, assume we have a nice base probability space $(\Omega, \mathcal{F}, P)$, which is rich enough that all the random variables we need exist.

1.1 So, What Is a Stochastic Process?

Definition 1 (Stochastic Process: As Collection of Random Variables) *A stochastic process $\{X_t\}_{t \in T}$ is a collection of random variables X_t, taking values in a common measure space $(\Xi, \mathcal{X})$, indexed by a set T.*

That is, for each $t \in T$, $X_t(\omega)$ is an $\mathcal{F}/\mathcal{X}$-measurable function from Ω to Ξ, which induces a probability measure on Ξ in the usual way.

It's sometimes more convenient to write $X(t)$ in place of X_t. Also, when $S \subset T$, X_s or $X(S)$ refers to that sub-collection of random variables.

Example 2 *Any single random variable is a (trivial) stochastic process. (Take $T = \{1\}$, say.)*

Example 3 *Let $T = \{1, 2, \ldots k\}$ and $\Xi = \mathbb{R}$. Then $\{X_t\}_{t \in T}$ is a random vector in $\mathbb{R}^k$.*

Example 4 *Let $T = \{1, 2, \ldots\}$ and Ξ be some finite set (or $\mathbb{R}$ or $\mathbb{C}$ or $\mathbb{R}^k \ldots$). Then $\{X_t\}_{t \in T}$ is a one-sided discrete (real, complex, vector-valued, ...) random sequence. Most of the stochastic processes you have encountered are probably of this sort: Markov chains, discrete-parameter martingales, etc.*

Example 5 *Let $T = \mathbb{Z}$ and Ξ be as in Example 4. Then $\{X_t\}_{t \in T}$ is a* two-*sided random sequence.*

Example 6 *Let $T = \mathbb{Z}^d$ and Ξ be as in Example 4. Then $\{X_t\}_{t \in T}$ is a d-dimensional spatially-discrete random field.*

Example 7 *Let $T = \mathbb{R}$ and $\Xi = \mathbb{R}$. Then $\{X_t\}_{t \in T}$ is a real-valued, continuous-time random process (or random motion or random signal).*

Vector-valued processes are an obvious generalization.

Example 8 *Let $T = \mathcal{B}$, the Borel field on the reals, and $\Xi = \overline{\mathbb{R}}^+$, the non-negative extended reals. Then $\{X_t\}_{t \in T}$ is a random set function on the reals.*

The definition of random set functions on $\mathbb{R}^d$ is entirely parallel. Notice that if we want not just a set function, but a measure or a probability measure, this will imply various forms of dependence among the random variables in the collection, e.g., a measure must respect finite additivity over disjoint sets. We will return to this topic in the next section.

Example 9 *Let $T = \mathcal{B} \times \mathbb{N}$ and $\Xi = \overline{\mathbb{R}}^+$. Then $\{X_t\}_{t \in T}$ is a one-sided random sequence of set functions.*

Example 10 (Empirical Processes) *Suppose Z_i, $= 1, 2, \ldots$ are independent, identically-distributed real-valued random variables. (We can see from Example 4 that this is a one-sided real-valued random sequence.) For each Borel set B and each n, define*

$$\hat{P}_n(B) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_B(Z_i)$$

i.e., the fraction of the samples up to time n which fall into that set. This is the empirical measure. $\hat{P}_n(B)$ is a one-sided random sequence of set functions — in fact, of probability measures. We would like to be able to say something about how it behaves. It would be very reassuring, for instance, to be able to show that it converges to the common distribution of the Z_i.

1.2 Random Functions

$X(t,\omega)$ has two arguments, t and ω. For each fixed value of t, $X_t(\omega)$ is straightforward random variable. For each fixed value of ω, however, $X(t)$ is a function from T to Ξ — a *random function*. The advantage of the random function perspective is that it lets us consider the realizations of stochastic processes as single objects, rather than large collections. This isn't just tidier; we will need to talk about *relations* among the variables in the collection or their realizations, rather than just properties of individual variables, and this will help us do so. In Example 10, it's important that we've got random *probability measures*, rather than just random *set functions*, so we need to require that, e.g., $\hat{P}_n(A \cup B) = \hat{P}_n(A) + \hat{P}_n(B)$ when A and B are disjoint Borel sets, and this is a relationship among the three random variables $\hat{P}_n(A)$, $\hat{P}_n(B)$ and $\hat{P}_n(A \cup B)$. Plainly, working out all the dependencies involved here is going to get rather tedious, so we'd like a way to talk about acceptable realizations of the whole stochastic process. This is what the random functions notion will let us do.

We'll make this more precise by defining a random function as a function-valued random variable. To do this, we need a measure space of functions, and a measurable mapping from $(\Omega, \mathcal{F}, P)$ to that function space. To get a measure space, we need a carrier set and a σ-field on it. The natural set to use is Ξ^T, the set of all functions from T to Ξ. (We'll see how to restrict this to just the functions we want presently.) Now, how about the σ-field?

Definition 11 (Cylinder Set) *Given an index set T and a collection of σ-fields $\mathcal{X}_t$ on spaces Ξ_t, $t \in T$. Pick any $t \in T$ and any $A_t \in \mathcal{X}_t$. Then $A_t \times \prod_{s \neq t} \Xi_s$ is a* one-dimensional cylinder set*.*

For any finite k, $k-$dimensional cylinder sets are defined similarly, and clearly are the intersections of k different one-dimensional cylinder sets. To see why they have this name, notice a cylinder, in Euclidean geometry, consists of all the points where the x and y coordinates fall into a certain set (the base), leaving the z coordinate unconstrained. Similarly, a cylinder set like $A_t \times \prod_{s \neq t} \Xi_s$ consists of all the functions in Ξ^T where $f(t) \in A_t$, and are otherwise unconstrained.

Definition 12 (Product σ-field) *The* product σ-field*, $\otimes_{t \in T} \mathcal{X}_t$, is the σ-field over Ξ^T generated by all the one-dimensional cylinder sets. If all the $\mathcal{X}_t$ are the same, $\mathcal{X}$, we write the product σ-field as $\mathcal{X}^T$.*

The product σ-field is enough to let us define a random function, and is going to prove to be *almost* enough for our purposes.

Definition 13 (Random Function; Sample Path) *A Ξ-valued random function on T is a map $X : \Omega \mapsto \Xi^T$ which is $\mathcal{F}/\mathcal{X}^T$-measurable. The realizations of X are functions $x(t)$ taking values in Ξ, called its* sample paths*.*

Definition 14 (Functional of the Sample Path) *Let $E, \mathcal{E}$ be a measure-space. A* functional of the sample path *is a mapping $f : \Xi^T \mapsto E$ which is $\mathcal{X}^T/\mathcal{E}$-measurable.*

Examples of useful and common functionals include maxima, minima, sample averages, etc. Notice that none of these are functions of any one random variable, and in fact their value cannot be determined from any part of the sample path smaller than the whole thing.

Definition 15 (Projection Operator, Coordinate Map) *A* projection operator *or* coordinate map *π_t is a map from Ξ^T to Ξ such that $\pi_t X = X(t)$.*

The projection operators are a convenient device for recovering the individual coordinates — the random variables in the collection — from the random function. Obviously, as t ranges over T, $\pi_t X$ gives us a collection of random variables, i.e., a stochastic process in the sense of our first definition. The following lemma lets us go back and forth between the collection-of-variables, coordinate view, and the entire-function, sample-path view.

Lemma 16 *X is $\mathcal{F}/\otimes_{t\in T}\mathcal{X}_t$-measurable iff $\pi_t X$ is $\mathcal{F}/\mathcal{X}_t$-measurable for every t.*

Proof: This follows from the fact that the one-dimensional cylinder sets generate the product σ-field. □

We have said before that we will want to constrain our stochastic processes to have certain properties — to be probability measures, rather than just set functions, or to be continuous, or twice differentiable, etc. Write the set of all functions in Ξ^T as U. Notice that U does not have to be an element of the product σ-field, and in general is *not.* (We will consider some of the reasons for this later.) As usual, by $U \cap \mathcal{X}^T$ we will mean the collection of all sets of the form $U \cap C$, where $C \in \mathcal{X}^T$. Notice that $(U, U \cap \mathcal{X}^T)$ is a measure space. What we want is to ensure that the sample path of our random function lies in U.

Definition 17 (Stochastic Process: As Random Function) *A Ξ-valued stochastic process on T with paths in U, $U \subseteq \Xi^T$, is a random function $X : \Omega \mapsto U$ which is $\mathcal{F}/U \cap \mathcal{X}^T$-measurable.*

Corollary 18 *A function X from Ω to U is $\mathcal{F}/U \cap \mathcal{X}^T$-measurable iff X_t is $\mathcal{F}/\mathcal{X}$-measurable for all t.*

Proof: Because $X(\omega) \in U$, $X(\omega)$ is $\mathcal{F}/U \cap \mathcal{X}^T$ iff it is $\mathcal{F}/\mathcal{X}^T$-measurable. Then apply Lemma 16. □

Example 19 (Random Measures) *Let $T = \mathcal{B}^d$, the Borel field on $\mathbb{R}^d$, and let $\Xi = \overline{\mathbb{R}}^+$, the non-negative extended reals. Then Ξ^T is the class of set functions on $\mathbb{R}^d$. Let M be the class of such set functions which are also measures (i.e., which are countably additive and give zero on the null set). Then a random set function X with paths in M is a* random measure.

Example 20 (Point Process) *Let X be a random measure, as in the previous example. If $X(B)$ is a finite integer for every bounded Borel set B, then X is a* point process. *If in addition $X(r) \leq 1$ for every $r \in \mathbb{R}^d$, then X is* simple. *The Poisson process is a simple point process.*

Example 21 *Let $T = \mathbb{R}^+$, $\Xi = \mathbb{R}^d$, and $\mathbf{C}(T)$ the class of continuous functions from T to Ξ (in the usual topology). Then a Ξ-valued random process on T with paths in $\mathbf{C}(T)$ is a* continuous random process. *The Wiener process, or Brownian motion, is an example. We will see that most sample paths in $\mathbf{C}(T)$ are not differentiable.*

1.3 Exercises

Exercise 1.1 (The product σ-field answers countable questions) *Let $\mathcal{D} = \bigcup_S \mathcal{X}^S$, where the union ranges over all countable subsets S of the index set T. For any event $D \in \mathcal{D}$, whether or not a sample path $x \in D$ depends on the value of x_t at only a countable number of indices t.*

(a) Show that $\mathcal{D}$ is a σ-field.

(b) Show that if $A \in \mathcal{X}^T$, then $A \in \mathcal{X}^S$ for some countable subset S of T.

Chapter 2

Building Infinite Processes from Finite-Dimensional Distributions

> Section 2.1 introduces the finite-dimensional distributions of a stochastic process, and shows how they determine its infinite-dimensional distribution.
>
> Section 2.2 considers the consistency conditions satisfied by the finite-dimensional distributions of a stochastic process, and the extension theorems (due to Daniell and Kolmogorov) which prove the existence of stochastic processes with specified, consistent finite-dimensional distributions.

2.1 Finite-Dimensional Distributions

So, we now have X, our favorite Ξ-valued stochastic process on T with paths in U. Like any other random variable, it has a probability law or distribution, which is defined over the entire set U. Generally, this is infinite-dimensional. Since it is inconvenient to specify distributions over infinite-dimensional spaces all in a block, we consider the *finite-dimensional distributions*.

Definition 22 (Finite-dimensional distributions) *The finite-dimensional distributions of X are the the joint distributions of $X_{t_1}, X_{t_2}, \ldots X_{t_n}$, $t_1, t_2, \ldots t_n \in T$, $n \in \mathbb{N}$.*

You will sometimes see "FDDs" and "fidis" as abbreviations for "finite-dimensional distributions". Please do not use "fidis".

We can at least hope to specify the finite-dimensional distributions. But we are going to want to ask a lot of questions about asymptotics, and global properties of sample paths, which go beyond any *finite* dimension, so you might worry that we'll still need to deal directly with the infinite-dimensional distribution. The next theorem says that this worry is unfounded; the finite-dimensional distributions specify the infinite-dimensional distribution (pretty much) uniquely.

Theorem 23 *Let X and Y be two Ξ-valued processes on T with paths in U. Then X and Y have the same distribution iff all their finite-dimensional distributions agree.*

Proof: "Only if": Since X and Y have the same distribution, applying the any given set of coordinate mappings will result in identically-distributed random vectors, hence all the finite-dimensional distributions will agree.

"If": We'll use the π-λ theorem.

Let $\mathcal{C}$ be the finite cylinder sets, i.e., all sets of the form

$$C = \left\{x \in \Xi^T | (x_{t_1}, x_{t_2}, \ldots x_{t_n}) \in B\right\}$$

where $n \in \mathbb{N}$, $B \in \mathcal{X}^n$, $t_1, t_2, \ldots t_n \in T$. Clearly, this is a π-system, since it is closed under intersection.

Now let $\mathcal{L}$ consist of all the sets $L \in \mathcal{X}^T$ where $\mathbb{P}(X \in L) = \mathbb{P}(Y \in L)$. We need to show that this is a λ-system, i.e., that it (i) includes Ξ^T, (ii) is closed under complementation, and (iii) is closed under monotone increasing limits. (i) is clearly true: $\mathbb{P}\left(X \in \Xi^T\right) = \mathbb{P}\left(Y \in \Xi^T\right) = 1$. (ii) is true because we're looking at a probability: if $L \in \mathcal{L}$, then $\mathbb{P}(X \in L^c) = 1 - \mathbb{P}(X \in L) = 1 - \mathbb{P}(Y \in L) = \mathbb{P}(Y \in L^c)$. To see (iii), let $L_n \uparrow L$ be a monotone-increasing sequence of sets in $\mathcal{L}$, and recall that, for any measure, $L_n \uparrow L$ implies $\mu L_n \uparrow \mu L$. So $\mathbb{P}(X \in L_n) \uparrow \mathbb{P}(X \in L)$, $\mathbb{P}(Y \in L_n) \uparrow \mathbb{P}(Y \in L)$, and (since $\mathbb{P}(X \in L_n) = \mathbb{P}(Y \in L_n)$), $\mathbb{P}(X \in L_n) \uparrow \mathbb{P}(Y \in L)$ as well. A sequence cannot have two limits, so $\mathbb{P}(X \in L) = \mathbb{P}(Y \in L)$, and $L \in \mathcal{L}$.

Since the finite-dimensional distributions match, $\mathbb{P}(X \in C) = \mathbb{P}(Y \in C)$ for all $C \in \mathcal{C}$, which means that $\mathcal{C} \subset \mathcal{L}$. Also, from the definition of the product σ-field, $\sigma(\mathcal{C}) = \mathcal{X}^T$. Hence, by the $\pi - \lambda$ theorem, $\mathcal{X}^T \subseteq \mathcal{L}$. □

A note of caution is in order here. If X is a Ξ-valued process on T whose paths are constrained to line in U, and Y is a similar process that it is not so constrained, it is nonetheless possible that X and Y agree in all their finite-dimensional distributions. The trick comes if U is not, itself, an element of $\mathcal{X}^T$. The most prominent instance of this is when $\Xi = \mathbb{R}$, $T = \mathbb{R}$, and the constraint is continuity of the sample paths: we will see that $U \notin \mathcal{B}^{\mathbb{R}}$. (This is the point of Exercise 1.1.)

2.2 Consistency and Extension

The finite-dimensional distributions of a given stochastic process are related to one another in the usual way of joint and marginal distributions. Take some collection of indices $t_1, t_2 \ldots t_n \in T$, and corresponding measurable sets $B_1 \in \mathcal{X}_1, B_2 \in \mathcal{X}_2, \ldots B_n \in \mathcal{X}_n$. Then, for any $m > n$, and any further indices

$t_{n+1}, t_{n_2}, \ldots t_m$, it must be the case that

$$\begin{aligned} & \mathbb{P}\left(X_{t_1} \in B_1, X_{t_2} \in B_2, \ldots X_{t_n} \in B_n\right) && (2.1) \\ = \quad & \mathbb{P}\left(X_{t_1} \in B_1, X_{t_2} \in B_2, \ldots X_{t_n} \in B_n, X_{t_{n+1}} \in \Xi, X_{t_{n+2}} \in \Xi, \ldots X_{t_m} \in \Xi\right) \end{aligned}$$

This is going to get really awkward to write over and over, so let's introduce some simplifying notation. $\mathrm{Fin}(T)$ will denote the class of all finite sub-sets of our index set T, and likewise $\mathrm{Denum}(T)$ all denumerable sub-sets. We'll indicate such sub-sets, for the moment, by capital letters like J, K, etc., and extend the definition of coordinate maps (Definition 15) so that π_J maps from Ξ^T to Ξ^J in the obvious way, and π_J^K maps from Ξ^K to Ξ^J, if $J \subset K$. If μ is the measure for the whole process, then the finite-dimensional distributions are $\{\mu_J | J \in \mathrm{Fin}(T)\}$. Clearly, $\mu_J = \mu \circ {\pi_J}^{-1}$.

Definition 24 (Projective Family of Distributions) *A family of distributions μ_J, $J \in \mathrm{Denum}(T)$, is* projective *when for every $J, K \in \mathrm{Denum}(T)$, $J \subset K$ implies*

$$\mu_J = \mu_K \circ \left(\pi_J^K\right)^{-1} \tag{2.2}$$

Such a family is also said to be consistent *or* compatible *(with one another).*

Lemma 25 (FDDs Form Projective Families) *The finite-dimensional distributions of a stochastic process always form a projective family.*

Proof: This is just the fact that we get marginal distributions by integrating out some variables from the joint distribution. But, to proceed formally: Letting J and K be finite sets of indices, $J \subset K$, we know that $\mu_K = \mu \circ {\pi_K}^{-1}$, that $\mu_J = \mu \circ {\pi_J}^{-1}$ and that $\pi_J = \pi_J^K \circ \pi_K$. Hence

$$\begin{aligned} \mu_J &= \mu \circ \left(\pi_J^K \circ \pi_K\right)^{-1} && (2.3) \\ &= \mu \circ \pi_K^{-1} \circ \left(\pi_J^K\right)^{-1} && (2.4) \\ &= \mu_K \circ \left(\pi_J^K\right)^{-1} && (2.5) \end{aligned}$$

as required. □

I claimed that the reason to care about finite-dimensional distributions is that if we specify them, we specify the distribution of the whole process. Lemma 25 says that a putative family of finite dimensional distributions must be consistent, if they are to let us specify a stochastic process. Theorem 23 says that there can't be more than one process distribution with all the same finite-dimensional marginals, but it doesn't guarantee that a given collection of consistent finite-dimensional distributions *can* be extended to a process distribution — it gives uniqueness but not existence. Proving the existence of an extension requires some extra assumptions. Either we need to impose topological conditions on Ξ, or we need to ensure that all the finite-dimensional distributions can be related through conditional probabilities. The first approach is due to Daniell and Kolmogorov, and will finish this lecture; the second is due to Ionescu-Tulcea, and will begin the next.

We'll start with Daniell's theorem on the existence of random sequences, i.e., where the index set is the natural numbers, which uses mathematical induction to extend the finite-dimensional family. To get *there*, we need a useful proposition about our ability to represent non-trivial random variables as functions of uniform random variables on the unit interval.

Proposition 26 (Randomization, transfer) *Let X and X' be identically-distributed random variables in a measurable space Ξ and Y a random variable in a Borel space Υ. Then there exists a measurable function $f : \Xi \times [0,1] \mapsto \Upsilon$ such that $\mathcal{L}(X', f(X', Z)) = \mathcal{L}(X, Y)$, when Z is uniformly distributed on the unit interval and independent of X'.*

Proof: See Kallenberg, Theorem 6.10 (p. 112–113). □

Basically what this says is that if we have two random variables with a certain joint distribution, we can always represent the pair by a copy of one of the variables (X), and a transformation of an independent random number. It is important that Υ be a Borel space here; the result, while very natural-sounding, does not hold for arbitrary measurable spaces, because the proof relies on having a regular conditional probability.

Theorem 27 (Daniell Extension Theorem) *For each $n \in \mathbb{N}$, let Ξ_n be a Borel space, and μ_n be a probability measure on $\prod_{i=1}^{n} \Xi_i$. If the μ_n form a projective family, then there exist random variables $X_i : \Omega \mapsto \Xi_i$, $i \in \mathbb{N}$, such that $\mathcal{L}(X_1, X_2, \ldots X_n) = \mu_n$ for all n, and a measure μ on $\prod_{i=1}^{\infty} \Xi_i$ such that μ_n is equal to the projection of μ onto $\prod i = 1^n \Xi_i$.*

Proof: For any fixed n, $X_1, X_2, \ldots X_n$ is just a random vector with distribution μ_n, and we can always construct such an object. The delicate part here is showing that, when we go to $n+1$, we can use the *same* random elements for the first n coordinates. We'll do this by using the representation-by-randomization proposition just introduced, starting with an IID sequence of uniform random variables on the unit interval, and then transforming them to get a sequence of variables in the Ξ_i which have the right joint distribution. (This is like the quantile transform trick for generating random variates.) The proof will go inductively, so first we'll take care of the induction step, and then go back to reassure ourselves about the starting point.

Induction: Assume we already have $X_1, X_2, \ldots X_n$ such that $\mathcal{L}(X_1, X_2, \ldots X_n) = \mu_n$, and that we have a $Z_{n+1} \sim U(0,1)$ and independent of all the X_i to date. As remarked, we can always get $Y_1, Y_2, \ldots Y_{n+1}$ such that $\mathcal{L}(Y_1, Y_2, \ldots Y_{n+1}) = \mu_{n+1}$. Because the μ_n form a projective family, $\mathcal{L}(Y_1, Y_2, \ldots Y_n) = \mathcal{L}(X_1, X_2, \ldots X_n)$. Hence, by Proposition 26, there is a measurable f such that, if we set $X_{n+1} = f(X_1, X_2, \ldots X_n, Z_{n+1})$, then $\mathcal{L}(X_1, X_2, \ldots X_n, X_{n+1}) = \mu_{n+1}$.

First step: We need there to be an X_1 with distribution μ_1, and we need a (countably!) unlimited supply of IID variables $Z_2, Z_3, \ldots$ all $\sim U(0,1)$. But the existence of X_1 is just the existence of a random variable with a well-defined distribution, which is unproblematic, and the existence of an infinite sequence of IID uniform random variates is too. (See 36-752, or Lemma 3.21 in Kallenberg.)

Finally, to convince yourself of the existence of the measure μ on the product space, recall Lemma 16. □

Remark: Kallenberg, Corollary 6.15, gives a somewhat more abstract version of this theorem.

Daniell's extension theorem works fine for one-sided random sequences, but we often want to work with larger and more interesting index sets. For this we need the full Kolmogorov extension theorem, where the index set T can be completely arbitrary. This in turn needs the Carathéodory Extension Theorem, which I re-state here for convenience.

Theorem 28 (Carathéodory Extension Theorem) *Let μ be a non-negative, finitely additive set function on a field $\mathcal{C}$ of subsets of some space Ω. If μ is also countably additive, then it extends to a measure on $\sigma(\mathcal{C})$, and, if $\mu(\Omega) < \infty$, the extension is unique.*

Proof: See 36-752 lecture notes (Theorem 50, Exercise 51), or Kallenberg, Theorem 2.5, pp. 26–27. Note that "extension" here means extending from a mere field to a σ-field, not from finite to infinite index sets. □

Theorem 29 (Kolmogorov Extension Theorem) *Let Ξ_t, $t \in T$, be a collection of Borel spaces, with σ-fields $\mathcal{X}_i$, and let μ_J, $J \in \mathrm{Fin}(T)$, be a projective family of finite-dimensional distributions on those spaces. Then there exist Ξ_t-valued random variables X_t such that $\mathcal{L}(X_J) = \mu_J$ for all $J \in \mathrm{Fin}(T)$.*

Proof: This will be easier to follow if we first consider the case there T is countable, which is basically Theorem 27 again, and then the general case, where we need Theorem 28.

Countable T: We can, by definition, put the elements of T in $1-1$ correspondence with the elements of $\mathbb{N}$. This in turn establishes a bijection between the product space $\bigotimes_{t \in T} \Xi_t = \Xi_T$ and the sequence space $\bigotimes_{i=1}^{\infty} \Xi_t$. This bijection also induces a projective family of distributions on finite sequences. The Daniell Extension Theorem (27) gives us a measure on the sequence space, which the bijection takes back to a measure on Ξ_T. To see that this μ does not depend on the order in which we arranged T, notice that any two arrangements must give identical results for any finite set J, and then use Theorem 23.

Uncountable T: For each countable $K \subset T$, the argument of the preceding paragraph gives us a measure μ_K on Ξ_K. And, clearly, these μ_K themselves form a projective family. Now let's define a set function μ on the countable cylinder sets, i.e., on the class $\mathcal{D}$ of sets of the form $A \times \Xi_{T \setminus K}$, for some $K \in \mathrm{Denum}(T)$ and some $A \in \mathcal{X}_K$. Specifically, $\mu : \mathcal{D} \mapsto [0, 1]$, and $\mu(A \times \Xi_{T \setminus K}) = \mu_K(A)$. We would like to use Carathéodory's theorem to extend this set function to a measure on the product σ-algebra $\mathcal{X}_T$. First, let's check that the countable cylinder sets form a field: (i) $\Xi_T \in \mathcal{D}$, clearly. (ii) The complement, in Ξ_T, of a countable cylinder $A \times \Xi_{T \setminus K}$ is another countable cylinder, $A^c \times \Xi_{T \setminus K}$. (iii)

The union of two countable cylinders $B_1 = A_1 \times \Xi_{T \setminus K_1}$ and $B_2 = A_2 \times \Xi_{T \setminus K_2}$ is another countable cylinder, since we can always write it as $A \times \Xi_{T \setminus K}$ for some $A \in \mathcal{X}_K$, where $K = K_1 \cup K_2$. Clearly, $\mu(\emptyset) = 0$, so we just need to check that μ is countably additive. So consider any sequence of disjoint cylinder sets $B_1, B_2, \ldots$. Because they're cylinder sets, each i, $B_i = A_i \times \Xi_{T \setminus K_i}$, for some $K_i \in \mathrm{Denum}(T)$, and some $A_i \in \mathcal{X}_{K_i}$. Now set $K = \bigcup_i K_i$; this is a countable union of countable sets, and so itself countable. Furthermore, say $C_i = A_i \times \Xi_{K \setminus K_i}$, so we can say that $\bigcup_i B_i = (\bigcup_i C_i) \times \Xi_{T \setminus K}$. With this notation in place,

$$\begin{aligned}
\mu \bigcup_i B_i &= \mu_K \bigcup_i C_i && (2.6)\\
&= \sum_i \mu_K C_i && (2.7)\\
&= \sum_i \mu_{K_i} A_i && (2.8)\\
&= \sum_i \mu B_i && (2.9)
\end{aligned}$$

where in the second line we've used the fact that μ_K is a probability measure on Ξ_K, and so countably additive on sets like the C_i. This proves that μ is countably additive, so by Theorem 28 it extends to a measure on $\sigma(\mathcal{D})$, the σ-field generated by the countable cylinder sets. But we know from Definition 12 that this σ-field *is* the product σ-field. Since $\mu(\Xi_T) = 1$, Theorem 28 further tells us that the extension is unique. □

Borel spaces are good enough for most of the situations we find ourselves modeling, so the Daniell-Kolmogorov Extension Theorem (as it's often known) see a lot of work. Still, some people dislike having to make topological assumptions to solve probabilistic problems; it seems inelegant. The Ionescu-Tulcea Extension Theorem provides a purely probabilistic solution, available if we can write down the FDDs recursively, in terms of regular conditional probability distributions, even if the spaces where the process has its coordinates are not nice and Borel. Doing this properly will involve our revisiting and extending some ideas about conditional probability, which you will have seen in 36-752, so it will be deferred to the next lecture.

Chapter 3

Building Infinite Processes from Regular Conditional Probability Distributions

Section 3.1 introduces the notion of a probability kernel, which is a useful way of systematizing and extending the treatment of conditional probability distributions you will have seen in 36-752.

Section 3.2 gives an extension theorem (due to Ionescu Tulcea) which lets us build infinite-dimensional distributions from a family of finite-dimensional distributions. Rather than assuming topological regularity of the space, as in Section 2.2, we assume that the FDDs can be derived from one another recursively, through applying probability kernels. This is the same as assuming regularity of the appropriate conditional probabilities.

3.1 Probability Kernels

Definition 30 (Probability Kernel) *A* probability kernel *from a measurable space $\Xi, \mathcal{X}$ to another measurable space $\Upsilon, \mathcal{Y}$ is a function $\kappa : \Xi \times \mathcal{Y} \mapsto [0, 1]$ such that*

1. *for any $Y \in \mathcal{Y}$, $\kappa(x, Y)$ is $\mathcal{X}$-measurable; and*

2. *for any $x \in \Xi$, $\kappa(x, Y) \equiv \kappa_x(Y)$ is a probability measure on $\Upsilon, \mathcal{Y}$. We will write the integral of a function $f : \Upsilon \mapsto \mathbb{R}$, with respect to this measure, as $\int f(y)\kappa(x, dy)$, $\int f(y)\kappa_x(dy)$, or, most compactly, $\kappa f(x)$.*

If condition 1 is satisfied and, for fixed x, $\kappa(x, Y)$ is a measure but not a probability measure, then κ is called a measure kernel *or even just a* kernel.

Notice that we can represent any distribution on Ξ as a kernel where the first argument is irrelevant: $\kappa(x_1, Y) = \kappa(x_2, Y)$ for all $x_1, x_2 \in \Xi$. The "kernels" in kernel density estimation are probability kernels, as are the stochastic transition matrices of Markov chains. (The kernels in support vector machines, however, generally are not.) Regular conditional probabilities, which you will remember from 36-752, are all probability kernels. This fact suggests how we define the composition of kernels.

Definition 31 (Composition of probability kernels) *Let κ_1 be a kernel from Ξ to Υ, and κ_2 a kernel from $\Xi \times \Upsilon$ to Γ. Then we define $\kappa_1 \otimes \kappa_2$ as the kernel from Ξ to $\Upsilon \times \Gamma$ such that*

$$(\kappa_1 \otimes \kappa_2)(x, B) = \int \kappa_1(x, dy) \int \kappa_2(x, y, dz) \mathbf{1}_B(y, z)$$

for every measurable $B \subseteq \Upsilon \times \Gamma$ (where z ranges over the space Γ).

Verbally, κ_1 gives us a distribution on Υ, from any starting point $x \in \Xi$. Given a pair of points $(x, y) \in \Xi \times \Upsilon$, κ_2 gives a distribution on Γ. So their composition says, basically, how to chain together conditional distributions, given a starting point.

3.2 Extension via Recursive Conditioning

With the machinery of probability kernels in place, we are in a position to give an alternative extension theorem, i.e., a different way of proving the existence of stochastic processes with specified finite-dimensional marginal distributions. In Section 2.2, we assumed some topological niceness in the sample spaces, namely that they were Borel spaces. Here, instead, we will assume probabilistic niceness in the FDDs themselves, namely that they can be obtained through composing probability kernels. This is the same as assuming that they can be obtained by chaining together regular conditional probabilities. The general form of this result is attributed in the literature to Ionescu Tulcea.

Just as proving the Kolmogorov Extension Theorem needed a measure-theoretic result, the Carathéodory Extension Theorem, our proof of the Ionescu Tulcea Extension Theorem will require a different measure-theoretic result, which is not, so far as I know, named after anyone.

Proposition 32 *Suppose μ is a finite, non-negative, additive set function on a field $\mathcal{A}$. If, for any sequence of sets $A_n \in \mathcal{A}$, $A_n \downarrow \emptyset \implies \mu A_n \to 0$, then (1) μ is countably additive on $\mathcal{A}$, and (2) μ extends uniquely to a measure on $\sigma(A)$.*

PROOF: Part (1) is a weaker version of Theorem F in Chapter 2, §9 of Halmos, *Measure Theory* (p. 39). (When reading his proof, remember that every field of sets is also a ring of sets.) Part (2) follows from part (1) and the Carathéodory Extension Theorem (28). □

With this preliminary out of the way, let's turn to the main event.

Theorem 33 (Ionescu Tulcea Extension Theorem) *Consider a sequence of measurable spaces $\Xi_n, \mathcal{X}_n$, $n \in \mathbb{N}$. Suppose that for each n, there exists a probability kernel κ_n from $\prod_{i=1}^{n-1} \Xi_i$ to Ξ_n (taking κ_1 to be a kernel insensitive to its first argument, i.e., a probability measure). Then there exists a sequence of random variables X_n, $n \in \mathbb{N}$, taking values in the corresponding Ξ_n, such that $\mathcal{L}(X_1, X_2, \ldots X_n) = \bigotimes_{i=1}^{n} \kappa_i$.*

Proof: As before, we'll be working with the cylinder sets, but now we'll make our life simpler if we consider cylinders where the base set rests in the first n spaces $\Xi_1, \ldots \Xi_n$. More specifically, set $\mathcal{B}_n = \bigotimes_{i=1}^{n} \mathcal{X}_i$ (these are the base sets), and $\mathcal{C}_n = \mathcal{B}_n \times \prod_{i=n+1}^{\infty} \Xi_i$ (these are the cylinder sets), and $\mathcal{C} = \bigcup_n \mathcal{C}_n$. $\mathcal{C}$ clearly contains all the finite cylinders, so it generates the product σ-field on infinite sequences. We will use it as the field in Proposition 32. (Checking that $\mathcal{C}$ is a field is entirely parallel to checking that the $\mathcal{D}$ appearing in the proof of Theorem 29 was a field.)

For each base set $A \in \mathcal{B}_n$, let $[A]$ be the corresponding cylinder, $[A] = A \times \prod_{i=n+1}^{\infty} \Xi_i$. Notice that for every set $C \in \mathcal{C}$, there is at least one A, in some $\mathcal{B}_n$, such that $C = [A]$. Now we define a set function μ on $\mathcal{C}$.

$$\mu([A]) \quad = \quad \left(\bigotimes_{i=1}^{n} \kappa_i\right) A \tag{3.1}$$

(Checking that this is well-defined is left as an exercise, 3.2.) Clearly, this is a finite, and finitely-additive, set function defined on a field. So to use Proposition 32, we just need to check continuity from above at $\emptyset$. Let A_n be any sequence of sets such that $[A_n] \downarrow \emptyset$ and $A_n \in \mathcal{B}_n$. (Any sequence of sets in $\mathcal{C} \downarrow \emptyset$ can be massaged into this form.) We wish to show that $\mu([A_n]) \downarrow 0$. We'll get this to work by considering functions which are (pretty much) conditional probabilities for these sets:

$$p_{n|k} \quad = \quad \left(\bigotimes_{i=k+1}^{n} \kappa_i\right) \mathbf{1}_{A_n},\ k \leq n \tag{3.2}$$

$$p_{n|n} \quad = \quad \mathbf{1}_{A_n} \tag{3.3}$$

Two facts follow immediately from the definitions:

$$p_{n|0} \quad = \quad \left(\bigotimes_{i=1}^{n} \kappa_i\right) \mathbf{1}_{A_n} = \mu([A_n]) \tag{3.4}$$

$$p_{n|k} \quad = \quad \kappa_{k+1} p_{n|k+1} \tag{3.5}$$

From the fact that the $[A_n] \downarrow \emptyset$, we know that $p_{n+1|k} \leq p_{n|k}$, for all k. This implies that $\lim_n p_{n|k} = m_k$ exists, for each k, and is approached from above. Applied to $p_{n|0}$, we see from 3.5 that $\mu([A_n]) \to m_0$. We would like $m_0 = 0$. Assume the contrary, that $m_0 > 0$. From 3.5 and the dominated convergence theorem, we can see that $m_k = \kappa_{k+1} m_{k+1}$. Hence if $m_0 > 0$, $\kappa_1 m_1 > 0$, which means (since that last expression is really an integral) that there is at least one point $x_1 \in \Xi_1$ such that $m_1(s_1) > 0$. Recursing our way down the line, we get a sequence $x = x_1, x_2, \ldots \in \Xi^{\mathbb{N}}$ such that $m_n(x_1, \ldots x_n) > 0$ for all n. But now look what we've done: for each n,

$$\begin{aligned} 0 &< m_n(x_1, \ldots x_n) && (3.6)\\ &\leq p_{n|n}(x_1, \ldots x_n) && (3.7)\\ &= \mathbf{1}_{A_n}(x_1, \ldots x_n) && (3.8)\\ &= \mathbf{1}_{[A_n]}(x) && (3.9)\\ x &\in [A_n] && (3.10) \end{aligned}$$

This is the same as saying that $x \in \bigcap_n [A_n]$. But $[A_n] \downarrow \emptyset$, so there can be no such x. Hence $m_0 = 0$, meaning that $\mu([A_n]) \rightarrow 0$, and μ is continuous at the empty set.

Since μ is finite, finitely-additive, non-negative and continuous at $\emptyset$, by Proposition 32 it extends uniquely to a measure on the product σ-field. □

Notes on the proof: It would seem natural that one could show $m_0 = 0$ directly, rather than by contradiction, but I can't think of a way to do it, and every book I've consulted does it in exactly this way.

To appreciate the simplification made possible by the notion of probability kernels, compare this proof to the one given by Fristedt and Gray (1997, §22.1).

Notice that the Daniell, Kolmogorov and Ionescu Tulcea Extension Theorems all give *sufficient* conditions for the existence of stochastic processes, not necessary ones. The necessary and sufficient condition for extending the FDDs to a process probability measure is something called σ-smoothness. (See Pollard (2002) for details.) Generally speaking, we will deal with processes which satisfy both the Kolmogorov and the Ionescu Tulcea type conditions, e.g., real-valued Markov process.

3.3 Exercises

Exercise 3.1 (Łomnick-Ulam Theorem on infinite product measures) *Let T be an uncountable index set, and $(\Xi_t, \mathcal{X}_t, \mu_t)$ a collection of probability spaces. Show that there exist independent random variables X_t in Ξ_t with distributions μ_t.* Hint: *use the Ionescu Tulcea theorem on countable subsets of T, and then imitate the proof of the Kolmogorov extension theorem.*

Exercise 3.2 *In the proof of the Ionescu Tulcea Theorem, we employed a set function on the finite cylinder sets, where the measure of an infinite-dimensional cylinder set $[A]$ is set equal to the measure of its finite-dimensional base set A. However, the same cylinder set can be specified by different base sets, so it is necessary to show that Equation 3.1 has a unique value on its right-hand side. In what follows, C is an arbitrary member of the class $\mathcal{C}$.*

(i) Show that, when $A, B \in \mathcal{B}_n$, $[A] = [B]$ iff $A = B$. That is, two cylinders generated by bases of equal dimensionality are equal iff their bases are equal.

(ii) Show that there is a smallest n such that $C = [A]$ for an $A \in \mathcal{B}_n$. Conclude that the right-hand side of Equation 3.1 could be made well-defined if we took n there to be this least possible n.

(iii) Suppose that $m < n$, $A \in \mathcal{B}_m$, $B \in \mathcal{B}_n$, and $[A] = [B]$. Show that $B = A \times \prod_{i=m+1}^{n} \Xi_i$.

(iv) Continuing the situation in (iii), show that

$$\left(\bigotimes_{i=1}^{m} \kappa_i\right) A = \left(\bigotimes_{i=1}^{n} \kappa_i\right) B$$

Conclude that the right-hand side of Equation 3.1 is well-defined, as promised.

Chapter 4

One-Parameter Processes, Usually Functions of Time

> Section 4.1 defines one-parameter processes, and their variations (discrete or continuous parameter, one- or two- sided parameter), including many examples.
>
> Section 4.2 shows how to represent one-parameter processes in terms of "shift" operators.

We've been doing a lot of pretty abstract stuff, but the point of this is to establish a common set of tools we can use across many different concrete situations, rather than having to build very similar, specialized tools for each distinct case. Today we're going to go over some examples of the kind of situation our tools are supposed to let us handle, and begin to see how they let us do so. In particular, the two classic areas of application for stochastic processes are dynamics (systems changing over time) and inference (conclusions changing as we acquire more and more data). Both of these can be treated as "one-parameter" processes, where the parameter is time in the first case and sample size in the second.

4.1 One-Parameter Processes

The index set T isn't, usually, an amorphous abstract set, but generally something with some kind of topological or geometrical structure. The number of (topological) dimensions of this structure is the number of *parameters* of the process.

Definition 34 (One-Parameter Process) *A process whose index set T has one dimension is a* one-parameter process*. A process whose index set has more than one dimension is a* multi-parameter process*. A one-parameter process is* discrete *or* continuous *depending on whether its index set is countable or uncountable. A one-parameter process where the index set has a minimal element, otherwise it is* two-sided*.*

$\mathbb{N}$ is a one-sided discrete index set, $\mathbb{Z}$ a two-sided discrete index set, $\mathbb{R}^+$ (including zero!) is a one-sided continuous index set, and $\mathbb{R}$ a two-sided continuous index set.

Most of this course will be concerned with one-parameter processes, which are intensely important in applications. This is because the one-dimensional parameter is usually either time (when we're doing dynamics) or sample size (when we're doing inference), or both at once. There are also some important cases where the single parameter is space.

Example 35 (Bernoulli process) *You all know this one: a one-sided infinite sequence of independent, identically-distributed binary variables, where $X_t = 1$ with probability p, for all t.*

Example 36 (Markov models) *Markov chains are discrete-parameter stochastic processes. They may be either one-sided or two-sided. So are Markov models of order k, and hidden Markov models. Continuous-time Markov processes are, naturally enough, continuous-parameter stochastic processes, and again may be either one-sided or two-sided.*

Instances of physical processes that may be represented by Markov models include: the positions and velocities of the planets; the positions and velocities of molecules in a gas; the pressure, temperature and volume of the gas; the position and velocity of a tracer particle in a turbulent fluid flow; the three-dimensional velocity field of a turbulent fluid; the gene pool of an evolving population. Instances of physical processes that may be represented by hidden Markov models include: the spike trains of neurons; the sonic wave-forms of human speech; many economic and social time-series; etc.

Example 37 ("White Noise") *For each $t \in \mathbb{R}^+$, let $X_t \sim \mathcal{N}(0,1)$, all mutually independent of one another. This is a process with a one-sided continuous parameter.*

It would be character building, at this point, to convince yourself that the process just described exists. (You will need the Kolmogorov Extension Theorem, 29).

Example 38 (Wiener Process) *Here $T = \mathbb{R}^+$ and $\Xi = \mathbb{R}$. The* Wiener process *is the continuous-parameter random process where (1) $W(0) = 0$, (2) for any three times, $t_1 < t_2 < t_3$, $W(t_2) - W(t_1)$ and $W(t_3) - W(t_2)$ are independent (the "independent increments" property), (3) $W(t_2) - W(t_1) \sim \mathcal{N}(0, t_2 - t_1)$ and (4) $W(t, \omega)$ is a continuous function of t for almost all ω. We will spend a* lot *of time with the Wiener process, because it turns out to play a role in the theory of stochastic processes analogous to that played by the Gaussian distribution in elementary probability — the easily-manipulated, formally-nice distribution delivered by limit theorems.*

When we examine the Wiener process in more detail, we will see that it almost never has a derivative. Nonetheless, in a sense which will be made clearer when we come to stochastic calculus, the Wiener process can be regarded as the integral over time of something *very like* white noise, as described in the preceding example.

Example 39 (Logistic Map) *Let $T = \mathbb{N}$, $\Xi = [0, 1]$, $X(0) \sim U(0, 1)$, and $X(t + 1) = aX(t)(1 - X(t))$, $a \in [0, 4]$. This is called the logistic map. Notice that all the randomness is in the initial value $X(0)$; given the initial condition, all later values $X(t)$ are fixed. Nonetheless, this is a Markov process, and we will see that, at least for certain values of a, it satisfies versions of the laws of large numbers and the central limit theorem. In fact, large classes of deterministic dynamical systems have such stochastic properties.*

Example 40 (Symbolic Dynamics of the Logistic Map) *Let $X(t)$ be the logistic map, as in the previous example, and let $S(t) = 0$ if $X(t) \in [0, 0.5)$ and $S(t) = 1$ if $X(t) = [0.5, 1]$. That is, we partition the state space of the logistic map, and record which cell of the partition the original process finds itself in. $X(t)$ is a Markov process, but these "symbolic" dynamics are not necessarily Markovian. We will want to know when functions of Markov processes are themselves Markov. We will also see that there is a sense in which, Markovian or not, this partition is exactly as informative as the original, continuous state — that it is* generating. *Finally, when $a = 4$ in the logistic map, the symbol sequence is actually a Bernoulli process, so that a deterministic function of a completely deterministic dynamical system provides a model of IID randomness.*

Here are some examples where the parameter is sample size.

Example 41 (IID Samples) *Let X_i, $i \in \mathbb{N}$ be samples from an IID distribution, and $Z_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the sample mean. Then Z_n is a one-parameter stochastic process. The point of the ordinary law of large numbers is to reassure us that $Z_n \rightarrow \mathbf{E}[X_n]$ a.s. The point of the central limit theorem is to reassure us that $\sqrt{n}(Z_n - \mathbf{E}[X])$ has constant average size, so that the sampling fluctuation $Z_n - \mathbf{E}[X]$ must be shrinking as $\sqrt{n}$ grows.*

If X_i is the indicator of a set, this convergence means that the relative frequency with which the set is occupied will converge on its true probability.

Example 42 (Non-IID Samples) *Let X_i be a non-IID one-sided discrete-parameter process, say a Markov chain, and again let Z_n be its sample mean, now often called its "time average". The usual machinery of the law of large numbers and the central limit theorem are now inapplicable, and someone who has just taken 36-752 has, strictly speaking, no idea as to whether or not their time averages will converge on expectations. Under the heading of ergodic theory, we will see when this will happen. Since this is the situation with all interesting time series, the application of statistical methods to situations where we cannot contrive to randomize depends crucially on ergodic considerations.*

Example 43 (Estimating Distributions) *Recall Example 10, where we looked at the sequence of empirical distributions $\hat{P}_n$ for samples from an IID data-source. We would like to be able to say that $\hat{P}_n$ converges on P. The usual way to do this, if our samples are of a real-valued random variable, is to consider the empirical cumulative distribution function, F_n. For each n, this may be regarded as a one-parameter random process ($T = \mathbb{R}$, $\Xi = [0, 1]$), and the difficulty is to show that this sequence of random processes converges to F. The usual way is to show that $\sqrt{n}(F_n - F)$, the* empirical process*, converges to a relative of the Wiener process, which in a sense we'll examine later has constant "size"; since $\sqrt{n}$ grows, it follows that $F_n - F$ must shrink. So theorizing even this elementary bit of statistical inference really requires* two *doses of stochastic process theory, one to get a grip on F_n at each n, and the other to get a grip on what happens to F_n as n grows.*

Example 44 (Doob's Martingale) *Let X be a random variable, and $\mathcal{F}_\rangle$, $i \in \mathbb{N}$, a sequence of increasing σ-algebras (i.e. a filtration). Then $Y_i = \mathbf{E}\left[X|\mathcal{F}_\rangle\right]$ is a one-sided discrete-parameter stochastic process, and in fact a martingale. In fact, martingales in general are one-parameter stochastic processes. Note that posterior mean parameter estimates, in Bayesian inference, are an example of Doob's martingale.*

Here are some examples where the one-dimensional parameter is not time or sample size.

Example 45 (The One-Dimensional Ising Model) *This system serves as a toy model of magnetism in theoretical physics. Atoms sit evenly spaced on the points of a regular, infinite, one-dimensional crystalline lattice. Each atom has a magnetic moment, which is either pointing north (+1) or south (−1). Atoms are more likely to point north if their neighbors point north, and vice-versa. The natural index here is $\mathbb{Z}$, so the parameter is discrete and two-sided.*

Example 46 (Text) *Text (at least in most writing systems!) can be represented by a sequence of discrete values at discrete, ordered locations. Since texts can be arbitrarily long, but they all start somewhere, they are discrete-parameter, one-sided processes. Or, more exactly, once we specify a distribution over sequences from the appropriate alphabet, we will have such a process.*

Example 47 (Polymer Sequences) *Similarly, DNA, RNA and proteins are all* heteropolymers *— compounds in which distinct constituent chemicals (the* monomers*) are joined in a sequence. Position along the sequence (chromosome, protein) provides the index, and the nature of the monomer at that position the value.*

Linguists believe that no Markovian model (with finitely many states) can capture human language. Whether this is true of DNA sequences is not known. In both cases, hidden Markov models are used extensively, even if they can only be approximately true of language.

4.2 Operator Representations of One-Parameter Processes

Consider our favorite discrete-parameter process, say X_t. If we try to relate X_t to its *history*, i.e., to the preceding values from the process, we will often get a remarkably complicated probabilistic expression. There is however an alternative, which represents the dynamical part of any process as a remarkably simple semi-group of operators.

Definition 48 (Shift Operators) *Consider Ξ^T, $T = \mathbb{N}$, $= \mathbb{Z}$, $= \mathbb{R}^+$ or $= \mathbb{R}$. The* shift-by-τ *operator Σ_τ, $\tau \geq 0$, maps Ξ^T into itself by shifting forward in time: $(\Sigma_\tau)x(t) = x(t + \tau)$. The collection of all shift operators is the* shift semi-group *or* time-evolution semi-group*.*

(A semi-group does not need to have an identity element, and one which does is technically called a "monoid". No one talks about the shift or time-evolution monoid, however.)

Before we had a Ξ-valued stochastic process X on T, i.e., our process was a random function from T to Ξ. To extract individual random variables, we used the projection operators π_t, which took X to X_t. With the shift operators, we simply have $\pi_t = \pi_0 \circ \Sigma_t$. To represent the passage of time, then, we just apply elements of this semi-group to the function space. Rather than having complicated dynamics which gets us from one value to the next, by working with shifts on function space, all of the complexity is shifted to the initial distribution. This will prove to be extremely useful when we consider stationary processes in the next lecture, and even more useful when, later on, we want to extend the limit theorems from IID sequences to dependent processes.

4.3 Exercises

Exercise 4.1 (Existence of proto-Wiener processes) *Use Theorem 29 and the properties of Gaussian distributions to show that processes exist which satisfy points (1)–(3) of Example 38 (but not necessarily continuity). You will want to begin by finding a way to write down the FDDs recursively.*

Exercise 4.2 (Time-Evolution Semi-Group) *These are all very easy, but worth the practice.*

1. *Verify that the time-evolution semi-group, as described, is a monoid, i.e., that it is closed under composition, that composition is associative, and that there is an identity element. What, in fact, is the identity?*

2. *Can a one-sided process have a shift* group*, rather than just a semi-group?*

3. *Verify that $\pi_\tau = \pi_0 \circ \Sigma_\tau$.*

4. *Verify that, for a discrete-parameter process, $\Sigma_t = (\Sigma_1)^t$, and so Σ_1 generates the semi-group. (For this reason it is often abbreviated to Σ.)*

Chapter 5

Stationary One-Parameter Processes

Section 5.1 describes the three main kinds of stationarity: strong, weak, and conditional.

Section 5.2 relates stationary processes to the shift operators introduced in the last chapter, and to measure-preserving transformations more generally.

5.1 Kinds of Stationarity

Stationary processes are those which are, in some sense, the same at different times — slightly more formally, which are invariant under translation in time. There are three particularly important forms of stationarity: strong or strict, weak, and conditional.

Definition 49 (Strong Stationarity) *A one-parameter process is* strongly stationary *or* strictly stationary *when all its finite-dimensional distributions are invariant under trnaslation of the indices. That is, for all $\tau \in T$, and all $J \in \mathrm{Fin}(T)$,*

$$\mathcal{L}(X_J) = \mathcal{L}(X_{J+\tau}) \tag{5.1}$$

Notice that when the parameter is discrete, we can get away with just checking the distributions of blocks of consecutive indices.

Definition 50 (Weak Stationarity) *A one-parameter process is* weakly stationary *or* second-order stationary *when, for all $t \in T$,*

$$\mathbf{E}[X_t] = \mathbf{E}[X_0] \tag{5.2}$$

and for all $t, \tau \in T$,

$$\mathbf{E}[X_\tau X_{\tau+t}] = \mathbf{E}[X_0 X_t] \tag{5.3}$$

At this point, you should check that a weakly stationary process has time-invariant correlations. (We will say much more about this later.) You should also check that strong stationarity implies weak stationarity. It will turn out that weak and strong stationarity coincide for Gaussian processes, but not in general.

Definition 51 (Conditional (Strong) Stationarity) *A one-parameter process is conditionally stationary if its conditional distributions are invariant under time-translation: $\forall n \in \mathbb{N}$, for every set of $n+1$ indices $t_1, \ldots t_{n+1} \in T$, $t_i < t_{i+1}$, and every shift τ,*

$$\mathcal{L}\left(X_{t_{n+1}} | X_{t_1}, X_{t_2} \ldots X_{t_n}\right) = \mathcal{L}\left(X_{t_{n+1}+\tau} | X_{t_1+\tau}, X_{t_2+\tau} \ldots X_{t_n+\tau}\right) \quad (5.4)$$

(a.s.).

Strict stationarity implies conditional stationarity, but the converse is not true, in general. (Homogeneous Markov processes, for instance, are all conditionally stationary, but most are not stationary.) Many methods which are normally presented using strong stationarity can be adapted to processes which are merely conditionally stationary.[1]

Strong stationarity will play an important role in what follows, because it is the natural generaliation of the IID assumption to situations with dependent variables — we allow for dependence, but the probabilistic set-up remains, in a sense, unchanging. This will turn out to be enough to let us learn a great deal about the process from observation, just as in the IID case.

5.2 Strictly Stationary Processes and Measure-Preserving Transformations

The shift-operator representation of Section 4.2 is particularly useful for strongly stationary processes.

Theorem 52 *A process X with measure μ is strongly stationary if and only if μ is shift-invariant, i.e., $\mu = \mu \circ \Sigma_\tau^{-1}$ for all Σ_τ in the time-evolution semi-group.*

Proof: "If" (invariant distributions imply stationarity): For any finite collection of indices J, $\mathcal{L}(X_J) = \mu \circ \pi_J^{-1}$ (Lemma 25), and similarly $\mathcal{L}(X_{J+\tau}) = \mu \circ \pi_{J+\tau}^{-1}$.

$$\begin{aligned}
\pi_{J+\tau} &= \pi_J \circ \Sigma_\tau & (5.5)\\
\pi_{J+\tau}^{-1} &= \Sigma_\tau^{-1} \circ \pi_J^{-1} & (5.6)\\
\mu \circ \pi_{J+\tau}^{-1} &= \mu \circ \Sigma_\tau^{-1} \circ \pi_J^{-1} & (5.7)\\
\mathcal{L}(X_{J+\tau}) &= \mu \circ \pi_J^{-1} & (5.8)\\
&= \mathcal{L}(X_J) & (5.9)
\end{aligned}$$

[1] For more on conditional stationarity, see Caires and Ferreira (2005).

"Only if": The statement that $\mu = \mu \circ \Sigma_\tau^{-1}$ really means that, for any set $A \in \mathcal{X}^T$, $\mu(A) = \mu(\Sigma_\tau^{-1} A)$. Suppose A is a finite-dimensional cylinder set. Then the equality holds, because all the finite-dimensional distributions agree (by hypothesis). But this means that X and $\Sigma_\tau X$ are two processes with the same finite-dimensional distributions, and so their infinite-dimensional distributions agree (Theorem 23), and the equality holds on all measurable sets A. □

This can be generalized somewhat.

Definition 53 (Measure-Preserving Transformation) *A measurable mapping F from a measurable space $\Xi, \mathcal{X}$ into itself preserves measure μ iff, $\forall A \in \mathcal{X}$, $\mu(A) = \mu(F^{-1}A)$, i.e., iff $\mu = \mu \circ F^{-1}$. This is true just when $F(X) \overset{d}{=} X$, when X is a Ξ-valued random variable with distribution μ. We will often say that F is measure-preserving, without qualification, when the context makes it clear which measure is meant.*

Remark on the definition. It is natural to wonder why we write the defining property as $\mu = \mu \circ F^{-1}$, rather than $\mu = \mu \circ F$. There is actually a subtle difference, and the former is stronger than the latter. To see this, unpack the statements, yielding respectively

$$\forall A \in \mathcal{X},\ \mu(A) = \mu(F^{-1}(A)) \tag{5.10}$$

$$\forall A \in \mathcal{X},\ \mu(A) = \mu(F(A)) \tag{5.11}$$

To see that Eq. 5.10 implies Eq. 5.11, pick any measurable set B, and then apply 5.10 to $F(B)$ (which is $\in \mathcal{X}$, because F is measurable). To go the other way, from 5.11 to 5.10, it would have to be the case that, $\forall A \in \mathcal{X}$, $\exists B \in \mathcal{X}$ such that $A = F(B)$, i.e., every measurable set would have to be the image, under F, of another measurable set. This is not necessarily the case; it would require, for starters, that F be onto (surjective).

Theorem 52 says that every stationary process can be represented by a measure-preserving transformation, namely the shift. Since measure-preserving transformations arise in many other ways, however, it is useful to know about the processes they generate.

Corollary 54 *If F is a measure-preserving transformation on Ξ and X is a Ξ-valued random variable, then the sequence $F^n(X)$, $n \in \mathbb{N}$ is strongly stationary.*

Proof: Consider shifting the sequence $F^n(X)$ by one: the n^{th} term in the shifted sequence is $F^{n+1}(X) = F^n(F(X))$. But since $\mathcal{L}(F(X)) = \mathcal{L}(X)$, by hypothesis, $\mathcal{L}\left(F^{n+1}(X)\right) = \mathcal{L}(F^n(X))$, and the measure is shift-invariant. So, by Theorem 52, the process $F^n(X)$ is stationary.

5.3 Exercises

Exercise 5.1 (Functions of Stationary Processes) *Use Corollary 54 to show that if g is any measurable function on Ξ, then the sequence $g(F^n(X))$ is also* *stationary.*

Exercise 5.2 (Continuous Measure-Preserving Families of Transformations) *Let F_t, $t \in \mathbb{R}^+$, be a semi-group of measure-preserving transformations, with F_0 being the identity. Prove the analog of Corollary 54, i.e., that $F_t(X)$, $t \in \mathbb{R}^+$, is a stationary process.*

Exercise 5.3 (The Logistic Map as an M.P.T.) *The logistic map with $a = 4$ is a measure-preserving transformation, and the measure it preserves has the density $1/\pi\sqrt{x(1-x)}$ (on the unit interval).*

1. *Verify that this density is invariant under the action of the logistic map.*
2. *Simulate the logistic map with* uniformly *distributed X_0. What happens to the density of X_t as $t \to \infty$?*

Chapter 6

Random Times and Their Properties

Section 6.1 recalls the definition of a filtration (a growing collection of σ-fields) and of "stopping times" (basically, measurable random times).

Section 6.2 defines various sort of "waiting' times, including hitting, first-passage, and return or recurrence times.

Section 6.3 proves the Kac recurrence theorem, which relates the finite-dimensional distributions of a stationary process to its mean recurrence times.

6.1 Reminders about Filtrations and Stopping Times

You will have seen these in 36-752 as part of martingale theory, though their application is more general, as we'll see.

Definition 55 (Filtration) *Let T be an ordered index set. A collection $\mathcal{F}_t$, $t \in T$ of σ-algebras is a* filtration *(with respect to this order) if it is non-decreasing, i.e., $f \in \mathcal{F}_t$ implies $f \in mathcalF_s$ for all $s > t$. We generally abbreviate this filtration by $\mathcal{F}$. Define $\mathcal{F}_t^+$ as $\bigcap_{s>t} \mathcal{F}_s$. If $\mathcal{F}^+ = \mathcal{F}$, then $\mathcal{F}$ is* right-continuous.

Recall that we generally think of a σ-algebra as representing available information — for any event $f \in \mathcal{F}$, we can answer the question "did f happen?" A filtration is a way of representing our information about a system growing over time. To see what right-continuity is about, imagine it failed, which would mean $\mathcal{F}_t \subset \bigcap_{s>t} \mathcal{F}_s$. Then there would have to be events which were detectable at *all* times after t, but not at t itself, i.e., some sudden jump in our information right after t. This is what right-continuity rules out.

Definition 56 (Adapted Process) *A stochastic process X on T is* adapted *to a filtration $\mathcal{F}$ if $\forall t$, X_t is $\mathcal{F}_t$-measurable. Any process is adapted to the filtration it induces, $\sigma\{X_s : s \leq t\}$.*

A process being adapted to a filtration just means that, at every time, the filtration gives us enough information to find the value of the process.

Definition 57 (Stopping Time, Optional Time) *An* optional time *or a* stopping time*, with respect to a filtration $\mathcal{F}$, is a T-valued random variable τ such that, for all t,*

$$\{\omega \in \Omega : \tau(\omega) \leq t\} \quad \in \quad \mathcal{F}_t \tag{6.1}$$

If Eq. 6.1 holds with $<$ instead of $\leq$, then τ is weakly optional *or a* weak stopping time.

Basically, all we're doing here is defining what we mean by "a random time at which something detectable happens".

6.2 Waiting Times

"Waiting times" are particular kinds of optional kinds: how much time must elapse before a given event happens, either from a particular starting point, or averaging over all trajectories? Often, these are of particular interest in themselves, and some of them can be related to other quantities of interest.

Definition 58 (Hitting Time) *Given a one-sided Ξ-valued process X, the hitting time τ_B of a measurable set $B \subset \Xi$ is the first time at which $X(t) \in B$;*

$$\tau_B \quad = \quad \inf\{t > 0 : X_t \in B\} \tag{6.2}$$

Example 59 (Fixation through Genetic Drift) *Consider the variation in a given locus (roughly, gene) in an evolving population. If there are k different versions of the gene ("alleles"), the state of the population can be represented by a vector $X(t) \in \mathbb{R}^k$, where at each time $X_i(t) \geq 0$ and $\sum_i X_i(t) = 1$. This set is known as the k-dimensional* probability simplex *S_k. We say that a certain allele has* been fixed *in the population or* gone to fixation *at t if $X_i(t) = 1$ for some i, meaning that all members of the population have that version of the gene. Fixation corresponds to $X(t) \in V$, where V consists of the vertices of the simplex. An important question in evolutionary theory is how long it takes the population to go to fixation. By comparing the actual rate of fixation to that expected under a model of adaptively-neutral genetic drift, it is possible to establish that some genes are under the influence of natural selection.*

Gillespie (1998) is a nice introduction to population genetics, including this problem among many others, using only elementary probability. More sophisticated models treat populations as measure-valued stochastic processes.

Example 60 (Stock Options) *A* stock option[1] *is a legal instrument giving the holder the right to buy a stock at a specified price (the* strike price*, c) before a certain expiration date t_e. The point of the option is that, if you exercise it at a time t when the price of the stock $p(t)$ is above c, you can turn around and sell the stock to someone else, making a profit of $p(t) - c$. When $p(t) > c$, the option is said to be* in money *or* above water*. Options can themselves be sold, and the value of an option depends on how much money it could be used to make, which in turn depends on the probability that it will be "in money" before time t_e. An important part of mathematical finance thus consists of problems of the form "assuming prices $p(t)$ follow a process distribution μ, what is the distribution of hitting times of the set $p(t) > c$?"*

While the financial industry is a major consumer of stochastics, and it has a legitimate role to play in capitalist society, I do hope you will find something more interesting to do with your new-found mastery of random processes, so I will not give many examples of this sort. If you want much, much more, read Shiryaev (1999).

Definition 61 (First Passage Time) *When $\Xi = \mathbb{R}$ or $\mathbb{Z}$, we call the hitting time of the origin the time of* first passage through the origin*, and similarly for other points.*

Definition 62 (Return Time, Recurrence Time) *Fix a set $B \in \Xi$. Suppose that $X(t_0) \in B$. Then the* return time *or* first return time *of B is* recurrence time *of B is* $\inf \{t > t_0 \ : \ X(t) \in B\}$*, and the* recurrence time *θ_B is the difference between the first return time and t_0.*

Note 1: If I'm to be honest with you, I should admit that "return time" and "recurrence time" are used more or less interchangeably in the literature to refer to either the time *coordinate* of the first return (what I'm calling the return time) or the time *interval* which elapses before that return (what I'm calling the recurrence time). I will try to keep these straight here. Check definitions carefully when reading papers!

Note 2: Observe that if we have a discrete-parameter process, and are interested in recurrences of a finite-length sequence of observations $w \in \Xi^k$, we can handle this situation by the device of working with the shift operator in sequence space.

The question of whether any of these waiting times is optional (i.e., measurable) must, sadly, be raised. The following result is generally enough for our purposes.

Proposition 63 *Let X be a Ξ-valued process on a one-sided parameter T, adapted to a filtration $\mathcal{F}$, and let B be an arbitrary measurable set in Ξ. Then τ_B is weakly $\mathcal{F}$-optional under any of the following (sufficient) conditions, and $\mathcal{F}$-optional under the first two:*

[1] Actually, this is just one variety of option (an "American call"), out of a huge variety. I will *not* go into details.

1. *T is discrete.*
2. *T is $\mathbb{R}^+$, Ξ is a metric space, B is closed, and $X(t)$ is a continuous function of t.*
3. *T is $\mathbb{R}^+$, Ξ is a topological space, B is open, and $X(t)$ is right-continuous as a function of t.*

Proof: See, for instance, Kallenberg, Lemma 7.6, p. 123.

6.3 Kac's Recurrence Theorem

For strictly stationary, discrete-parameter sequences, a very pretty theorem, due to Mark Kac (1947), relates the probability of seeing a particular event to the mean time between recurrences of the event. Throughout, we consider an arbitrary Ξ-valued process X, subject only to the requirements of stationarity and a discrete parameter.

Fix an arbitrary measurable set $A \in \Xi$ with $\mathbb{P}(X_1 \in A) > 0$, and consider a new process $Y(t)$, where $Y_t = 1$ if $X_t \in A$ and $Y_t = 0$ otherwise. By Exercise 5.1, Y_t is also stationary. Thus $\mathbb{P}(X_1 \in A, X_2 \not\in A) = \mathbb{P}(Y_1 = 1, Y_2 = 0)$. Let us abbreviate $\mathbb{P}(Y_1 = 0, Y_2 = 0, \ldots Y_{n_1} = 0, Y_n = 0)$ as w_n; this is the probability of making n consecutive observations, none of which belong to the event A. Clearly, $w_n \geq w_{n+1}$. Similarly, let $e_n = \mathbb{P}(Y_1 = 1, Y_2 = 0, \ldots Y_n = 0)$ and $r_n = \mathbb{P}(Y_1 = 1, Y_2 = 0, \ldots Y_n = 1)$ — these are, respectively, the probabilities of starting in A and not returning within $n-1$ steps, and of starting in A and returning for the first time after $n-2$ steps. (Set e_1 to $\mathbb{P}(Y_1 = 1)$, and $w_0 = e_0 = 1$.)

Lemma 64 *The following recurrence relations hold among the probabilities w_n, e_n and r_n:*

$$\begin{aligned} e_n &= w_{n-1} - w_n, \ n \geq 1 && (6.3)\\ r_n &= e_{n-1} - e_n, \ n \geq 2 && (6.4)\\ r_n &= w_{n-2} - 2w_{n-1} + w_n, \ n \geq 2 && (6.5) \end{aligned}$$

Proof: To see the first equality, notice that

$$\begin{aligned} &\mathbb{P}(Y_1 = 0, Y_2 = 0, \ldots Y_{n-1} = 0) && (6.6)\\ &= \mathbb{P}(Y_2 = 0, Y_3 = 0, \ldots Y_n = 0)\\ &= \mathbb{P}(Y_1 = 1, Y_2 = 0, \ldots Y_n = 0) + \mathbb{P}(Y_1 = 0, Y_2 = 0, \ldots Y_n = 0) && (6.7) \end{aligned}$$

using first stationarity and then total probability. To see the second equality, notice that, by total probability,

$$\begin{aligned} &\mathbb{P}(Y_1 = 1, Y_2 = 0, \ldots Y_{n-1} = 0) && (6.8)\\ &= \mathbb{P}(Y_1 = 1, Y_2 = 0, \ldots Y_{n-1} = 0, Y_n = 0) + \mathbb{P}(Y_1 = 1, Y_2 = 0, \ldots Y_{n-1} = 0, Y_n = 1) \end{aligned}$$

The third relationship follows from the first two. □

Theorem 65 (Recurrence in Stationary Processes) *Let X be a Ξ-valued discrete-parameter stationary process. For any set A with $\mathbb{P}(X_1 \in A) > 0$, for almost all ω such that $X_1(\omega) \in A$, there exists a τ for which $X_\tau(\omega) \in A$.*

$$\sum_{k=1}^{\infty} \mathbb{P}(\theta_A = k | X_1 \in A) = 1 \tag{6.9}$$

Proof: The event $\{\theta_A = k, X_1 \in A\}$ is the same as the event $\{Y_1 = 1, Y_2 = 0, \ldots Y_{k+1} = 1\}$. Since $\mathbb{P}(X_1 \in A) > 0$, we can handle the conditional probabilities in an elementary fashion:

$$\begin{aligned}
\mathbb{P}(\theta_A = k | X_1 \in A) &= \frac{\mathbb{P}(\theta_A = k, X_1 \in A)}{\mathbb{P}(X_1 \in A)} && (6.10)\\
&= \frac{\mathbb{P}(Y_1 = 1, Y_2 = 0, \ldots Y_{k+1} = 1)}{\mathbb{P}(Y_1 = 1)} && (6.11)\\
\sum_{k=1}^{\infty} \mathbb{P}(\theta_A = k | X_1 \in A) &= \frac{\sum_{k=1}^{\infty} \mathbb{P}(Y_1 = 1, Y_2 = 0, \ldots Y_{k+1} = 1)}{\mathbb{P}(Y_1 = 1)} && (6.12)\\
&= \frac{\sum_{k=2}^{\infty} r_k}{e_1} && (6.13)
\end{aligned}$$

Now consider the finite sums, and apply Eq. 6.5.

$$\begin{aligned}
\sum_{k=2}^{n} r_k &= \sum_{k=2}^{n} w_{k-2} - 2w_{k-1} + w_k && (6.14)\\
&= \sum_{k=0}^{n-2} w_k + \sum_{k=2}^{n} w_k - 2\sum_{k=1}^{n-1} w_k && (6.15)\\
&= w_0 + w_n - w_1 - w_{n-1} && (6.16)\\
&= (w_0 - w_1) - (w_{n-1} - w_n) && (6.17)\\
&= e_1 - (w_{n-1} - w_n) && (6.18)
\end{aligned}$$

where the last line uses Eq. 6.4. Since $w_{n-1} \geq w_n$, there exists a $\lim_n w_n$, which is ≥ 0 since every individual w_n is. Hence $\lim_n w_{n-1} - w_n = 0$.

$$\begin{aligned}
\sum_{k=1}^{\infty} \mathbb{P}(\theta_A = k | X_1 \in A) &= \frac{\sum_{k=2}^{\infty} r_k}{e_1} && (6.19)\\
&= \lim_{n \to \infty} \frac{e_1 - (w_{n-1} - w_n)}{e_1} && (6.20)\\
&= \frac{e_1}{e_1} && (6.21)\\
&= 1 && (6.22)
\end{aligned}$$

which was to be shown. □

Corollary 66 (Poincaré Recurrence Theorem) *Let F be a transformation which preserves measure μ. Then for any measurable set A, for μ-almost-all $x \in A$, $\exists n \geq 1$ such that $F^n(x) \in A$.*

Proof: A direct application of the theorem, given the relationship between stationary processes and measure-preserving transformations we established in the last lecture. □

Corollary 67 ("Nietzsche") *In the set-up of the previous theorem, if $X_1(\omega) \in A$, then $X_t \in A$ for infinitely many t (a.s.).*

Proof: Repeated application of the theorem yields an infinite sequence of times $\tau_1, \tau_2, \tau_3, \ldots$ such that $X_{\tau_i}(\omega) \in A$, for almost all ω such that $X_1(\omega) \in A$ in the first place. □

Now that we've established that once something happens, it will happen again and again, we would like to know how long we have to wait between recurrences.

Theorem 68 (Kac's Recurrence Theorem) *Continuing the previous notation, $\mathbf{E}[\theta_A | X_1 \in A] = 1/\mathbb{P}(X_1 \in A)$ if and only if $\lim_n w_n = 0$.*

Proof: "If": Unpack the expectation:

$$\mathbf{E}[\theta_A | X_1 \in A] = \sum_{k=1}^{\infty} k \frac{\mathbb{P}(Y_1 = 1, Y_2 = 0, \ldots Y_{k+1} = 1)}{\mathbb{P}(Y_1 = 1)} \tag{6.23}$$

$$= \frac{1}{\mathbb{P}(X_1 \in A)} \sum_{k=1}^{\infty} k r_{k+1} \tag{6.24}$$

so we just need to show that the last series above sums to 1. Using Eq. 6.5 again,

$$\sum_{k=1}^{n} k r_{k+1} = \sum_{k=1}^{n} k(w_{k-1} - 2w_k + w_{k+1}) \tag{6.25}$$

$$= \sum_{k=1}^{n} k w_{k-1} + \sum_{k=1}^{n} k w_{k+1} - 2 \sum_{k=1}^{n} k w_k \tag{6.26}$$

$$= \sum_{k=0}^{n-1} (k+1) w_k + \sum_{k=2}^{n+1} (k-1) w_k - 2 \sum_{k=1}^{n} k w_k \tag{6.27}$$

$$= w_0 + n w_{n+1} - (n+1) w_n \tag{6.28}$$

$$= 1 - w_n - n(w_n - w_{n+1}) \tag{6.29}$$

We therefore wish to show that $\lim_n w_n = 0$ implies $\lim_n w_n + n(w_n - w_{n+1}) = 0$. By hypothesis, it is enough to show that $\lim_n n(w_n - w_{n+1}) = 0$. The partial sums on the left-hand side of Eq. 6.25 are non-decreasing, so $w_n + n(w_n - w_{n+1})$ is non-increasing. Since it is also ≥ 0, the limit $\lim_n w_n + n(w_n - w_{n+1})$ exists;

using $w_n \rightarrow 0$ again, so does $\lim_n w_n + n(w_n - w_{n+1})$. Since $\lim_n w_n$ exists, the series $\sum_{n=1}^{\infty} w_n - w_{n+1}$ must converge, and so $w_n - w_{n+1}$ must be at most $o(n^{-1})$. Hence $\lim_n n(w_n - w_{n+1}) = 0$, as was to be shown.

"Only if": From Eq. 6.29 in the "if" part, we see that the hypothesis is equivalent to

$$1 = \lim_n 1 - w_n - n(w_n - w_{n+1}) \tag{6.30}$$

Since $w_n \geq w_{n+1}$, $1 - w_n - n(w_n - w_{n+1}) \leq 1 - w_n$. We know from the proof of Theorem 65 that $\lim_n w_n$ exists, whether or not it is zero. If it is not zero, then $\lim_n 1 - w_n - n(w_n - w_{n+1}) \leq 1 - \lim_n w_n < 1$. Hence $w_n \rightarrow 0$ is a necessary condition. □

Example 69 *One might imagine that the condition $w_n \rightarrow 0$ in Kac's Theorem is redundant, given the assumption of stationarity. Here is a counter-example. Consider a homogeneous Markov chain on a finite space Ξ, which is partitioned into two non-communicating components, Ξ_1 and Ξ_2. Each component is, internally, irreducible and aperiodic, so there will be an invariant measure μ_1 supported on Ξ_1, and another invariant measure μ_2 supported on Ξ_2. But then, for any $s \in [0,1]$, $s\mu_1 + (1-s)\mu_2$ will also be invariant. (Why?) Picking $A \subset \Xi_2$ gives $\lim_n w_n = s$, the probability that the chain begins in the wrong component to ever reach A.*

Kac's Theorem turns out to be the foundation for a fascinating class of methods for learning the distributions of stationary processes, and for "universal" prediction and data compression. There is also an interesting interaction with large deviations theory. This subject is one possibility for further discussion at the end of the course. Whether or not we get there, let me recommend some papers in the footnote.[2]

6.4 Exercises

Exercise 6.1 (Weakly Optional Times and Right-Continuous Filtrations) *Show that a random time τ is weakly $\mathcal{F}$-optional iff it is $\mathcal{F}^+$-optional.*

Exercise 6.2 (Kac's Theorem for the Logistic Map) *First, do Exercise 5.3. Then, using the same code, suitably modified, numerically check Kac's Theorem for the logistic map with $a = 4$. Pick any interval $I \subset [0,1]$ you like, but be sure not to make it* too *small.*

1. *Generate n initial points in I, according to the invariant measure $\frac{1}{\pi\sqrt{x(1-x)}}$. For each point x_i, find the first t such that $F^t(x_i) \in I$, and take the mean over the sample. What happens to this* space average *as n grows?*

[2] Kontoyiannis *et al.* (1998); "How Sampling Reveals a Process" (Ornstein and Weiss, 1990); Algoet (1992).

2. *Generate a single point x_0 in I, according to the invariant measure. Iterate it N times. Record the successive times $t_1, t_2, \ldots$ at which $F^t(x_0) \in I$, and find the mean of $t_i - t_{i-1}$ (taking $t_0 = 0$). What happens to this* time average *as N grows?*

Chapter 7

Continuity of Stochastic Processes

Section 7.1 describes the leading kinds of continuity for stochastic processes, which derive from the modes of convergence of random variables. It also defines the idea of versions of a stochastic process.

Section 7.2 explains why continuity of sample paths is often problematic, and why we need the whole "paths in U" song-and-dance. As an illustration, we consider a Gausssian process which is close to the Wiener process, except that it's got a nasty non-measurability.

Section 7.3 introduces separable random functions.

7.1 Kinds of Continuity for Processes

Continuity is a convergence property: a continuous function is one where convergence of the inputs implies convergence of the outputs. But we have several kinds of convergence for random variables, so we may expect to encounter several kinds of continuity for random processes. Note that the following definitions are stated broadly enough that the index set T does not have to be one-dimensional.

Definition 70 (Continuity in Mean) *A stochastic process X is* continuous in the mean at t_0 *if $t \to t_0$ implies $\mathbf{E}\left[|X(t) - X(t_0)|^2\right] \to 0$. X is continuous in the mean if this holds for all t_0.*

It would, of course, be more natural to refer to this as "continuity in mean *square*", or even "continuity in L_2", and one can define continuity in L_p for arbitrary p.

Definition 71 (Continuity in Probability) *X is* continuous in probability *at t_0 if $t \to t_0$ implies $X(t) \xrightarrow{P} X(t_0)$. X is* continuous in probability *or* stochastically continuous *if this holds for all t_0.*

Note that neither L_p-continuity nor stochastic continuity says that the individual sample paths, themselves, are continuous.

Definition 72 (Continuous Sample Paths) *A process X is* continuous *at t_0 if, for almost all ω, $t \rightarrow t_0$ implies $X(t, \omega) \rightarrow X(t_0, \omega)$. A process is* continuous *if, for almost all ω, $X(\cdot, \omega)$ is a continuous function.*

Obviously, continuity of sample paths implies stochastic continuity and L_p-continuity.

A weaker pathwise property than strict continuity, frequently used in practice, is the combination of continuity from the right with limits from the left. This is usually known by the term "cadlag", abbreviating the French phrase "continues à droite, limites à gauche"; "rcll" is an unpronounceable synonym.

Definition 73 (Cadlag) *A sample function x on a well-ordered set T is* cadlag *if it is continuous from the right and limited from the left at every point. That is, for every $t_0 \in T$, $t \downarrow t_0$ implies $x(t) \rightarrow x(t_0)$, and for $t \uparrow t_0$, $\lim_{t \uparrow t_0} x(t)$ exists, but need not be $x(t_0)$. A stochastic process X is cadlag if almost all its sample paths are cadlag.*

As we will see, it will not be easy to show that our favorite random processes have any of these desirable properties. What will be easy will be to show that they are, in some sense, easily modified into ones which do have good regularity properties, without loss of probabilistic content. This is made more precise by the notion of *versions* of a stochastic process, related to that of versions of conditional probabilities.

Definition 74 (Versions of a Stochastic Process) *Two stochastic processes X and Y with a common index set T are called* versions *of one another if*

$$\forall t \in T,\ \mathbb{P}\left(\omega :\ X(t, \omega) = Y(t, \omega)\right) = 1$$

Such processes are also said to be stochastically equivalent.

Lemma 75 *If X and Y are versions of one another, they have the same finite-dimensional distributions.*

Proof: Clearly it will be enough to show that $\mathbb{P}(X_J = Y_J) = 1$ for arbitrary finite collections of indices J. Pick any such collection $J = \{t_1, t_2, \ldots t_j\}$. Then

$$\begin{aligned}
\mathbb{P}(X_J = Y_J) &= \mathbb{P}\left(X_{t_1} = Y_{t_1}, \ldots X_{t_j} = Y_{t_j}\right) && (7.1)\\
&= 1 - \mathbb{P}\left(\bigcup_{t_i \in J} X_{t_i} \neq Y_{t_i}\right) && (7.2)\\
&\geq 1 - \sum_{t_i \in J} \mathbb{P}\left(X_{t_i} \neq Y_{t_i}\right) && (7.3)\\
&= 1 && (7.4)
\end{aligned}$$

using only *finite* sub-additivity. □

There is a stronger notion of similarity between processes than that of versions, which will sometimes be useful.

Definition 76 (Indistinguishable Processes) *Two stochastic processes X and Y are* indistinguishable*, or* equivalent up to evanescence*, when*

$$\mathbb{P}\left(\omega : \ \forall t, X(t,\omega) = Y(t,\omega)\right) = 1$$

Notice that saying X and Y are indistinguishable means that their sample paths are equal almost surely, while saying they are versions of one another means that, at any time, they are almost surely equal. Indistinguishable processes are versions of one another, but not necessarily the reverse. (Look at where the quantifier and the probability statements go.) However, if $T = \mathbb{R}^d$, then any two right-continuous versions of the same process are indistinguishable (Exercise 7.2).

7.2 Why Continuity Is an Issue

In many situations, we want to use stochastic processes to model dynamical systems, where we know that the dynamics are continuous in time (i.e. the index set is $\mathbb{R}$, or maybe $\mathbb{R}^+$ or $[0,T]$ for some real T).[1] This means that we ought to restrict the sample paths to be continuous functions; in some cases we'd even want them to be differentiable, or yet more smooth. As well as being a matter of physical plausibility or realism, it is also a considerable mathematical convenience, as the following shows.

Proposition 77 *Let $X(t,\omega)$ be a real-valued continuous-parameter process with continuous sample paths. Then on any finite interval I, $M(\omega) \equiv \sup_{t\in I} X(t,\omega)$ and $m(\omega) \equiv \inf_{t\in I} X(t,\omega)$ are measurable random variables.*

Proof: It'll be enough to prove this for the supremum function M; the proof for m is entirely parallel. First, notice that $M(\omega)$ must be finite, because the sample paths $X(\cdot,\omega)$ are continuous functions, and continuous functions are bounded on bounded intervals. Next, notice that $M(\omega) > a$ if and only if $X(t,\omega) > a$ for some $t \in I$. But then, by continuity, there will be some rational $t' \in I \cap \mathbf{Q}$ such that $X(t',\omega) > a$; countably many, in fact.[2] Hence

$$\{\omega : \ M(\omega) > a\} = \bigcup_{t \in I\cap\mathbf{Q}} \{\omega : X(t,\omega) > a\}$$

[1] Strictly speaking, we don't *really* know that space-time is a continuum, but the discretization, if there is one, is so fine that it might as well be.

[2] Continuity means that we can pick a δ such that, for all t' within δ of t, $X(t',\omega)$ is within $\frac{1}{2}(X(t,\omega) - a)$ of $X(t,\omega)$. And there are countably many rational numbers within any real interval. — There is nothing special about the rational numbers here; any countable, dense subset of the real numbers would work as well.

Since, for each t, $X(t,\omega)$ is a random variable, the sets in the union on the right-hand side are all measurable, and the union of a countable collection of measurable sets is itself measurable. Since intervals of the form (a,∞) generate the Borel σ-field on the reals, we have shown that $M(\omega)$ is a measurable function from Ω to the reals, i.e., a random variable. □

Continuity raises some very tricky technical issues. The product σ-field is the usual way of formalizing the notion that what we know about a stochastic process are values observed at certain particular times. What we saw in Exercise 1.1 is that "the product σ-field answers countable questions": for any measurable set A, whether $x(\cdot,\omega) \in A$ depends only on the value of $x(t,\omega)$ at countably many indices t. It follows that the class of all continuous sample paths is *not* product-σ-field measurable, because $x(\cdot,\omega)$ is continuous at t iff $x(t_n,\omega) \rightarrow x(t,\omega)$ along *every* sequence $t_n \rightarrow t$, and this is involves the value of the function at uncountably many coordinates. It is further true that the class of differentiable functions is not product σ-field measurable. For that matter, neither is the class of piecewise linear functions! (See Exercise 7.1.)

You might think that, on the basis of Theorem 23, this should not *really* be much of an issue: that even if the class of continuous sample paths (say) isn't strictly measurable, it could be well-approximated by measurable sets, and so getting the finite-dimensional distributions right is all that matters. This would make the theory of stochastic processes in continuous time much simpler, but unfortunately it's not quite the case. Here is an example to show just how bad things can get, even when all the finite-dimensional distributions agree.[3]

Example 78 (A Horrible Version of the proto-Wiener Process) *Example 38 defined the Wiener process by four requirements: starting at the origin, independent increments, a Gaussian distribution of increments, and continuity of sample paths. Take a Wiener process $W(t,\omega)$ and consider $M(\omega) \equiv \sup_{t\in[0,1]} W(t,\omega)$, its supremum over the unit interval. By the preceding proposition, we know that M is a measurable random variable. But we can construct a version of W for which the supremum is not measurable.*

For starters, assume that Ω can be partitioned into an uncountable collection of disjoint measurable sets, one for each $t \in [0,1]$. (This can be shown as an exercise in real analysis.) Select any non-measurable real-valued function $B(\omega)$, so long as $B(\omega) > M(\omega)$ for all ω. (There are uncountably many suitable functions.) Set $W^*(t,\omega) = W(t,\omega)$ if $\omega \not\in \Omega_t$, and $= B(\omega)$ if $\omega \in \Omega_t$. Now, at every t, $\mathbb{P}\left(W(t,\omega) = W^*(t,\omega)\right) = 1$. W^* is a version of W, and all their finite-dimensional distributions. But, for every ω, there is a t such that $W^*(t,\omega) = B(\omega) > \sup_t W(t,\omega)$, so $\sup_t W^*(t,\omega) = B(\omega)$, which by design is non-measurable.[4]

[3] I stole this example from Pollard (2002, p. 214).

[4] Note that the Wiener process is an important model for the price of a stock in financial theory (more exactly, for the log of its price), and its maximum over an interval is closely related to the value of an option on that stock, so this is something you really want to be able to make probability statements about.

Fundamentally, the issues with continuity are symptoms of a deeper problem. The reason the supremum function is non-measurable in the example is that it involves uncountably many indices. A countable collection of ill-behaved sets of measure zero is a set of measure zero, and may be ignored, but an uncountable collection of them can have probability 1. Fortunately, there are standard ways of evading these measure-theoretic issues, by showing that one can always find random functions which not only have prescribed finite-dimensional distributions (what we did in Lectures 2 and 3), but also are regular enough that we can take suprema, or integrate over time, or force them to be continuous. This hinges on the notion of *separability* for random functions.

7.3 Separable Random Functions

The basic idea of a *separable* random function is one whose properties can be handled by dealing only with a countable, dense subset, just as, in the proof of Proposition 77, we were able to get away with only looking at $X(t)$ at only rational values of t. Because a space with a countable, dense subset is called a "separable" space, we will call such functions "separable functions".

Definition 79 (Separable Functions) *Let Ξ and T be metric spaces, and D be a countable, dense subset of T. A function $x : T \mapsto \Xi$ is* D-separable *or* separable with respect to D *if, $forall t \in T$, there exists a sequence $t_i \in D$ such that $t_i \to t$ and $x(t_i) \to x(t)$.*

Lemma 80 *The following conditions are sufficient for separability:*

1. *T is countable.*
2. *x is continuous.*
3. *T is well-ordered and x is right-continuous.*

Proof: (1) Take the separating set to be T itself. (2) Pick any countable dense D. By density, for every t there will be a sequence $t_i \in D$ such that $t_i \to t$. By continuity, along any sequence converging to t, $x(t_i) \to t$. (3) Just like (2), only be sure to pick the $t_i > t$. (You can do this, again, for any countable dense D.) □

Definition 81 (Separable Process) *A Ξ-valued process X on T is* separable with respect to D *if D is a countable, dense subset of T, and there is a measure-zero set $N \subset \Omega$ such that for every $\omega \notin N$, $X(\cdot, \omega)$ is D-separable. That is, $X(\cdot, \omega)$ is almost surely D-separable.*

We cannot easily guarantee that a process is separable. What we can easily do is go from one process, which may or may not be separable, to a separable process with the same finite-dimensional distributions. This is known as a *separable modification* of the original process. Combined with the extension theorems (Theorems 27, 29 and 33), this tells that we can always construct a separable process with desired finite-dimensional distributions. We shall therefore feel entitled to assume that our processes *are* separable, without further ado. The proofs of the existence of separable and continuous versions of general processes are, however, somewhat involved, and so postponed to the next lecture.

7.4 Exercises

Exercise 7.1 *Consider real-valued functions on the unit interval (i.e., $\Xi = \mathbb{R}$, $T = [0,1]$, $\mathcal{X} = \mathcal{B}$). The product σ-field is thus $\mathcal{B}^{[0,1]}$. In many circumstances, it would be useful to constrain sample paths to be piece-wise linear functions of the index. Let $\mathbf{PL}([0,1])$ denote this class of functions. Use the argument of Exercise 1.1 to show that $\mathbf{PL}([0,1]) \not\in \mathcal{B}^{[0,1]}$.*

Exercise 7.2 *Show that, if X and Y are versions of one another, with index set $\mathbb{R}^d$, and both are right-continuous, then they are indistinguishable.*

Chapter 8

More on Continuity

> Section 8.1 constructs separable modifications of reasonable but non-separable random functions, and explains how separability relates to non-denumerable properties like continuity.
>
> Section 8.2 constructs versions of our favorite one-parameter processes where the sample paths are measurable functions of the parameter.
>
> Section 8.3 gives conditions for the existence of cadlag versions.
>
> Section 8.4 gives some criteria for continuity, and for the existence of "continuous modifications" of discontinuous processes.

Recall the story so far: last time we saw that the existence of processes with given finite-dimensional distributions does not guarantee that they have desirable and natural properties, like continuity, and in fact that one can construct discontinuous versions of processes which *ought* to be continuous. We therefore need extra theorems to guarantee the existence of *continuous* versions of processes with specified FDDs. To get there, we will first prove the existence of *separable* versions. This will require various topological conditions on both the index set T and the value space Ξ.

In the interest of space (or is it time?), Section 8.1 will provide complete and detailed proofs. The other sections will simply state results, and refer proofs to standard sources, mostly Gikhman and Skorokhod (1965/1969). (They in turn follow Doob (1953), but are explicit about what he regarded as obvious generalizations and extensions, *and* they cost about \$20, whereas Doob costs \$120 in paperback.)

8.1 Separable Versions

We can show that separable versions of our favorite stochastic processes exist under quite general conditions, but first we will need some preliminary results, living at the border between topology and measure theory. This starts by recalling some facts about compact spaces.

Definition 82 (Compactness, Compactification) *A set A in a topological space Ξ is* compact *if every covering of A by open sets contains a finite sub-cover. Ξ is a* compact space *if it is itself a compact set. Every non-compact topological space Ξ is a sub-space of some compact topological space $\tilde{\Xi}$. The super-space $\tilde{\Xi}$ is a* compactification *of Ξ. Every compact metric space is separable.*[1]

Example 83 *The real numbers $\mathbb{R}$ are not compact: they have no finite covering by open intervals (or other open sets). The extended reals, $\overline{\mathbb{R}} \equiv \mathbb{R} \cup +\infty \cup -\infty$, are compact, since intervals of the form $(a, \infty]$ and $[-\infty, a)$ are open. This is a* two-point compactification *of the reals. There is also a one-point compactification, with a single point at $\pm\infty$, but this has the undesirable property of making big negative and positive numbers close to each other.*

Recall that a random function is separable if its value at any arbitrary index can be determined almost surely by examining its values on some fixed, countable collection of indices. The next lemma states an alternative characterization of separability. The lemma after that gives conditions under which a weaker property holds — the almost-sure determination of whether $X(t, \omega) \in B$, for a specific t and set B, by the behavior of $X(t_n, \omega)$ at countably many t_n. The final lemma extends this to large collections of sets, and then the proof of the theorem puts all the parts together.

Lemma 84 *Let T be a separable set, Ξ a compact metric space, and D a countable dense subset of T. Define V as the class of all open balls in T centered at points in D and with rational radii. For any $G \subset T$, let*

$$R(G, \omega) \equiv \text{closure}\left(\bigcup_{t \in G \cap D} X(t, \omega)\right) \quad (8.1)$$

$$R(t, \omega) \equiv \bigcap_{S:\ S \in V,\ t \in S} R(S, \omega) \quad (8.2)$$

Then $X(t, \omega)$ is D-separable if and only if there exists a set $N \subset \Omega$ such that

$$\omega \notin N \Rightarrow \forall t,\ X(t, \omega) \in R(t, \omega) \quad (8.3)$$

and $\mathbb{P}(N) = 0$.

Proof: Roughly speaking, $R(t, \omega)$ is what we'd think the range of the function would be, in the vicinity of t, if it we went just by what it did at points in the separating set D. The actual value of the function falling into this range (almost surely) is necessary and sufficient for the function to be separable. But let's speak less roughly.

"Only if": Since $X(t, \omega)$ is D-separable, for almost all ω, for any t there is some sequence $t_n \in D$ such that $t_n \to t$ and $X(t_n, \omega) \to X(t, \omega)$. For any

[1] This last statement requires the axiom of choice.

ball S centered at t, there is some N such that $t_n \in S$ if $n \geq N$. Hence the values of $x(t_n)$ are eventually confined to the set $\bigcup_{t \in S \cap D} X(t,\omega)$. Recall that the closure of a set A consists of the points x such that, for some sequence $x_n \in A$, $x_n \rightarrow x$. As $X(t_n,\omega) \rightarrow X(t,\omega)$, it must be the case that $X(t,\omega) \in \text{closure}\left(\bigcup_{t \in S \cap D} X(t,\omega)\right)$. Since this applies to all S, $X(t,\omega)$ must be in the intersection of all those closures, hence $X(t,\omega) \in R(t,\omega)$ — unless we are on one of the probability-zero bad sample paths, i.e., unless $\omega \in N$.

"If": Assume that, with probability 1, $X(t,\omega) \in R(t,\omega)$. Thus, for any $S \in V$, we know that there exists a sequence of points $t_n \in S \cap D$ such that $X(t_n,\omega) \rightarrow X(t,\omega)$. However, this doesn't say that $t_n \rightarrow t$, which is what we need for separability. We will now build such a sequence. Consider a series of spheres $S_k \in V$ such that (i) every point in S_k is within a distance 2^{-k} of t and (ii) $S_{k+1} \subset S_k$. For each S_k, there is a sequence $t_n^{(k)} \in S_k$ such that $X(t_n^{(k)},\omega) \rightarrow X(t,\omega)$. In fact, for any $m > 0$, $|X(t_n^{(k)},\omega) - X(t,\omega)| < 2^{-m}$ if $n \geq N(k,m)$, for some $N(k,m)$. Our final sequence of indices t_i then consists of the following points: $t_n^{(1)}$ for n from $N(1,1)$ to $N(1,2)$; $t_n^{(2)}$ for n from $N(2,2)$ to $N(2,3)$; and in general $t_n^{(k)}$ for n from $N(k,k)$ to $N(k,k+1)$. Clearly, $t_i \rightarrow t$, and $X(t_i,\omega) \rightarrow X(t,\omega)$. Since every $t_i \in D$, we have shown that $X(t,\omega)$ is D-separable. □

Lemma 85 *Let T be a separable index set, Ξ a compact space, X a random function from T to Ξ, and B be an arbitrary Borel set of Ξ. Then there exists a denumerable set of points $t_n \in T$ such that, for any $t \in T$, the set*

$$N(t,B) \equiv \{\omega : X(t,\omega) \notin B\} \cap \left(\bigcap_{n=1}^{\infty} \{\omega : X(t_n,\omega) \in B\}\right) \tag{8.4}$$

has probability 0.

Proof: We proceed recursively. The first point, t_1, can be whatever we like. Suppose $t_1, t_2, \ldots t_n$ are already found, and define the following:

$$M_n \equiv \bigcap_{k=1}^{n} \{\omega : X(t_k,\omega) \in B\} \tag{8.5}$$

$$L_n(t) \equiv M_n \cap \{\omega : X(t,\omega) \notin B\} \tag{8.6}$$

$$p_n \equiv \sup_t \mathbb{P}(L_n(t)) \tag{8.7}$$

M_n is the set where the random function, evaluated at the first n indices, gives a value in our favorite set; it's clearly measurable. $L_n(t)$, also clearly measurable, gives the collection of points in Ω where, if we chose t for the next point in the collection, this will break down. p_n is the worst-case probability of this happening. For each t, $L_{n+1}(t) \subseteq L_n(t)$, so $p_{n+1} \leq p_n$. Suppose $p_n = 0$; then we've found the promised denumerable sequence, and we're done. Suppose instead that $p_n > 0$. Pick any t such that $\mathbb{P}(L_n(t)) \geq \frac{1}{2}p_n$, and call it t_{n+1}. (There has to be such a point, or else p_n wouldn't be the supremum.) Now notice that $L_1(t_2)$, $L_2(t_3)$, $\ldots L_n(t_{n+1})$ are all mutually exclusive, but not necessarily jointly exhaustive. So

$$\begin{aligned}
1 &= \mathbb{P}(\Omega) \qquad (8.8)\\
&\geq \mathbb{P}\left(\bigcup_n L_n(t_{n+1})\right) \qquad (8.9)\\
&= \sum_n \mathbb{P}(L_n(t_{n+1})) \qquad (8.10)\\
&\geq \sum_n \frac{1}{2} p_n > 0 \qquad (8.11)
\end{aligned}$$

so $p_n \to 0$ as $n \to \infty$.

We saw that $L_n(t)$ is a monotone-decreasing sequence of sets, for each t, so a limiting set exists, and in fact $\lim_n L_n(t) = N(t, B)$. So, by monotone convergence,

$$\begin{aligned}
\mathbb{P}(N(t,B)) &= \mathbb{P}\left(\lim_n L_n(t)\right) \qquad (8.12)\\
&= \lim_n \mathbb{P}(L_n(t)) \qquad (8.13)\\
&\leq \lim_n p_n \qquad (8.14)\\
&= 0 \qquad (8.15)
\end{aligned}$$

as was to be shown. □

Lemma 86 *Let $\mathcal{B}_0$ be any countable class of Borel sets in Ξ, and $\mathcal{B}$ the closure of $\mathcal{B}_0$ under countable intersection. Under the hypotheses of the previous lemma, there is a denumerable sequence t_n such that, for every $t \in T$, there exists a set $N(t) \subset \Omega$ with $\mathbb{P}(N(t)) = 0$, and, for all $B \in \mathcal{B}$,*

$$\{\omega :\ X(t,\omega) \notin A\} \cap \left(\bigcap_{n=1}^{\infty} \{\omega :\ X(t_n,\omega) \in A\}\right) \subseteq N(t) \qquad (8.16)$$

Proof: For each $B \in \mathcal{B}_0$, construct the sequence of indices as in the previous lemma. Since there only countably many sets in B, if we take the union of all of these sequences, we will get another countable sequence, call it t_n. Then we have that, $\forall B \in \mathcal{B}_0$, $\forall t \in T$, $\mathbb{P}(X(t_n,\omega) \in B, n \geq 1,\ X(t,\omega) \notin B) = 0$. Take this set to be $N(t,B)$, and define $N(t) \equiv \bigcup_{B \in \mathcal{B}_0} N(t,B)$. Since $N(t)$ is a countable union of probability-zero events, it is itself a probability-zero event. Now, take any $B \in \mathcal{B}$, and any $B_0 \in \mathcal{B}_0$ such that $B \subseteq B_0$. Then

$$\begin{aligned}
&\{X(t,\omega) \notin B_0\} \cap \left(\bigcap_{n=1}^{\infty} \{X(t_n,\omega) \in B\}\right) \qquad (8.17)\\
&\subseteq \{X(t,\omega) \notin B_0\} \cap \left(\bigcap_{n=1}^{\infty} \{X(t_n,\omega) \in B_0\}\right)\\
&\subseteq N(t) \qquad (8.18)
\end{aligned}$$

Since $B = \bigcap_k B_0^{(k)}$ for some sequence of sets $B_0^{(k)} \in \mathcal{B}_0$, it follows (via De Morgan's laws and the distributive law) that

$$\{X(t,\omega) \not\in B\} \quad = \quad \bigcup_{k=1}^{\infty} \left\{X(t,\omega) \not\in B_0^{(k)}\right\} \tag{8.19}$$

$$\begin{aligned}
&\{X(t,\omega) \not\in B\} \cap \left(\bigcap_{n=1}^{\infty} \{X(t_n,\omega) \in B\}\right) \\
&= \bigcup_{k=1}^{\infty} \left\{X(t,\omega) \not\in B_0^{(k)}\right\} \cap \left(\bigcap_{n=1}^{\infty} \{X(t_n,\omega) \in B\}\right) && (8.20)\\
&\subseteq \bigcup_{n=1}^{\infty} N(t) && (8.21)\\
&= N(t) && (8.22)
\end{aligned}$$

which was to be shown. □

Theorem 87 (Separable Versions, Separable Modifications) *Suppose that Ξ is a compact metric space and T is a separable metric space. Then, for any Ξ-valued stochastic process X on T, there exists a separable version $\tilde{X}$. This is called a* separable modification *of X.*

Proof: Let D be a countable dense subset of T, and V the class of open spheres of rational radius centered at points in D. Any open subset of T is a union of countably many sets from V, which is itself countable. Similarly, let C be a countable dense subset of Ξ, and let $\mathcal{B}_0$ consist of the complements of spheres centers at points in D with rational radii, and (as in the previous lemma) let $\mathcal{B}$ be the closure of $\mathcal{B}_0$ under countable intersection. Every closed set in Ξ belongs to $\mathcal{B}$.[2] For every $S \in V$, consider the restriction of $X(t,\omega)$ to $t \in S$, and apply Lemma 86 to the random function $X(t,\omega)$ to get a sequence of indices $I(S) \subset T$, and, for every $t \in S$, a measure-zero set $N_S(t) \subset \Omega$ where things can go wrong. Set $I = \bigcup_{S \in V} I(S)$ and $N(t) = \bigcup_{S \in V} N_S(t)$. Because V is countable, I is still a countable set of indices, and $N(t)$ is still of measure zero. I is going to be our separating set, and we're going to show that we have uncountably many sets $N(t)$ won't be a problem.

Define $\tilde{X}(t,\omega) = X(t,\omega)$ if $t \in I$ or $\omega \not\in N(t)$ — if we're at a time in the separating set, or we're at some other time but have avoided the bad set, we'll just copy our original random function. What to do otherwise, when $t \not\in I$ and $\omega \in N(t)$? Construct $R(t,\omega)$, as in the proof of Lemma 84, and let $\tilde{X}(t,\omega)$ take *any* value in this set. Since $R(t,\omega)$ depends only on the value of the function at indices in the separating set, it doesn't matter whether we build it from X or from $\tilde{X}$. In fact, for all t and ω, $\tilde{X}(t,\omega) \in R(t,\omega)$, so, by Lemma 84,

[2] *You* show this.

$\tilde{X}(t,\omega)$ is separable. Finally, for every t, $\left\{\tilde{X}(t,\omega) = X(t,\omega)\right\} \subseteq N(t)$, so $\forall t$, $\mathbb{P}\left(\tilde{X}(t) = X(t)\right)$, and $\tilde{X}$ is a version of X (Definition 74). □

Corollary 88 *If the situation is as in the previous theorem, but Ξ is not compact, there exists a separable version of X in some compactification $\tilde{\Xi}$ of Ξ.*

Proof: Because Ξ is a sub-space of any of its compactifications $\tilde{\Xi}$, X is also a process with values in $\tilde{\Xi}$.[3] Since $\tilde{\Xi}$ is compact, X has a separable modification $\tilde{X}$ with values in $\tilde{\Xi}$, but (with probability 1) $\tilde{X}(t) \in \Xi$. □

Corollary 89 *Let Ξ be a compact metric space, T a separable index set, and μ_J, $J \in \mathrm{Fin}(T)$ a projective family of probability distributions. Then there is a separable stochastic process with finite-dimensional distributions given by μ_J.*

Proof: Combine Theorem 87 with the Kolmogorov Extension Theorem 29. □

8.2 Measurable Versions

It would be nice for us if $X(t)$ is a measurable function *of* t, because we are going to want to write down things like

$$\int_{t=a}^{t=b} X(t)dt$$

and have them mean something. Irritatingly, this will require another modification.

Definition 90 (Measurable sample paths) *Let $T, \mathcal{T}, \tau$ be a measurable space, its σ-field and a measure defined thereon. A random function X on T with values in $\Xi, \mathcal{X}$ has* measurable sample paths *or* is measurable *if $X : T \times \Omega \mapsto \Xi$ is $\widetilde{\mathcal{T} \times \mathcal{F}}/\mathcal{X}$ measurable, where $\mathcal{T} \times \mathcal{F}$ is the product σ-field on $T \times \Omega$, and $\widetilde{\mathcal{T} \times \mathcal{F}}$ its completion by the null sets of the product measure $\tau \times \mathbb{P}$.*

It would seem more natural to simply define measurable sample paths by saying that $X(\cdot,\omega)$ is a $\mathcal{T}$-measurable function of t for $\mathbb{P}$-almost-all ω. However, Definition 90 implies this version, via Fubini's Theorem, and facilitates the proofs of the two following theorems.

Theorem 91 *If $X(t)$ is measurable, and $\mathbf{E}[X(t)]$ is integrable (with respect to the measure τ on T), then for any set $I \in \mathcal{T}$,*

$$\int_I \mathbf{E}[X(t)]\,\tau(dt) \quad = \quad \mathbf{E}\left[\int_I X(t)\tau(dt)\right] \tag{8.23}$$

[3] If you want to be really picky, define a 1-1 function $h : \Xi \mapsto \tilde{\Xi}$ taking points to their counterparts. Then X and $h^{-1}(X)$ are indistinguishable. Do I need to go on?

PROOF: This is just Fubini's Theorem! □

Theorem 92 (Measurable Separable Modifications) *Suppose that T and Ξ are both compact. If $X(t,\omega)$ is continuous in probability at τ-almost-all t, then it has a version which is both separable and measurable, its* measurable separable modification.

PROOF: See Gikhman and Skorokhod (1965/1969, ch. IV, sec. 3, thm. 1, p. 157). □

8.3 Cadlag Versions

Theorem 93 *Let X be a separable random process with $T = [a,b] \subseteq \mathbb{R}$, and Ξ a complete metric space with metric ρ. Suppose that $X(t)$ is continuous in probability on T, and there are real constants $p, q, C \geq 0$, $r > 1$ such that, for any three indices $t_1 < t_2 < t_3 \in T$,*

$$\mathbf{E}\left[\rho^p(X(t_1), X(t_2))\rho^q(X(t_2), X(t_3))\right] \leq C|t_3 - t_1|^r \tag{8.24}$$

The there is a version of X whose sample paths are cadlag (a.s.).

PROOF: Combine Theorem 1 and Theorem 3 of Gikhman and Skorokhod (1965/1969, ch. IV, sec. 4, pp. 159–169). □

8.4 Continuous Modifications

Theorem 94 *Let X be a separable stochastic process with $T = [a,b] \subseteq \mathbb{R}$, and Ξ a complete metric space with metric ρ. Suppose that there are constants $C, p > 0$, $r > 1$ such that, for any $t_1 < t_2 \in T$,*

$$\mathbf{E}\left[\rho^p(X(t_1), X(t_2))\right] \leq C|t_2 - t_1|^r \tag{8.25}$$

Then $X(t)$ has a continuous version.

PROOF: See Gikhman and Skorokhod (1965/1969, ch. IV, sec. 5, thm. 2, p. 170), and the first remark following the theorem. □

A slightly more refined result requires two preliminary definitions.

Definition 95 (Modulus of continuity) *For any function x from a metric space T, d to a metric space Ξ, ρ, the* modulus of continuity *is the function $m_x(r) : \mathbb{R}^+ \mapsto \mathbb{R}^+$ given by*

$$m_x(r) = \sup\{\rho(x(s), x(t)) : s, t \in T, d(s,t) \leq r\} \tag{8.26}$$

Lemma 96 *x is uniformly continuous if and only if its modulus of continuity $\to 0$ as $r \to 0$.*

Proof: Obvious from Definition 95 and the definition of uniform continuity.

Definition 97 (Hölder-continuous) *Continuing the notation of Definition 95, we say that x is* Hölder-continuous with exponent c *if there are positive constants c, γ such that $m_x(r) \leq \gamma r^c$ for all sufficiently small r; i.e., $m_x(r) = O(r^c)$. If this holds on every bounded subset of T, then the function is* locally Hölder-continuous.

Theorem 98 *Let T be $\mathbb{R}^d$ and Ξ a complete metric space with metric ρ. If there are constants $p, q, \gamma > 0$, such that, for any $t_1, t_2 \in T$,*

$$\mathbf{E}\left[\rho^p(X(t_1), X(t_2))\right] \leq \gamma |t_1 - t_2|^{d+q} \tag{8.27}$$

then X has a continuous version $\tilde{X}$, and almost all sample paths of $\tilde{X}$ are locally Hölder-continuous for any exponent between 0 and q/p exclusive.

Proof: See Kallenberg, theorem 3.23 (pp. 57–58). Note that part of Kallenberg's proof is a restricted case of what we've already done in prove the existence of a separable version! □

This lecture, the last, and even a lot of the one before have all been pretty hard and abstract. As a reward for our labor, however, we now have a collection of very important tools — operator representations, filtrations and optional times, recurrence times, and finally existence theorems for continuous processes. These are the devices which will let us take the familiar theory of elementary Markov chains, with finitely many states in discrete time, and produce the general theory of Markov processes with continuous states and/or continuous time. The next lecture will begin this work, starting with the operators.

Chapter 9

Markov Processes

This lecture begins our study of Markov processes.

Section 9.1 is mainly "ideological": it formally defines the Markov property for one-parameter processes, and explains why it is a natural generalization of both complete determinism and complete statistical independence.

Section 9.2 introduces the description of Markov processes in terms of their transition probabilities and proves the existence of such processes.

9.1 The Correct Line on the Markov Property

The Markov property is the independence of the future from the past, given the present. Let us be more formal.

Definition 99 (Markov Property) *A one-parameter process X is a Markov process with respect to a filtration $\mathcal{F}$ when X_t is adapted to the filtration, and, for any $s > t$, X_s is independent of $\mathcal{F}_t$ given X_t, $X_s \perp\!\!\!\perp \mathcal{F}_t | X_t$. If no filtration is mentioned, it may be assumed to be the natural one generated by X. If X is also conditionally stationary, then it is a* time-homogeneous *(or just* homogeneous*) Markov process.*

Lemma 100 *Let X_t^+ stand for the collection of X_u, $u > t$. If X is Markov, then $X_t^+ \perp\!\!\!\perp \mathcal{F}_t | X_t$.*

Proof: See Exercise 9.1. □

There are two routes to the Markov property. One is the path followed by Markov himself, of desiring to weaken the assumption of strict statistical independence between variables to mere conditional independence. In fact, Markov specifically wanted to show that independence was *not* a necessary condition for the law of large numbers to hold, because his arch-enemy claimed that it was, and used that as grounds for believing in free will and Christianity.[1] It turns out that all the key limit theorems of probability — the weak and strong laws of large numbers, the central limit theorem, etc. — work perfectly well for Markov processes, as well as for IID variables.

The other route to the Markov property begins with completely deterministic systems in physics and dynamics. The *state* of a deterministic dynamical system is some variable which fixes the value of all present and future observables. As a consequence, the present state determines the state at all future times. However, strictly deterministic systems are rather thin on the ground, so a natural generalization is to say that the present state determines the *distribution* of future states. This is precisely the Markov property.

Remarkably enough, it is possible to represent any one-parameter stochastic process X as a noisy function of a Markov process Z. The shift operators give a trivial way of doing this, where the Z process is not just homogeneous but actually fully deterministic. An equally trivial, but slightly more probabilistic, approach is to set $Z_t = X_t^-$, the complete past up to and including time t. (This is not necessarily homogeneous.) It turns out that, subject to mild topological conditions on the space X lives in, there is a unique *non-trivial* representation where $Z_t = \epsilon(X_t^-)$ for some function ϵ, Z_t is a homogeneous Markov process, and $X_u \perp\!\!\!\perp \sigma(\{X_t, t \leq u\})|Z_t$. (See Knight (1975, 1992).) We may explore such *predictive Markovian representations* at the end of the course, if time permits.

9.2 Transition Probability Kernels

The most obvious way to specify a Markov process is to say what its transition probabilities are. That is, we want to know $\mathbb{P}(X_s \in B|X_t = x)$ for every $s > t$, $x \in \Xi$, and $B \in \mathcal{X}$. Probability kernels (Definition 30) were invented to let us do just this.

Definition 101 (Product of Probability Kernels) *Let μ and ν be two probability kernels from Ξ to Ξ. Then their product $\mu\nu$ is a kernel from Ξ to Ξ, defined by*

$$(\mu\nu)(x, B) \equiv \int \mu(x, dy)\nu(y, B) \tag{9.1}$$

$$= (\mu \otimes \nu)(x, \Xi \times B) \tag{9.2}$$

Intuitively, all the product does is say that the probability of starting at the point x and landing in the set B is equal the probability of first going to y and then ending in B, integrated over all intermediate points y. (Strictly speaking, there is an abuse of notation in Eq. 9.2, since the second kernel in a composition $\otimes$ should be defined over a product space, here $\Xi \times \Xi$. So suppose we have such a

[1] I am not making this up. See Basharin *et al.* (2004) for a nice discussion of the origin of Markov chains and of Markov's original, highly elegant, work on them. There is a translation of Markov's original paper in an appendix to Howard (1971), and I dare say other places as well.

kernel ν', only $\nu'((x, y), B) = \nu(y, B)$.) Finally, observe that if $\mu(x, \cdot) = \delta_x$, the delta function at x, then $(\mu\nu)(x, B) = \nu(x, B)$, and similarly that $(\nu\mu)(x, B) = \nu(x, B)$.

Definition 102 *For every $(t, s) \in T \times T$, $s \geq t$, let $\mu_{t,s}$ be a probability kernel from Ξ to Ξ. These probability kernels form a* transition semi-group *when*

1. *For all t, $\mu_{t,t}(x, \cdot) = \delta_x$.*
2. *For any $t \leq s \leq u \in T$, $\mu_{t,u} = \mu_{t,s}\mu_{s,u}$.*

A transition semi-group for which $\forall t \leq s \in T$, $\mu_{t,s} = \mu_{0,s-t} \equiv \mu_{s-t}$ is homogeneous*.*

As with the shift semi-group, this is really a monoid (because $\mu_{t,t}$ acts as the identity).

The major theorem is the existence of Markov processes with specified transition kernels.

Theorem 103 *Let $\mu_{t,s}$ be a transition semi-group and ν_t a collection of distributions on a Borel space Ξ. If*

$$\nu_s = \nu_t\mu_{t,s} \tag{9.3}$$

then there exists a Markov process X such that

$$\forall t,\ \mathcal{L}(X_t) = \nu_t \tag{9.4}$$

$$\forall t_1 \leq t_2 \leq \ldots \leq t_n,\ \mathcal{L}(X_{t_1}, X_{t_2} \ldots X_{t_n}) = \nu_{t_1} \otimes \mu_{t_1,t_2} \otimes \ldots \otimes \mu_{t_{n-1},t_n} \tag{9.5}$$

Conversely, if X is a Markov process with values in Ξ, then there exist distributions ν_t and a transition kernel semi-group $\mu_{t,s}$ such that Equations 9.4 and 9.3 hold, and

$$\mathbb{P}(X_s \in B|\mathcal{F}_t) = \mu_{t,s} \ a.s. \tag{9.6}$$

Proof: (*From transition kernels to a Markov process.*) For any finite set of times $J = \{t_1, \ldots t_n\}$ (in ascending order), define a distribution on Ξ_J as

$$\nu_J \equiv \nu_{t_1} \otimes \mu_{t_1,t_2} \otimes \ldots \otimes \mu_{t_{n-1},t_n} \tag{9.7}$$

It is easily checked, using point (2) in the definition of a transition kernel semi-group (Definition 102), that the ν_J form a projective family of distributions. Thus, by the Kolmogorov Extension Theorem (Theorem 29), there exists a stochastic process whose finite-dimensional distributions are the ν_J. Now pick a J of size n, and two sets, $B \in \mathcal{X}^{n-1}$ and $C \in \mathcal{X}$.

$$\begin{aligned} \mathbb{P}(X_J \in B \times C) &= \nu_J(B \times C) && (9.8)\\ &= \mathbf{E}[\mathbf{1}_{B\times C}(X_J)] && (9.9)\\ &= \mathbf{E}\left[\mathbf{1}_B(X_{J\setminus t_n})\mu_{t_{n-1},t_n}(X_{t_{n-1}}, C)\right] && (9.10)\end{aligned}$$

Set $\mathcal{F}_t$ to be the natural filtration, $\sigma(\{X_u, u \leq s\})$. If $A \in \mathcal{F}_s$ for some $s \leq t$, then by the usual generating class arguments we have

$$\mathbb{P}\left(X_t \in C, X_s^- \in A\right) = \mathbf{E}\left[\mathbf{1}_A \mu_{s,t}(X_s, C)\right] \tag{9.11}$$

$$\mathbb{P}\left(X_t \in C | \mathcal{F}_s\right) = \mu_{s,t}(X_s, C) \tag{9.12}$$

i.e., $X_t \perp\!\!\!\perp \mathcal{F}_s | X_s$, as was to be shown.

(*From the Markov property to the transition kernels.*) From the Markov property, for any measurable set $C \in \mathcal{X}$, $\mathbb{P}\left(X_t \in C|\mathcal{F}_s\right)$ is a function of X_s alone. So define the kernel $\mu_{s,t}$ by $\mu_{s,t}(x, C) = \mathbb{P}\left(X_t \in C|X_s = x\right)$, with a possible measure-0 exceptional set from (ultimately) the Radon-Nikodym theorem. (The fact that Ξ is Borel guarantees the existence of a regular version of this conditional probability.) We get the semi-group property for these kernels thus: pick any three times $t \leq s \leq u$, and a measurable set $C \subseteq \Xi$. Then

$$\mu_{t,u}(X_t, C) = \mathbb{P}\left(X_u \in C|\mathcal{F}_t\right) \tag{9.13}$$

$$= \mathbb{P}\left(X_u \in C, X_s \in \Xi|\mathcal{F}_t\right) \tag{9.14}$$

$$= (\mu_{t,s} \otimes \mu_{s,u})(X_t, \Xi \times C) \tag{9.15}$$

$$= (\mu_{t,s}\mu_{s,u})(X_t, C) \tag{9.16}$$

The argument to get Eq. 9.3 is similar. □

Note: For one-sided discrete-parameter processes, we could use the Ionescu-Tulcea Extension Theorem 33 to go from a transition kernel semi-group to a Markov process, even if Ξ is not a Borel space.

Definition 104 *Let X be a homogeneous Markov process with transition kernels μ_t. A distribution ν on Ξ is* invariant *when, $\forall t$, $\nu = \nu\mu_t$, i.e.,*

$$(\nu\mu_t)(B) \equiv \int \nu(dx)\mu_t(x, B) \tag{9.17}$$

$$= \nu(B) \tag{9.18}$$

ν is also called an equilibrium distribution*.*

The term “equilibrium” comes from statistical physics, where however its meaning is a bit more strict, in that “detailed balance” must also be satisified: for any two sets $A, B \in \mathcal{X}$,

$$\int \nu(dx)\mathbf{1}_A\mu_t(x, B) = \int \nu(dx)\mathbf{1}_B\mu_t(x, A) \tag{9.19}$$

i.e., the flow of probability from A to B must equal the flow in the opposite direction. Much confusion has resulted from neglecting the distinction between equilibrium in the strict sense of detailed balance and equilibrium in the weaker sense of invariance.

Theorem 105 *Suppose X is homogeneous, and $\mathcal{L}(X_t) = \nu$, where ν is an invariant distribution. Then the process X_t^+ is stationary.*

Proof: Exercise 9.4. □

9.3 Exercises

Exercise 9.1 *Prove Lemma 100.*

Exercise 9.2 *Show that if X is a Markov process, then, for any $t \in T$, X_t^+ is a one-sided Markov process.*

Exercise 9.3 *Let X be a continuous-parameter Markov process, and t_n a countable set of strictly increasing indices. Set $Y_n = X_{t_n}$. Is Y_n a Markov process? If X is homogeneous, is Y also homogeneous? Does either answer change if $t_n = nt$ for some constant interval $t > 0$?*

Exercise 9.4 *Prove Theorem 105.*

Chapter 10

Alternate Characterizations of Markov Processes

This lecture introduces two ways of characterizing Markov processes other than through their transition probabilities.

Section 10.1 addresses a question raised in the last class, about when being Markovian relative to one filtration implies being Markov relative to another.

Section 10.2 describes discrete-parameter Markov processes as transformations of sequences of IID uniform variables.

Section 10.3 describes Markov processes in terms of measure-preserving transformations (Markov operators), and shows this is equivalent to the transition-probability view.

10.1 The Markov Property Under Multiple Filtrations

In the last lecture, we defined what it is for a process to be Markovian relative to a given filtration $\mathcal{F}_t$. The question came up in class of when knowing that X Markov with respect to one filtration $\mathcal{F}_t$ will allow us to deduce that it is Markov with respect to another, say $\mathcal{G}_t$.

To begin with, let's introduce a little notation.

Definition 106 (Natural Filtration) *The natural filtration for a stochastic process X is $\mathcal{F}_t^X \equiv \sigma(\{X_u,\ u \leq t\})$. Obviously, every process X is adapted to $\mathcal{F}_t^X$.*

Definition 107 (Comparison of Filtrations) *A filtration $\mathcal{G}_t$ is* finer than *or* more refined than *or* a refinement of *$\mathcal{F}_t$, $\mathcal{F}_t \prec \mathcal{G}_t$, if, for all t, $\mathcal{F}_t \subseteq G_t$, and at least sometimes the inequality is strict. $\mathcal{F}_t$ is* coarser *or* less fine *than $\mathcal{G}_t$. If $\mathcal{F}_t \prec \mathcal{G}_t$ or $\mathcal{F}_t = \mathcal{G}_t$, we write $\mathcal{F}_t \preceq \mathcal{G}_t$.*

Lemma 108 *If X is adapted to $\mathcal{G}_t$, then $\mathcal{F}_t^X \preceq \mathcal{G}_t$.*

PROOF: For each t, X_t is $\mathcal{G}_t$ measurable. But $\mathcal{F}_t^X$ is, by construction, the smallest σ-algebra with respect to which X_t is measurable, so, for every t, $\mathcal{F}_t^X \subseteq \mathcal{G}_t$, and the result follows. □

Theorem 109 *If X is Markovian with respect to $\mathcal{G}_t$, then it is Markovian with respect to any coarser filtration to which it is adapted, and in particular with respect to its natural filtration.*

PROOF: Use the smoothing property of conditional expectations: For any two σ-fields $\mathcal{F} \subset \mathcal{G}$ and random variable Y, $\mathbf{E}\left[Y|\mathcal{F}\right] = \mathbf{E}\left[\mathbf{E}\left[Y|\mathcal{G}\right]|\mathcal{F}\right]$ a.s. So, if $\mathcal{F}_t$ is coarser than $\mathcal{G}_t$, and X is Markovian with respect to the latter, for any function $f \in L_1$ and time $s > t$,

$$\begin{aligned}
\mathbf{E}\left[f(X_s)|\mathcal{F}_t\right] &= \mathbf{E}\left[\mathbf{E}\left[f(X_s)|\mathcal{G}_t\right]|\mathcal{F}_t\right] \ a.s. && (10.1)\\
&= \mathbf{E}\left[\mathbf{E}\left[f(X_s)|X_t\right]|\mathcal{F}_t\right] && (10.2)\\
&= \mathbf{E}\left[f(X_s)|X_t\right] && (10.3)
\end{aligned}$$

where the last line uses the facts that (i) $\mathbf{E}\left[f(X_s)|X_t\right]$ is a function X_t, (ii) X is adapted to $\mathcal{F}_t$, so X_t is $\mathcal{F}_t$-measurable, and (iii) if Y is $\mathcal{F}$-measurable, then $\mathbf{E}\left[Y|\mathcal{F}\right] = Y$. Since this holds for all $f \in L_1$, it holds in particular for $\mathbf{1}_A$, where A is any measurable set, and this established the conditional independence which constitutes the Markov property. Since (Lemma 108) the natural filtration is the coarsest filtration to which X is adapted, the remainder of the theorem follows. □

The converse is false, as the following example shows.

Example 110 *We revert to the symbolic dynamics of the logistic map, Examples 39 and 40. Let S_1 be distributed on the unit interval with density $1/\pi\sqrt{s(1-s)}$, and let $S_n = 4S_{n-1}(1-S_{n-1})$. Finally, let $X_n = \mathbf{1}_{[0.5,1.0]}(S_n)$. It can be shown that the X_n are a Markov process with respect to their natural filtration; in fact, with respect to that filtration, they are independent and identically distributed Bernoulli variables with probability of success $1/2$. However, $\mathbb{P}\left(X_{n+1}|\mathcal{F}_n^S, X_n\right) \neq \mathbb{P}\left(X_{n+1}|X_n\right)$, since X_{n+1} is a* deterministic *function of S_n. But, clearly, $\mathcal{F}_n^S$ is a refinement of $\mathcal{F}_n^X$.*

The issue can be illustrated with graphical models (Spirtes *et al.*, 2001; Pearl, 1988). A discrete-time Markov process looks like Figure 10.1*a*. X_n blocks all the pasts from the past to the future (in the diagram, from left to right), so it produces the desired conditional independence. Now let's add another variable which actually drives the X_n (Figure 10.1*b*). If we can't measure the S_n variables, just the X_n ones, then it can still be the case that we've got the conditional independence among what we can see. But if we can see X_n as well as S_n — which is what refining the filtration amounts to — then simply conditioning on X_n does not block all the paths from the past of X to its future, and, generally speaking, we will lose the Markov property. Note that knowing

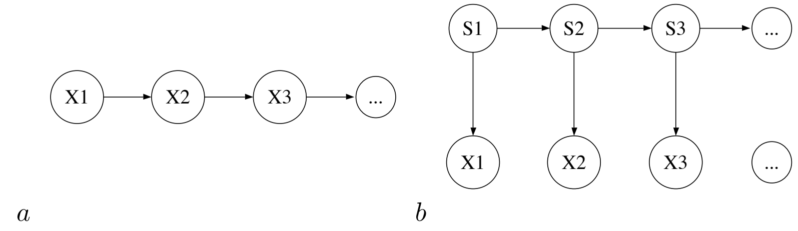

Figure 10.1: (*a*) Graphical model for a Markov chain. (*b*) Refining the filtration, say by conditioning on an additional random variable, can lead to a failure of the Markov property.

S_n *does* block all paths from past to future — so this remains a *hidden* Markov model. Markovian representation theory is about finding conditions under which we can get things to look like Figure 10.1*b*, even if we can't get them to look like Figure 10.1*a*.

10.2 Markov Sequences as Transduced Noise

A key theorem says that discrete-time Markov processes can be viewed as the result of applying a certain kind of filter to pure noise.

Theorem 111 *Let X be a one-sided discrete-parameter process taking values in a Borel space Ξ. X is Markov iff there are measurable functions $f_n : \Xi \times [0,1] \mapsto \Xi$ such that, for IID random variables $Z_n \sim U(0,1)$, all independent of X_1, $X_{n+1} = f_n(X_n, Z_n)$ almost surely. X is homogeneous iff $f_n = f$ for all n.*

Proof: Kallenberg, Proposition 8.6, p. 145. Notice that, in order to get the "only if" direction to work, Kallenberg invokes what we have as Proposition 26, which is where the assumptions that Ξ is a Borel space comes in. You should verify that the "if" direction does not require this assumption. □

Let us stick to the homogeneous case, and consider the function f in somewhat more detail.

In engineering or computer science, a *transducer* is an apparatus — really, a function — which takes a stream of inputs of one kind and produces a stream of outputs of another kind.

Definition 112 (Transducer) *A (deterministic) transducer is a sextuple $\langle \Sigma, \Upsilon, \Xi, f, h, s_0 \rangle$ where Σ, Υ and Ξ are, respectively, the state, input and output spaces, $f : \Sigma \times \Xi \mapsto \Sigma$ is the* state update function *or* state transition function*, $h : \Sigma \times \Upsilon \mapsto \Xi$ is the* measurement *or* observation *function, and $s_0 \in \Sigma$ is the* starting state*. (We shall assume both f and h are always measurable.) If h does not depend on its state argument, the transducer is* memoryless*. If f does not depend on its state argument, the transducer is* without after-effect*.*

It should be clear that if a memoryless transducer is presented with IID inputs, its output will be IID as well. What Theorem 111 says is that, if we have a transducer with memory (so that h depends on the state) but is without after-effect (so that f does not depend on the state), IID inputs will produce Markovian outputs, and conversely any reasonable Markov process can be represented in this way. Notice that if a transducer is without memory, we can replace it with an equivalent with a single state, and if it is without after-effect, we can identify Σ and Ξ.

Notice also that the two functions f and h determine a transition function where we use the input to update the state: $g : \Sigma \times \Upsilon \mapsto \Sigma$, where $g(s, y) = f(s, h(s, y))$. Thus, if the inputs are IID and uniformly distributed, then (Theorem 111) the successive states of the transducer are always Markovian. The question of which processes can be produced by noise-driven transducers is this intimately bound up with the question of Markovian representations. While, as mentioned, quite general stochastic processes *can* be put in this form (Knight, 1975, 1992), it is not necessarily possible to do this with a finite internal state space Σ, even when Ξ is finite. The distinction between finite and infinite Σ is crucial to theoretical computer science, and we might come back to it later, but

Two issues suggest themselves in connection with this material. One is whether, given a *two*-sided process, we can pull the same trick, and represent a Markovian X as a transformation of an IID sequence extending into the infinite past. (Remember that the theorem is for one-sided processes, and starts with an initial X_1.) This is more subtle than it seems at first glance, or even than it seemed to Norbert Wiener when he first posed the question (Wiener, 1958); for a detailed discussion, see Rosenblatt (1971), and, for recent set of applications, Wu (2005). The other question is whether the same trick can be pulled in continuous time; here much less is known.

10.3 Time-Evolution (Markov) Operators

Let's look again at the evolution of the one-dimensional distributions for a Markov process:

$$\nu_s = \nu_t \mu_{t,s} \quad (10.4)$$

$$\nu_s(B) = \int \nu_t(dx) \mu_{t,s}(x, B) \quad (10.5)$$

The transition kernels define linear operators taking distributions on Ξ to distributions on Ξ. This can be abstracted.

Definition 113 (Markov Operator) *Take any probability space $\Xi, \mathcal{X}, \mu$, and let L_1 be as usual the class of all μ-integrable generalized functions on Ξ. A linear operator $P : L_1 \mapsto L_1$ is a* Markov operator *when:*

1. *If $f \geq 0$ (a.e. μ), $Pf \geq 0$ (a.e. μ).*

2. *If $f \leq M$ (a.e. μ), $Pf \leq M$ (a.e. μ).*
3. *$P\mathbf{1}_\Xi = \mathbf{1}_\Xi$.*
4. *If $f_n \downarrow 0$, then $Pf_n \downarrow 0$.*

Lemma 114 *Every probability kernel κ from Ξ to Ξ induces a Markov operator K,*

$$Kf(x) = \int \kappa(x, dy) f(y) \tag{10.6}$$

and conversely every operator defines a transition probability kernel,

$$\kappa(x, B) = K\mathbf{1}_B(x) \tag{10.7}$$

Proof: Exercise 10.1. □

Clearly, if κ is part of a transition kernel semi-group, then the collection of induced Markov operators also forms a semi-group.

Theorem 115 (Markov operator semi-groups and Markov processes) *Let X be a Markov process with transition kernels $\mu_{t,s}$, and let $K_{t,s}$ be the corresponding semi-group of operators. Then for any $f \in L_1$,*

$$\mathbf{E}\left[f(X_s)|\mathcal{F}_t\right] = (K_{t,s}f)(X_t) \tag{10.8}$$

Conversely, let X be any stochastic process, and let $K_{t,s}$ be a semi-group of Markov operators such that Equation 10.8 is valid (a.s.). Then X is a Markov process.

Proof: Exercise 10.2. □

Remark. The proof works because the expectations of all L_1 functions together determine a probability measure. If we knew of another collection of functions which also sufficed to determine a measure, then linear operators on that collection would work just as well, in the theorem, as do the Markov operators, which by definition apply to all of L_1. In particular, it is sometimes possible to define operators only on much smaller, more restricted collections of functions, which can have technical advantages. See Ethier and Kurtz (1986, ch. 4, sec. 1) for details.

The next two lemmas will prove useful in establishing asymptotic results.

Lemma 116 (Markov Operators are Contractions) *For any Markov operator P and $f \in L_1$,*

$$\|Pf\| \leq \|f\| \tag{10.9}$$

Proof (after Lasota and Mackey (1994, prop. 3.1.1, pp. 38–39)): First, notice that $(Pf(x))^+ \leq Pf^+(x)$, because $(Pf(x))^+ = (Pf^+ - Pf^-)^+ = \max(0, Pf^+ - Pf^-) \leq \max(0, Pf^+) = Pf^+$. Similarly $(Pf(x))^- \leq Pf^-(x)$. Therefore $|Pf| \leq P|f|$, and then the statement follows by integration. □

Lemma 117 *For any Markov operator, and any $f, g \in L_1$, $\|P^n f - P^n g\|$ is non-increasing.*

Proof: By linearity, $\|P^n f - P^n g\| = \|P^n(f - g)\|$. By the definition of P^n, $\|P^n(f - g)\| = \|PP^{n-1}(f - g)\|$. By the contraction property (Lemma 116), $\|PP^{n-1}(f - g)\| \leq \|P^{n-1}(f - g)\| = \|P^{n-1}f - P^{n-1}g\|$ (by linearity again). □

Theorem 118 *A probability measure ν is invariant for a homogeneous Markov process iff it is a fixed point of all the transition operators, $\nu K_t = \nu$.*

Proof: Clear from the definitions! □

10.4 Exercises

Exercise 10.1 *Prove Lemma 114. Hint: you will want to use the fact that $\mathbf{1}_B \in L_1$ for all measurable sets B.*

Exercise 10.2 *Prove Theorem 115. Hint: in showing that a collection of operators determines a Markov process, try using mathematical induction on the finite-dimensional distributions.*

Chapter 11

Markov Examples

Section 11.1 finds the transition kernels for the Wiener process, as an example of how to manipulate such things.

Section 11.2 looks at the evolution of densities under the action of the logistic map; this shows how deterministic dynamical systems can be brought under the sway of the theory we've developed for Markov processes.

11.1 Transition Kernels for the Wiener Process

We have previously defined the Wiener process (Examples 38 and 78) as the real-valued process on $\mathbb{R}^+$ with the following properties:

1. $W(0) = 0$;
2. For any three times $t_1 \leq t_2 \leq t_3$, $W(t_3) - W(t_2) \perp\!\!\!\perp W(t_2) - W(t_1)$ (independent increments);
3. For any two times $t_1 \leq t_2$, $W(t_2) - W(t_1) \sim \mathcal{N}(0, t_2 - t_1)$ (Gaussian increments);
4. Continuous sample paths (in the sense of Definition 72).

Here we will use the Gaussian increment property to construct a transition kernel, and then use the independent increment property to show that these keernels satisfy the Chapman-Kolmogorov equation, and hence that there exist *Markov* processes with the desired finite-dimensional distributions.

First, notice that the Gaussian increments property gives us the transition probabilities:

$$\begin{aligned}
\mathbb{P}\left(W(t_2) \in B | W(t_1) = w_1\right) &= \mathbb{P}\left(W(t_2) - W(t_1) \in B - w_1\right) && (11.1)\\
&= \int_{B-w_1} du \frac{1}{\sqrt{2\pi(t_2-t_1)}} e^{-\frac{u^2}{2(t_2-t_1)}} && (11.2)\\
&= \int_{B} dw_2 \frac{1}{\sqrt{2\pi(t_2-t_1)}} e^{-\frac{(w_2-w_1)^2}{2(t_2-t_1)}} && (11.3)\\
&\equiv \mu_{t_1,t_2}(w_1, B) && (11.4)
\end{aligned}$$

To show that $W(t)$ is a Markov process, we must show that, for any three times $t_1 \leq t_2 \leq t_3$, $\mu_{t_1,t_2}\mu_{t_2,t_3} = \mu_{t_1,t_3}$.

Notice that $W(t_3) - W(t_1) = (W(t_3) - W(t_2)) + (W(t_2) - W(t_1))$. Because increments are independent, then, $W(t_3) - W(t_1)$ is the sum of two independent random variables, $W(t_3) - W(t_2)$ and $W(t_2) - W(t_1)$. The distribution of $W(t_3) - W(t_1)$ is then the convolution of distributions of $W(t_3) - W(t_2)$ and $W(t_2) - W(t_1)$. Those are $\mathcal{N}(0, t_3 - t_2)$ and $\mathcal{N}(0, t_2 - t_1)$ respectively. The convolution of two Gaussian distributions is a third Gaussian, summing their parameters, so according to this argument, we must have $W(t_3) - W(t_1) \sim \mathcal{N}(0, t_3 - t_1)$. But this is precisely what we should have, by the Gaussian-increments property. Since the trick we used above to get the transition kernel from the increment distribution can be applied again, we conclude that $\mu_{t_1,t_2}\mu_{t_2,t_3} = \mu_{t_1,t_3}$ and the Chapman-Kolmogorov property is satisfied; therefore (Theorem 103), $W(t)$ is a Markov process (with respect to its natural filtration).

To see that $W(t)$ has, or can be made to have, continuous sample paths, invoke Theorem 94.

11.2 Probability Densities in the Logistic Map

Let's revisit the first part of Exercise 5.3, from the point of view of what we now know about Markov processes. The exercise asks us to show that the density $\frac{1}{\pi\sqrt{x(1-x)}}$ is invariant under the action of the logistic map with $a = 4$.

Let's write the mapping as $F(x) = 4x(1-x)$. Solving a simple quadratic equation gives us the fact that $F^{-1}(x)$ is the set $\left\{\frac{1}{2}\left(1 - \sqrt{1-x}\right), \frac{1}{2}\left(1 + \sqrt{1-x}\right)\right\}$. Notice, for later use, that the two solutions add up to 1. Notice also that $F^{-1}([0,x]) = \left[0, \frac{1}{2}\left(1 - \sqrt{1-x}\right)\right] \cup \left[\frac{1}{2}\left(1 + \sqrt{1-x}\right), 1\right]$. Now we consider $\mathbb{P}(X_{n+1} \leq x)$, the cumulative distribution function of X_{n+1}.

$$\begin{aligned}
&\mathbb{P}\left(X_{n+1} \leq x\right) \\
&= \mathbb{P}\left(X_{n+1} \in [0,x]\right) && (11.5)\\
&= \mathbb{P}\left(X_n \in F^{-1}\left([0,x]\right)\right) && (11.6)\\
&= \mathbb{P}\left(X_n \in \left[0, \frac{1}{2}\left(1-\sqrt{1-x}\right)\right] \cup \left[\frac{1}{2}\left(1+\sqrt{1-x}\right), 1\right]\right) && (11.7)\\
&= \int_0^{\frac{1}{2}\left(1-\sqrt{1-x}\right)} \rho_n\left(y\right) dy + \int_{\frac{1}{2}\left(1+\sqrt{1-x}\right)}^1 \rho_n\left(y\right) dy && (11.8)
\end{aligned}$$

where ρ_n is the density of X_n. So we have an integral equation for the evolution of the density,

$$\int_0^x \rho_{n+1}\left(y\right) dy = \int_0^{\frac{1}{2}\left(1-\sqrt{1-x}\right)} \rho_n\left(y\right) dy + \int_{\frac{1}{2}\left(1+\sqrt{1-x}\right)}^1 \rho_n\left(y\right) dy \quad (11.9)$$

This sort of integral equation is complicated to solve directly. Instead, take the derivative of both sides with respect to x; we can do this through the fundamental theorem of calculus. On the left hand side, this will just give $\rho_{n+1}\left(x\right)$, the density we want.

$$\begin{aligned}
&\rho_{n+1}\left(x\right) && (11.10)\\
&= \frac{d}{dx}\int_0^{\frac{1}{2}\left(1-\sqrt{1-x}\right)} \rho_n\left(y\right) dy + \frac{d}{dx}\int_{\frac{1}{2}\left(1+\sqrt{1-x}\right)}^1 \rho_n\left(y\right) dy \\
&= \rho_n\left(\frac{1}{2}\left(1-\sqrt{1-x}\right)\right)\frac{d}{dx}\left(\frac{1}{2}\left(1-\sqrt{1-x}\right)\right) && (11.11)\\
&\quad -\rho_n\left(\frac{1}{2}\left(1+\sqrt{1-x}\right)\right)\frac{d}{dx}\left(\frac{1}{2}\left(1+\sqrt{1-x}\right)\right) \\
&= \frac{1}{4\sqrt{1-x}}\left(\rho_n\left(\frac{1}{2}\left(1-\sqrt{1-x}\right)\right) + \rho_n\left(\frac{1}{2}\left(1+\sqrt{1-x}\right)\right)\right) && (11.12)
\end{aligned}$$

Notice that this defines a linear operator taking densities to densities. (You should verify the linearity.) In fact, this is a Markov operator, by the terms of Definition 113. Markov operators of this sort, derived from deterministic maps, are called *Perron-Frobenius* or *Frobenius-Perron* operators, and accordingly denoted by P. Thus an invariant density is a ρ^* such that $\rho^* = P\rho^*$. All the problem asks us to do is to verify that $\frac{1}{\pi\sqrt{x(1-x)}}$ is such a solution.

$$\begin{aligned}
& \rho^* \left(\frac{1}{2}\left(1-\sqrt{1-x}\right)\right) && (11.13)\\
= \ & \frac{1}{\pi}\left(\frac{1}{2}\left(1-\sqrt{1-x}\right)\left(1-\left(\frac{1}{2}\left(1-\sqrt{1-x}\right)\right)\right)\right)^{-1/2} \\
= \ & \frac{1}{\pi}\left(\frac{1}{2}\left(1-\sqrt{1-x}\right)\frac{1}{2}\left(1+\sqrt{1-x}\right)\right)^{-1/2} && (11.14)\\
= \ & \frac{2}{\pi\sqrt{x}} && (11.15)
\end{aligned}$$

Since $\rho^*(x) = \rho^*(1-x)$, it follows that

$$\begin{aligned}
P\rho^* &= 2\frac{1}{4\sqrt{1-x}}\rho^*\left(\frac{1}{2}\left(1-\sqrt{1-x}\right)\right) && (11.16)\\
&= \frac{1}{\pi\sqrt{x(1-x)}} && (11.17)\\
&= \rho^* && (11.18)
\end{aligned}$$

as desired.

By Lemma 117, for any distribution ρ, $\|P^n\rho - P^n\rho^*\|$ is a non-increasing function of n. However, $P^n\rho^* = \rho^*$, so the iterates of *any* distribution, under the map, approach the invariant distribution monotonically. It would be very handy if we could show that any initial distribution ρ eventually converged on ρ^*, i.e. that $\|P^n\rho - \rho^*\| \to 0$. When we come to ergodic theory, we will see conditions under which such distributional convergence holds, as it does for the logistic map, and learn how such convergence in distribution is connected to both pathwise convergence properties, and to the decay of correlations.

11.3 Exercises

Exercise 11.1 (Brownian Motion with Constant Drift) *Consider a process $X(0)$ which, like the Wiener process, has $X(0) = 0$ and independent increments, but where $X(t_2) - X(t_1) \sim \mathcal{N}(a(t_2-t_1), \sigma^2(t_2-t_1))$. a is called the* drift rate *and σ^2 the* diffusion constant. *Show that $X(t)$ is a Markov process, following the argument for the standard Wiener process ($a = 0$, $\sigma^2 = 1$) above. Do such processes have continuous modifications for all (finite) choices of a and σ^2? If so, prove it; if not, give at least one counter-example.*

Exercise 11.2 (Perron-Frobenius Operators) *Verify that P defined in the section on the logistic map above is a Markov operator.*

Chapter 12

Generators of Markov Processes

> This lecture is concerned with the infinitessimal generator of a Markov process, and the sense in which we are able to write the evolution operators of a homogeneous Markov process as exponentials of their generator.

Take our favorite continuous-time homogeneous Markov process, and consider its semi-group of time-evolution operators K_t. They obey the relationship $K_{t+s} = K_t K_s$. That is, multiplication of the operators corresponds to addition of their parameters, and vice versa. This is reminiscent of the exponential functions on the reals, where, for any $k \in \mathbb{R}$, $k^{(t+s)} = k^t k^s$. In the discrete-parameter case, in fact, $K_t = (K_1)^t$, where integer powers of operators are defined in the obvious way, through iterated composition, i.e., $K^2 f = K \circ (Kf)$. It would be nice if we could extend this analogy to continuous-parameter Markov processes. One approach which suggests itself is to notice that, for any k, there's another real number g such that $k^t = e^{tg}$, and that e^{tg} has a nice representation involving integer powers of g:

$$e^{tg} = \sum_{i=0}^{\infty} \frac{(tg)^i}{i!}$$

The strategy this suggests is to look for some other operator G such that

$$K_t = e^{tG} \equiv \sum_{i=0}^{\infty} \frac{t^i G^i}{i!}$$

Such an operator G is called the *generator* of the process, and the purpose of this section is to work out the conditions under which this analogy can be carried through.

In the exponential function case, we notice that g can be extracted by taking the derivative at zero: $\frac{d}{dt} e^{tg}\big|_{t=0} = g$. This suggests the following definition.

Definition 119 (Infinitessimal Generator) *Let K_t be a continuous-parameter semi-group of homogeneous Markov operators. Say that a function $f \in L_1$ belongs to* $\mathrm{Dom}(G)$ *if the limit*

$$\lim_{h \downarrow 0} \frac{K_h f - K_0 f}{h} \equiv Gf \tag{12.1}$$

exists in an L_1-norm sense, i.e., there exists some element of L_1, which we shall call Gf, such that

$$\lim_{h \downarrow 0} \left\| \frac{K_h f - K_0 f}{h} - Gf \right\| = 0 \tag{12.2}$$

The operator G defined through Eq. 12.1 is called the infinitessimal generator of the semi-group K_t.

Lemma 120 *G is a linear operator.*

Proof: Exercise 12.1. □

Lemma 121 *If μ is an invariant distribution of the semi-group K_t, then, $\forall f \in$* $\mathrm{Dom}(G)$*, $\mu G f = 0$.*

Proof: Since μ is invariant, $\mu K_t = \mu$ for all t, hence $\mu K_h f = \mu f$ for all $h \geq 0$ and all f. Since taking expectations with respect to a measure is a linear operator, $\mu(K_h f - f) = 0$, and obviously then $\mu G f = 0$. □

Remark: The converse statement, that if $\mu G f = 0$ for all f, then μ is an invariant measure, requires extra conditions.

You will usually see the definition of the generator written with f instead of $K_0 f$, but I chose this way of doing it to emphasize that G is, basically, the derivative at zero, that $G = dK/dt|_{t=0}$. Recall, from calculus, that the exponential function can k^t be defined by the fact that $\frac{d}{dt}k^t \propto k^t$ (and e can be defined as the k such that the constant of proportionality is 1). As part of our program, we will want to extend this differential point of view. The next lemma builds towards it, by showing that if $f \in \mathrm{Dom}(G)$, then $K_t f$ is too.

Lemma 122 *If G is the generator of the semi-group K_t, and f is in the domain of G, then K_t and G commute, for all t:*

$$\begin{aligned} K_t G f &= \lim_{t' \to t} \frac{K_{t'} f - K_t f}{t' - t} && (12.3) \\ &= G K_t f && (12.4) \end{aligned}$$

Proof: Exercise 12.2. □

Definition 123 (Time Derivative in Function Space) *For every $t \in T$, let $u(t,x)$ be a function in L_1. When the limit*

$$u'(t_0,x) = \lim_{t \rightarrow t_0} \frac{u(t,x) - u(t_0,x)}{t - t_0} \tag{12.5}$$

exists in the L_1 sense, then we say that $u'(t_0)$ is the time derivative *or* strong derivative *of $u(t)$ at t_0.*

Lemma 124 *Let K_t be a homogeneous semi-group of Markov operators with generator G. Let $u(t) = K_t f$ for some $f \in \mathrm{Dom}(G)$. Then $u(t)$ is differentiable at $t = 0$, and its derivative there is Gf.*

Proof: Obvious from the definitions. □

Theorem 125 *Let K_t be a homogeneous semi-group of Markov operators with generator G, and let $u(t,x) = (K_t f)(x)$, for fixed $f \in \mathrm{Dom}(G)$. Then $u'(t)$ exists for all t, and is equal to $Gu(t)$.*

Proof: Since $f \in \mathrm{Dom}(G)$, $K_t G f$ exists, but then, by Lemma 122, $K_t G f = G K_t f = Gu(t)$, so $u(t) \in \mathrm{Dom}(G)$ for all t. Now let's consider the time derivative of $u(t)$ at some arbitrary t_0, working from above:

$$\frac{(u(t) - u(t_0)}{t - t_0} = \frac{K_{t-t_0} u(t_0) - u(t_0)}{t - t_0} \tag{12.6}$$

$$= \frac{K_h u_(t_0) - u(t_0)}{h} \tag{12.7}$$

Taking the limit as $h \downarrow 0$, we get that $u'(t_0) = Gu(t_0)$, which exists, because $u(t_0) \in \mathrm{Dom}(G)$. □

Corollary 126 (Initial Value Problems in Function Space) $u(t) = K_t f$, $f \in \mathrm{Dom}(G)$, *solves the initial value problem $u(0) = f$, $u'(t) = Gu(t)$.*

Proof: Immediate from the theorem. □

Remark: Such initial value problems are sometimes called *Cauchy problems*, especially when G takes the form of a differential operator.

We are now almost ready to state the sense in which K_t is the result of exponentiating G. This is given by the remarkable Hille-Yosida theorem, which in turn involves a family of operators related to the time-evolution operators, the "resolvents", again built by analogy to the exponential functions. Notice that, for any positive constant λ,

$$\int_{t=0}^{\infty} e^{-\lambda t} e^{tg} dt = \frac{1}{\lambda - g} \tag{12.8}$$

from which we could recover g. The left-hand side is just the Laplace transform of e^{tg}.

Definition 127 (Continuous Functions Vanishing at Infinity) *Let Ξ be a locally compact and separable metric space. The class of functions C_0 will consist of functions $f : \Xi \mapsto R$ which are continuous and for which $x \rightarrow \infty$ implies $f(x) \rightarrow 0$.*

Definition 128 (Resolvents) *Given a continuous-parameter time-homogeneous Markov semi-group K_t, for each $\lambda > 0$, the* resolvent operator *or* resolvent *R_λ is the "Laplace transform" of K_t: for every $f \in C_0$,*

$$(R_\lambda f)(x) \equiv \int_{t=0}^{\infty} e^{-\lambda t} (K_t f)(x) dt \tag{12.9}$$

Remark 1: The name "resolvent", like some of the other ideas an terminology of Markov operators, comes from the theory of integral equations; invariant densities (when they exist) are solutions of homogeneous linear Fredholm integral equations of the second kind. Rather than pursue this analogy, or even explain what that means, I will refer you to the classic treatment of integral equations by Courant and Hilbert (1953, ch. 3), which everyone else seems to follow *very closely.*

Remark 2: When the function f is a value (loss, benefit, utility, ...) function, $(K_t f)(x)$ is the expected value at time t when starting the process in state x. $(R_\lambda f)(x)$ can be thought of as the *net present expected value* when starting at x and applying a discount rate λ.

Definition 129 (Yosida Approximation of Operators) *The* Yosida approximation *to a semi-group K_t with generator G is given by*

$$K_t^\lambda \equiv e^{tG^\lambda} \tag{12.10}$$

$$G^\lambda \equiv \lambda G R_\lambda = \lambda(\lambda R_\lambda - I) \tag{12.11}$$

The domain of G^λ contains all C_0 functions, not just those in Dom(G).

Theorem 130 (Hille-Yosida Theorem) *Let G be a linear operator on some linear subspace $\mathcal{D}$ of L_1. G is the generator of a continuous semi-group of contractions K_t if and only if*

1. *$\mathcal{D}$ is dense in L_1;*
2. *For every $f \in L_1$ and $\lambda > 0$, there exists a unique $g \in \mathcal{D}$ such that $\lambda g - Gg = f$;*
3. *For every $g \in \mathcal{D}$ and positive λ, $\|\lambda g - Gg\| \geq \lambda \|g\|$.*

Under these conditions, the resolvents of K_t are given by $R_\lambda = (\lambda - G)^{-1}$, and K_t is the limit of the Yosida approximations as $\lambda \rightarrow \infty$:

$$K_t f = \lim_{\lambda \rightarrow \infty} K_t^\lambda f, \ \forall f \in L_1 \tag{12.12}$$

PROOF: See Kallenberg, Theorem 19.11. □

12.1 Exercises

Exercise 12.1 *Prove Lemma 120.*

Exercise 12.2 *Prove Lemma 122.*

a *Prove Equation 12.3, restricted to $t' \downarrow t$ instead of $t' \to t$. Hint: Write T_t in terms of an integral over the corresponding transition kernel, and find a reason to exchange integration and limits.*

b *Show that the limit as $t' \uparrow t$ also exists, and is equal to the limit from above. Hint: Re-write the quotient inside the limit so it only involves positive time-differences.*

c *Prove Equation 12.4.*

Chapter 13

The Strong Markov Property and Martingale Problems

Section 13.1 introduces the strong Markov property — independence of the past and future conditional on the state at *random* (optional) times.

Section 13.2 describes "the martingale problem for Markov processes", explains why it would be nice to solve the martingale problem, and how solutions are strong Markov processes.

13.1 The Strong Markov Property

A process is Markovian, with respect to a filtration $\mathcal{F}$, if for any *fixed* time t, the future of the process is independent of $\mathcal{F}_t$ given X_t. This is not necessarily the case for a *random* time τ, because there could be subtle linkages between the random time and the evolution of the process. If these can be ruled out, we have a *strong* Markov process.

Definition 131 (Strongly Markovian at a Random Time) *Let X be a Markov process with respect to a filtration $\mathcal{F}$, with transition kernels $\mu_{t,s}$ and evolution operators $K_{t,s}$. Let τ be an $\mathcal{F}$-optional time which is almost surely finite. Then X is* strongly Markovian at τ *when either of the two following (equivalent) conditions hold*

$$\begin{aligned}
\mathbb{P}\left(X_{t+\tau} \in B|\mathcal{F}_\tau\right) &= \mu_{\tau,\tau+t}(X_\tau, B) && (13.1)\\
\mathbf{E}\left[f(X_{\tau+t})|\mathcal{F}_\tau\right] &= (K_{\tau,\tau+t}f)(X_\tau) && (13.2)
\end{aligned}$$

for all $t \geq 0$, $B \in \mathcal{X}$ and bounded measurable functions f.

Definition 132 (Strong Markov Property) *If X is Markovian with respect to $\mathcal{F}$, and strongly Markovian at every $\mathcal{F}$-optional time which is almost surely finite, then it is a* strong Markov process *with respect to $\mathcal{F}$.*

If the index set T is discrete, then the strong Markov property is implied by the ordinary Markov property (Definition 99). If time is continuous, this is not necessarily the case. It is generally true that, if X is Markov and τ takes on only *countably* many values, X is strongly Markov at τ (Exercise 13.1). We would like to find conditions under which a process is strongly Markovian for all optional times, however.

13.2 Martingale Problems

One approach to getting strong Markov processes is through martingales, and more specifically through what is known as the martingale problem.

Notice the following consequence of Theorem 125:

$$K_t f(x) - f(x) = \int_0^t K_s G f(x) ds \tag{13.3}$$

for any $t \geq 0$ and $f \in \text{Dom}(G)$. The relationship between $K_t f$ and the conditional expectation of f suggests the following definition.

Definition 133 (Martingale Problem) *Let Ξ be a Polish space, $\mathcal{D}$ a class of bounded, continuous, real-valued functions on Ξ, and G an operator from $\mathcal{D}$ to bounded, measurable functions on Ξ. A Ξ-valued stochastic process on $\mathbb{R}^+$ is a* solution to the martingale problem for G and $\mathcal{D}$ *if, for all $f \in \mathcal{D}$,*

$$f(X_t) - \int_0^t G f(X_s) ds \tag{13.4}$$

is a martingale with respect to $\mathcal{F}^X$, the natural filtration of X.

Proposition 134 *Suppose X is a cadlag solution to the martingale problem for $G, \mathcal{D}$. Then for any $f \in \mathcal{D}$, the stochastic process given by Eq. 13.4 is also cadlag.*

Proof: Follows from the assumption that f is continuous. □

Lemma 135 *X is a solution to the martingale problem for $G, \mathcal{D}$ if and only if, for all $t, s \geq 0$,*

$$\mathbf{E}\left[f(X_{t+s})|\mathcal{F}_t^X\right] - \mathbf{E}\left[\int_t^{t+s} G f(X_u) du|\mathcal{F}_t^X\right] = f(X_t) \tag{13.5}$$

PROOF: Take the definition of a martingale and re-arrange the terms in Eq. 13.4. □

Martingale problems are important because of the two following theorems (which can both be refined considerably).

Theorem 136 (Markov Processes Solve Martingale Problems) *Let X be a homogeneous Markov process with generator G and cadlag sample paths, and let $\mathcal{D}$ be the continuous functions in* $\mathrm{Dom}(G)$. *Then X solves the martingale problem for $G, \mathcal{D}$.*

PROOF: Exercise 13.2. □

Theorem 137 (Solutions to the Martingale Problem are Strongly Markovian) *Suppose that for each $x \in \Xi$, there is a unique cadlag solution to the martingale problem for $G, \mathcal{D}$ such that $X_0 = x$. Then the collection of these solutions is a homogeneous strong Markov family X, and the generator is equal to G on $\mathcal{D}$.*

PROOF: Exercise 13.3. □

The main use of Theorem 136 is that it lets us prove convergence of some functions of Markov processes, by showing that they can be cast into the form of Eq. 13.4, and then applying the martingale convergence devices. The other use is in conjunction with Theorem 137. We will often want to show that a *sequence* of Markov processes converges on a limit which is, itself, a Markov process. One approach is to show that the terms in the sequence solve martingale problems (via Theorem 136), argue that then the limiting process does too, and finally invoke Theorem 137 to argue that the limiting process must itself be strongly Markovian. This is often *much* easier than showing directly that the limiting process is strongly Markovian. Theorem 137 itself is often a convenient way of showing that the strong Markov property holds.

13.3 Exercises

Exercise 13.1 (Strongly Markov at Discrete Times) *Let X be a homogeneous Markov process with respect to a filtration $\mathcal{F}$ and τ be an $\mathcal{F}$-optional time. Prove that if $\mathbb{P}(\tau < \infty) = 1$, and τ takes on only countably many values, then X is strongly Markovian at τ. (*Note: *the requirement that X be homogeneous can be lifted, but requires some more technical machinery I want to avoid.)*

Exercise 13.2 (Markovian Solutions of the Martingale Problem) *Prove Theorem 136. Hints: Use Lemma 135, bounded convergence, and Theorem 125.*

Exercise 13.3 (Martingale Solutions are Strongly Markovian) *Prove Theorem 137. Hint: use the Optional Sampling Theorem (from 36-752, or from chapter 7 of Kallenberg).*

Chapter 14

Feller Processes

Section 14.1 fulfills the demand, made last time, for an example of a Markov process which is not strongly Markovian.

Section 14.2 makes explicit the idea that the transition kernels of a Markov process induce a kernel over sample paths, mostly to fix notation for later use.

Section 14.3 defines Feller processes, which link the cadlag and strong Markov properties.

14.1 An Example of a Markov Process Which Is Not Strongly Markovian

This is taken from Fristedt and Gray (1997, pp. 626–627).

Example 138 *We will construct an $\mathbb{R}^2$-valued Markov process on $[0, \infty)$ which is not strongly Markovian. Begin by defining the following map from $\mathbb{R}$ to $\mathbb{R}^2$:*

$$f(w) = \begin{cases} (w, 0) & w \leq 0 \\ (\sin w, 1 - \cos w) & 0 < w < 2\pi \\ (w - 2\pi, 0) & w \geq 2\pi \end{cases} \tag{14.1}$$

When w is less than zero or above 2π, $f(w)$ moves along the x axis of the plane; in between, it moves along a circle of radius 1, centered at $(0, 1)$, which it enters and leaves at the origin. Notice that f is invertible everywhere except at the origin, which is ambiguous between $w = 0$ and $w = 2\pi$.

Let $X(t) = f(W(t) + \pi)$, where $W(t)$ is a standard Wiener process. At all t, $\mathbb{P}(W(t) + \pi = 0) = \mathbb{P}(W(t) + \pi = 2\pi) = 0$, so, with probability 1, $X(t)$ can be inverted to get $W(t)$. Since $W(t)$ is a Markov process, it follows that $\mathbb{P}(X(t+h) \in B | X(t) = x) = \mathbb{P}\left(X(t+h) \in B | \mathcal{F}_t^X\right)$ almost surely, i.e., X is Markov. Now consider $\tau = \inf_t X(t) = (0, 0)$, the hitting time of the origin. This is clearly an F^X-optional time, and equally clearly almost surely finite, because, with probability 1, $W(t)$ will leave the interval $(-\pi, \pi)$ within a finite *time. But, equally clearly, the future behavior of X will be very different if it hits the origin because $W = \pi$ or because $W = -\pi$, which cannot be determined* just *from X. Hence, there is at least one optional time at which X is not strongly Markovian, so X is not a strong Markov process.*

14.2 Markov Families

We have been fairly cavalier about the idea of a Markov process having a particular initial state or initial distribution, basically relying on our familiarity with these ideas from elementary courses on stochastic processes. For future purposes, however, it is helpful to bring this notions formally within our general framework, and to fix some notation.

Definition 139 (Initial Distribution, Initial State) *Let Ξ be a Borel space with σ-field $\mathcal{X}$, T be a one-sided index set, and $\mu_{t,s}$ be a collection of Markovian transition kernels on Ξ. Then the* Markov process with initial distribution ν, *X_ν, is the Markov process whose finite-dimensional distributions are given by the action of $\mu_{t,s}$ on ν. That is, for all $0 \leq t_1 \leq t_2 \leq \ldots \leq t_n$,*

$$X_\nu(0), X_\nu(t_1), X_\nu(t_2), \ldots X_\nu(t_n) \quad \sim \quad \nu \otimes \mu_{0,t_1} \otimes \mu_{t_1,t_2} \otimes \ldots \otimes \mu_{t_{n-1},t_n} \tag{14.2}$$

If $\nu = \delta(x - a)$, the delta distribution at a, then we write X_a and call it the Markov process with initial state a.

The existence of processes with given initial distributions and initial states is a trivial consequence of Theorem 103, our general existence result for Markov processes.

Lemma 140 *For every initial state x, there is a probability distribution P_x on $\Xi^T, \mathcal{X}^T$. The function $P_x(A) : \Xi \times \mathcal{X}^T \rightarrow [0, 1]$ is a probability kernel.*

Proof: The initial state fixes all the finite-dimensional distributions, so the existence of the probability distribution follows from Theorem 23. The fact that $P_x(A)$ is a kernel is a straightforward application of the definition of kernels (Definition 30). □

Definition 141 *The* Markov family *corresponding to a given set of transition kernels $\mu_{t,s}$ is the collection of all P_x.*

That is, rather than thinking of a different stochastic process for each initial state, we can simply think of different distributions over the path space Ξ^T. This suggests the following definition.

Definition 142 *For a given initial distribution ν on Ξ, we define a distribution on the paths in a Markov family as, $\forall A \in \mathcal{X}^T$,*

$$P_\nu(A) \quad \equiv \quad \int_\Xi P_x(A)\nu(dx) \tag{14.3}$$

In physical contexts, we sometimes refer to distributions ν as *mixed states*, as opposed to the *pure states* x, because the path-space distributions induced by the former are mixtures of the distributions induced by the latter. You should check that the distribution over paths given by a Markov process with initial distribution ν, according to Definition 139, agrees with that given by Definition 142.

14.3 Feller Processes

Working in the early 1950s, Feller showed that, by imposing very reasonable conditions on the semi-group of evolution operators corresponding to a homogeneous Markov process, one could obtain very powerful results about the near-continuity of sample paths (namely, the existence of cadlag versions), about the strong Markov property, etc. Ever since, processes with such nice semi-groups have been known as *Feller processes*, or sometimes as *Feller-Dynkin processes*, in recognition of Dynkin's work in extending Feller's original approach. Unfortunately, to first order there are as many definitions of a Feller semi-group as there are books on Markov processes. I am going to try to follow Kallenberg as closely as possible, because his version is pretty clearly motivated, and you've already got it.

One point to notice is that, in developing the theory of Feller operators, we need to switch from operators on L_1, where we have been working before, to operators on L_∞. The L_∞ norm, $\sup_x |f(x)|$, is much stronger than the L_1 norm, $\int |f(x)|\mu(dx)$, and the former will let us make some regularity arguments which just aren't possible in the latter, at least not without a lot of extra machinery and assumptions.

As usual, we warm up with some definitions.

Definition 143 (Positive Operator) *An operator O is* positive *when $f \geq 0$ a.e. implies $Of \geq 0$ a.e.*

Definition 144 (Contraction Operator) *An operator O is an L_p-*contraction *when $\|Of\|_p \leq \|f\|_p$.*

Definition 145 (Strongly Continuous Semigroup) *A semigroup of operators O_t is* strongly continuous in the L_p sense *on a set of functions L when, $\forall f \in L$*

$$\lim_{t \to 0} \|O_t f - f\|_p = 0 \tag{14.4}$$

In the two preceding definitions, the p in L_p should be understood to be anything from 1 to ∞ inclusive.

Definition 146 (Conservative Operator) *An operator O is* conservative *when $O\mathbf{1}_\Xi = \mathbf{1}_\Xi$.*

In these terms, our earlier Markov operators are linear, positive, conservative L_1 contractions.

Lemma 147 *If O_t is a strongly continuous semigroup of linear L_p contractions, then, for each f, $O_t f$ is a continuous function of t.*

PROOF: Continuity here means that $\lim_{t' \rightarrow t} \|O_{t'} f - O_t\|_p = 0$ — we are using the L_p norm as our metric in function space. Consider first the limit from above:

$$\begin{aligned} \|O_{t+h} f - O_t f\|_p &= \|O_t(O_h f - f)\|_p && (14.5) \\ &\leq \|O_h f - f\|_p && (14.6) \end{aligned}$$

since the operators are contractions. Because they are strongly continuous, $\|O_h f - f\|_p$ can be made smaller than any $\epsilon > 0$ by taking h sufficiently small. Hence $\lim_{h \downarrow 0} O_{t+h} f$ exists and is $O_t f$. Similarly, for the limit from below,

$$\begin{aligned} \|O_{t-h} f - O_t f\|_p &= \|O_t f - O_{t-h} f\|_p && (14.7) \\ &= \|O_{t-h}(O_h f - f)\|_p && (14.8) \\ &\leq \|O_h f - f\|_p && (14.9) \end{aligned}$$

using the contraction property again. So $\lim_{h \downarrow 0} O_{t-h} f = O_t f$, also, and we can just say that $\lim_{t' \rightarrow t} O_{t'} f = O_t f$. □

Remark: The result actually holds if we just assume strong continuity, without contraction, but the proof isn't so pretty; see Ethier and Kurtz (1986, ch. 1, corollary 1.2, p. 7).

Definition 148 (Feller Semigroup) *A semigroup of linear, positive, conservative L_∞ contraction operators K_t is a* Feller semigroup *if, for every $f \in C_0$ and $x \in \Xi$, (Definition 127),*

$$\begin{aligned} K_t f &\in C_0 && (14.10) \\ \lim_{t \rightarrow 0} K_t f(x) &= f(x) && (14.11) \end{aligned}$$

Remark: Some authors omit the requirement that K_t be conservative. Also, this is just the homogeneous case, and one can define inhomogeneous Feller semigroups. However, the homogeneous case will be plenty of work enough for us!

Definition 149 (Feller Process) *A homogeneous Markov family X is a* Feller process *when, for all $x \in \Xi$,*

$$\begin{aligned} \forall t,\ y \rightarrow x &\Rightarrow X_y(t) \xrightarrow{d} X_x(t) && (14.12) \\ t \rightarrow 0 &\Rightarrow X_x(t) \xrightarrow{P} x && (14.13) \end{aligned}$$

Lemma 150 *Eq. 14.10 holds if and only if Eq. 14.12 does.*

PROOF: Exercise 14.2. □

Lemma 151 *Eq. 14.11 holds if and only if Eq. 14.13 does.*

PROOF: Exercise 14.3. □

Theorem 152 *A Markov process is a Feller process if and only if its evolution operators form a Feller semigroup.*

PROOF: Combine the lemmas. □

Feller semigroups in continuous time have generators, as in Chapter 12. In fact, the generator is *especially* useful for Feller semigroups, as seen by this theorem.

Theorem 153 (Generator of a Feller Semigroup) *If K_t and H_t are Feller semigroups with generator G, then $K_t = H_t$.*

PROOF: Because Feller semigroups consist of contractions, the Hille-Yosida Theorem 130 applies, and, for every positive λ, the resolvent $R_\lambda = (\lambda I - G)^{-1}$. Hence, if K_t and H_t have the same generator, they have the same resolvent operators. But this means that, for every $f \in C_0$ and x, $K_t f(x)$ and $H_t f(x)$ have the same Laplace transforms. Since, by Eq. 14.11 $K_t f(x)$ and $H_t f(x)$ are both right-continuous, their Laplace transforms are unique, so $K_t f(x) = H_t f(x)$. □

Theorem 154 *Every Feller semigroup K_t with generator G is strongly continuous on* Dom(G).

PROOF: From Corollary 126, we have, as seen in Chapter 13, for all $t \geq 0$,

$$K_t f - f = \int_0^t K_s G f ds \qquad (14.14)$$

Clearly, the right-hand side goes to zero as $t \to 0$. □

The two most important properties of Feller processes is that they are cadlag (or, rather, always have cadlag versions), and that they are strongly Markovian. First, let's look at the cadlag property. We need a result which I really should have put in Chapter 8.

Proposition 155 *Let Ξ be a locally compact, separable metric space with metric ρ, and let X be a separable Ξ-valued stochastic process on T. For given $\epsilon, \delta > 0$, define $\alpha(\epsilon, \delta)$ to be*

$$\inf_{\Gamma \in \mathcal{F}_s^X :\ \mathbb{P}(\Gamma)=1} \sup_{s,t \in T:\ s \leq t \leq s+\delta} \mathbb{P}\left(\omega :\ \rho(X(s,\omega), X(t,\omega)) \geq \epsilon,\ \omega \in \Gamma | \mathcal{F}_s^X\right) \qquad (14.15)$$

If, for all ϵ,

$$\lim_{\delta \to 0} \alpha(\epsilon, \delta) = 0 \qquad (14.16)$$

then X has a cadlag version.

Proof: Combine Theorem 2 and Theorem 3 of Gikhman and Skorokhod (1965/1969, Chapter IV, Section 4). □

Lemma 156 *Let X be a separable homogeneous Markov process. Define*

$$\alpha(\epsilon, \delta) = \sup_{t \in T:\ 0 \leq t \leq \delta;\ x \in \Xi} \mathbb{P}\left(\rho(X_x(t), x) \geq \epsilon\right) \tag{14.17}$$

If, for every $\epsilon > 0$,

$$\lim_{\delta \rightarrow 0} \alpha(\epsilon, \delta) = 0 \tag{14.18}$$

then X has a cadlag version.

Proof: The α in this lemma is clearly the α in the preceding proposition, using the fact that X is Markovian with respect to its natural filtration and homogeneous. □

Lemma 157 *A separable homogeneous Markov process X has a cadlag version if*

$$\lim_{\delta \downarrow 0} \sup_{x \in \Xi,\ 0 \leq t \leq \delta} \mathbf{E}\left[\rho(X_x(t), x)\right] = 0 \tag{14.19}$$

Proof: Start with the Markov inequality.

$$\forall x, t > 0, \epsilon > 0,\ \mathbb{P}\left(\rho(X_x(t), x) \geq \epsilon\right) \leq \frac{\mathbf{E}\left[\rho(X_x(t), x)\right]}{\epsilon} \tag{14.20}$$

$$\forall x, \delta > 0, \epsilon > 0,\ \sup_{0 \leq t \leq \delta} \mathbb{P}\left(\rho(X_x(t), x) \geq \epsilon\right) \leq \sup_{0 \leq t \leq \delta} \frac{\mathbf{E}\left[\rho(X_x(t), x)\right]}{\epsilon} \tag{14.21}$$

$$\forall \delta > 0, \epsilon > 0,\ \sup_{x,\ 0 \leq t \leq \delta} \mathbb{P}\left(\rho(X_x(\delta), x) \geq \epsilon\right) \leq \frac{1}{\epsilon} \sup_{x,\ 0 \leq t \leq \delta} \mathbf{E}\left[\rho(X_x(\delta), x)\right] \tag{14.22}$$

Taking the limit as $\delta \downarrow 0$, we have, for all $\epsilon > 0$,

$$\lim_{\delta \rightarrow 0} \alpha(\epsilon, \delta) \leq \frac{1}{\epsilon} \lim_{\delta \downarrow 0} \sup_{x,\ 0 \leq t \leq \delta} \mathbf{E}\left[\rho(X_x(\delta), x)\right] = 0 \tag{14.23}$$

So the preceding lemma applies. □

Theorem 158 (Feller Implies Cadlag) *Every Feller process X has a cadlag version.*

Proof: First, by the usual arguments, we can get a separable version of X. Next, we want to show that the last lemma is satisfied. Notice that, because Ξ is compact, $\lim_x \rho(x_n, x) = 0$ if and only if $f_k(x_n) \rightarrow f_k(x)$, for all f_k in some countable dense subset of the continuous functions on the state space.[1] Since the Feller semigroup is strongly continuous on the domain of its generator (Theorem 154), and that domain is dense in C_0 by the Hille-Yosida Theorem (130), we can pick our f_k to be in this class. The strong continuity is with respect to the L_∞ norm, so $\sup_x |K_t f(x) - K_s f(x)| = \sup_x |K_s(K_{t-s}f(x) - f(x))| \rightarrow 0$ as $t-s \rightarrow 0$, for every $f \in C_0$. But $\sup_x |K_t f(x) - K_s f(x)| = \sup_x \mathbf{E}\left[|f(X_x(t)) - f(X_x(s))|\right]$. So $\sup_{x,\ 0 \leq t \leq \delta} \mathbf{E}\left[|f(X_x(t)) - f(x)|\right] \rightarrow 0$ as $\delta \rightarrow 0$. Now Lemma 157 applies. □

Remark: Kallenberg (Theorem 19.15, p. 379) gives a different proof, using the existence of cadlag paths for certain kinds of supermartingales, which he builds using the resolvent operator. This seems to be the favored approach among modern authors, but obscures, somewhat, the work which the Feller properties do in getting the conclusion.

Theorem 159 (Feller Processes are Strongly Markovian) *Any Feller process X is strongly Markovian with respect to $\mathcal{F}^{X+}$, the right-continuous version of its natural filtration.*

PROOF: The strong Markov property holds if and only if, for all bounded, continuous functions f, $t \geq 0$ and $\mathcal{F}^{X+}$-optional times τ,

$$\mathbf{E}\left[f(X(\tau+t))|\mathcal{F}^{X+}_\tau\right] = K_t f(X(\tau)) \tag{14.24}$$

We'll show this holds for arbitrary, fixed choices of f, t and τ. First, we discretize time, to exploit the fact that the Markov and strong Markov properties coincide for discrete parameter processes. For every $h > 0$, set

$$\tau_h \equiv \inf_u \{u \geq \tau :\ u = kh,\ k \in \mathbb{N}\} \tag{14.25}$$

Now τ_h is almost surely finite (because τ is), and $\tau_h \rightarrow \tau$ a.s. as $h \rightarrow 0$. We construct the discrete-parameter sequence $X_h(n) = X(nh)$, $n \in \mathbb{N}$. This is a Markov sequence with respect to the natural filtration, i.e., for every bounded continuous f and $m \in \mathbb{N}$,

$$\mathbf{E}\left[f(X_h(n+m))|\mathcal{F}^X_n\right] = K_{mh} f(X_h(n)) \tag{14.26}$$

Since the Markov and strong Markov properties coincide for Markov sequences, we can now assert that

$$\mathbf{E}\left[f(X(\tau_h+mh))|\mathcal{F}^X_{\tau_h}\right] = K_{mh} f(X(\tau_h)) \tag{14.27}$$

Since $\tau_h \geq \tau$, $\mathcal{F}^X_\tau \subseteq \mathcal{F}^X_{\tau_h}$. Now pick any set $B \in \mathcal{F}^{X+}_\tau$ and use smoothing:

$$\mathbf{E}\left[f(X(\tau_h+t))\mathbf{1}_B\right] = \mathbf{E}\left[K_t f(X(\tau_h))\mathbf{1}_B\right] \tag{14.28}$$

$$\mathbf{E}\left[f(X(\tau+t))\mathbf{1}_B\right] = \mathbf{E}\left[K_t f(X(\tau))\mathbf{1}_B\right] \tag{14.29}$$

[1] Roughly speaking, if $f(x_n) \rightarrow f(x)$ for *all* continuous functions f, it should be obvious that there is no way to avoid having $x_n \rightarrow x$. Picking a countable dense subset of functions is still enough.

where we let $h \downarrow 0$, and invoke the fact that $X(t)$ is right-continuous (Theorem 158) and $K_t f$ is continuous. Since this holds for arbitrary $B \in \mathcal{F}_\tau^{X+}$, and $K_t f(X(\tau))$ has to be $\mathcal{F}_\tau^{X+}$-measurable, we have that

$$\mathbf{E}\left[f(X(\tau+t))|\mathcal{F}_\tau^{X+}\right] = K_t f(X(\tau)) \tag{14.30}$$

as required. □

Here is a useful consequence of Feller property, related to the martingale-problem properties we saw last time.

Theorem 160 (Dynkin's Formula) *Let X be a Feller process with generator G. Let α and β be two almost-surely-finite $\mathcal{F}$-optional times, $\alpha \leq \beta$. Then, for every continuous $f \in \mathrm{Dom}(G)$,*

$$\mathbf{E}\left[f(X(\beta)) - f(X(\alpha))\right] = \mathbf{E}\left[\int_\alpha^\beta Gf(X(t))dt\right] \tag{14.31}$$

Proof: Exercise 14.4. □

Remark: A large number of results very similar to Eq. 14.31 are *also* called "Dynkin's formula". For instance, Rogers and Williams (1994, ch. III, sec. 10, pp. 253–254) give that name to *three* different equations. Be careful about what people mean!

14.4 Exercises

Exercise 14.1 (Yet Another Interpretation of the Resolvents) *Consider again a homogeneous Markov process with transition kernel μ_t. Let τ be an exponentially-distributed random variable with rate λ, independent of X. Show that $\mathbf{E}[K_\tau f(x)] = \lambda R_\lambda f(x)$.*

Exercise 14.2 (The First Pair of Feller Properties) *Prove Lemma 150. Hint: you may use the fact that, for measures, $\nu_t \rightarrow \nu$ if and only if $\nu_t f \rightarrow \nu f$, for every bounded, continuous f.*

Exercise 14.3 (The Second Pair of Feller Properties) *Prove Lemma 151.*

Exercise 14.4 (Dynkin's Formula) *Prove Theorem 160*

Chapter 15

Convergence of Feller Processes

This chapter looks at the convergence of sequences of Feller processes to a limiting process.

Section 15.1 lays some ground work concerning weak convergence of processes with cadlag sample paths.

Section 15.2 states and proves the central theorem about the convergence of sequences of Feller processes.

Section 15.3 examines a particularly important special case, the approximation of ordinary differential equations by pure-jump Markov processes.

15.1 Weak Convergence of Processes with Cadlag Paths (The Skorokhod Topology)

Recall that a sequence of random variables $X_1, X_2, \ldots$ converges in distribution on X, or weakly converges on X, $X_n \xrightarrow{d} X$, if and only if $\mathbf{E}\left[f(X_n)\right] \rightarrow \mathbf{E}\left[f(X)\right]$, for all bounded, continuous functions f. This is still true when X_n are random functions, i.e., stochastic processes, only now the relevant functions f are functionals of the sample paths.

Definition 161 (Convergence in Finite-Dimensional Distribution) *Random processes X_n on T* converge in finite-dimensional distribution *on X, $X_n \xrightarrow{fd} X$, when, $\forall J \in \mathrm{Fin}(T)$, $X_n(J) \xrightarrow{d} X(J)$.*

Proposition 162 *Convergence in finite-dimensional distribution is necessary but not sufficient for convergence in distribution.*

PROOF: Necessity is obvious: the coordinate projections π_t are continuous functionals of the sample path, so they must converge if the distributions converge. Insufficiency stems from the problem that, even if a sequence of X_n all have sample paths in some set U, the limiting process might not: recall our example (78) of the version of the Wiener process with unmeasurable suprema. □

Definition 163 (The Space D) *By* $\mathbf{D}(T, \Xi)$ *we denote the space of all cadlag functions from* T *to* Ξ*. By default,* $\mathbf{D}$ *will mean* $\mathbf{D}(\mathbb{R}^+, \Xi)$*.*

$\mathbf{D}$ admits of multiple topologies. For most purposes, the most convenient one is the *Skorokhod topology*, a.k.a. the J_1 *topology* or the *Skorokhod* J_1 *topology*, which makes $\mathbf{D}(\Xi)$ a complete separable metric space when Ξ is itself complete and separable. (See Appendix A2 of Kallenberg.) For our purposes, we need only the following notion and theorem.

Definition 164 (Modified Modulus of Continuity) *The* modified modulus of continuity *of a function* $x \in \mathbf{D}(T, \Xi)$ *at time* $t \in T$ *and scale* $h > 0$ *is given by*

$$w(x, t, h) \equiv \inf_{(I_k)} \max_k \sup_{r,s \in I_k} \rho(x(s), x(r)) \tag{15.1}$$

where the infimum is over partitions of $[0, t)$ *into half-open intervals whose length is at least* h *(except possibly for the last one). Because* x *is cadlag, for fixed* x *and* t*,* $w(x, t, h) \rightarrow 0$ *as* $h \rightarrow 0$*.*

Theorem 165 (Weak Convergence in $\mathbf{D}(\mathbb{R}^+, \Xi)$) *Let* Ξ *be a complete, separable metric space. Then a sequence of random functions* $X_1, X_2, \ldots \in \mathbf{D}(\mathbb{R}^+, \Xi)$ *converges in distribution to* $X \in \mathbf{D}$ *if and only if*

i The set $T_c = \{t \in T : X(t) = X(t^-)\}$ *has a countable dense subset* T_0*, and the finite-dimensional distributions of the* X_n *converge on those of* X *on* T_0*.*

ii For every t*,*

$$\lim_{h \rightarrow 0} \limsup_{n \rightarrow \infty} \mathbf{E}\left[w(X_n, t, h) \wedge 1\right] = 0 \tag{15.2}$$

PROOF: See Kallenberg, Theorem 16.10, pp. 313–314. □

Theorem 166 (Sufficient Condition for Weak Convergence) *The following three conditions are all equivalent, and all imply condition (ii) in Theorem 165.*

1. *For any sequence of a.s.-finite* $\mathcal{F}^{X_n}$*-optional times* τ_n *and positive constants* $h_n \rightarrow 0$*,*

$$\rho(X_n(\tau_n), X_n(\tau_n + h_n)) \xrightarrow{P} 0 \tag{15.3}$$

2. *For all $t > 0$, for all*

$$\lim_{h \to 0} \limsup_{n \to \infty} \sup_{\sigma, \tau} \mathbf{E}\left[\rho(X_n(\sigma), X_n(\tau)) \wedge 1\right] = 0 \tag{15.4}$$

where σ and τ are $\mathcal{F}^{X_n}$-optional times $\sigma, \tau \leq t$, with $\sigma \leq \tau \leq \tau + h$.

3. *For all $t > 0$,*

$$\lim_{\delta \to 0} \limsup_{n \to \infty} \sup_{\tau \leq t} \sup_{0 \leq h \leq \delta} \mathbf{E}\left[\rho(X_n(\tau), X_n(\tau + h)) \wedge 1\right] = 0 \tag{15.5}$$

where the supremum in τ runs over all F^{X_n}-optional times $\leq t$.

Proof: See Kallenberg, Theorem 16.11, pp. 314–315. □

15.2 Convergence of Feller Processes

We need some technical notions about generators.

Definition 167 (Closed and Closable Generators, Closures) *A linear operator O on a Banach space $\mathcal{B}$ is* closed *if its* graph *— $\left\{f, g \in \mathcal{B}^2 : f \in \mathrm{Dom}(O), g = Of\right\}$ — is a closed set. An operator is* closable *if the closure of its graph is a function (and not just a relation). The* closure *of a closable operator is that function.*

Notice, by the way, that because O is linear, it is closable iff $f_n \to 0$ and $Af_n \to g$ implies $g = 0$.

Definition 168 (Core of an Operator) *Let O be a closed linear operator on a Banach space $\mathcal{B}$. A linear subspace $D \subseteq \mathrm{Dom}(O)$ is a* core *of O if the closure of O restricted to D is, again O.*

The idea of a core is that we can get away with knowing how the operator works on a linear subspace, which is often much easier to deal with, rather than controlling how it acts on its whole domain.

Proposition 169 *The generator of every Feller semigroup is closed.*

Proof: We need to show that the graph of G contains all of its limit points, that is, if $f_n \in \mathrm{Dom}(G)$ converges (in L_∞) on f, and $Gf_n \to g$, then $f \in \mathrm{Dom}(G)$ and $Gf = g$. First we show that $f \in \mathrm{Dom}(G)$.

$$\lim_{n \to \infty} (I - G)f_n = \lim_n f_n - \lim_n Gf_n \tag{15.6}$$

$$= f - g \tag{15.7}$$

But $(I - G)^{-1} = R_1$. Since this is a bounded linear operator, we can exchange applying the inverse and taking the limit, i.e.,

$$\begin{aligned} R_1 \lim_n (I - G) f_n &= R_1(f - g) && (15.8)\\ \lim_n R_1 (I - G) f_n &= R_1(f - g) && (15.9)\\ \lim_n f_n &= R_1(f - g) && (15.10)\\ f &= R_1(f - g) && (15.11) \end{aligned}$$

Since the range of the resolvents is contained in the domain of the generator, $f \in \mathrm{Dom}(G)$. We can therefore say that $f - g = (I - G)f$, which implies that $Gf = g$. Hence, the graph of G contains all its limit points, and G is closed. □

Theorem 170 *Let X_n be a sequence of Feller processes with semigroups $K_{n,t}$ and generators G_n, and X be another Feller process with semigroup K_t and a generator G containing a core D. Then the following are equivalent.*

1. *If $f \in D$, there exists a sequence of $f_n \in \mathrm{Dom}(G_n)$ such that $\|f_n - f\|_\infty \rightarrow 0$ and $\|A_n f_n - Af\|_\infty \rightarrow 0$.*
2. *$K_{n,t} \rightarrow K_t$ for every $t > 0$*
3. *$\|K_{n,t} f - K_t f\|_\infty \rightarrow 0$ for each $f \in C_0$, uniformly for bounded positive t*
4. *If $X_n(0) \xrightarrow{d} X(0)$ in Ξ, then $X_n \xrightarrow{d} X$ in $\mathbf{D}$.*

Proof: See Kallenberg, Theorem 19.25, p. 385. □

Remark. The important versions of the property above are the second — convergence of the semigroups — and the fourth — converge in distribution of the processes. The other two are there to simplify the proof.

15.3 Approximation of Ordinary Differential Equations by Markov Processes

The following result, due to Kurtz (1970, 1971), is essentially an application of Theorem 170.

First, recall that continuous-time, discrete-state Markov processes work essentially like a combination of a Poisson process (giving the time of transitions) with a Markov chain (giving the state moved to on transitions). This can be generalized to continuous-time, continuous-state processes, of what are called "pure jump" type.

Definition 171 (Pure Jump Markov Process) *A continuous-parameter Markov process is a* pure jump *process when its sample paths are piece-wise constant. For each state, there is an exponential distribution of times spent in that state, whose parameter is denoted $\lambda(x)$, and a transition probability kernel or exit distribution $\mu(x, B)$.*

Observe that pure-jump Markov processes always have cadlag sample paths. Also observe that the average amount of time the process spends in state x, once it jumps there, is $1/\lambda(x)$. So the time-average "velocity", i.e., rate of change, starting from x,

$$\lambda(x) \int_{\Xi} (y-x)\mu(x,dy)$$

Theorem 172 *Let X_n be a sequence of pure-jump Markov processes with state spaces Ξ_n, holding time parameters λ_n and transition probabilities μ_n. Suppose that, for all n Ξ_n is a Borel-measurable subset of $\mathbb{R}^k$ for some k. Let Ξ be another measurable subset of $\mathbb{R}^k$, on which there exists a function $F(x)$ such that $|F(x)-F(y)| \leq M|x-y|$ for some constant M. Suppose all of the following conditions holds.*

1. *The time-averaged rate of change is always finite:*

$$\sup_n \sup_{x \in \Xi_n \cap \Xi} \lambda_n(x) \int_{\Xi_n} |y-x|\mu_n(x,dy) \quad < \quad \infty \tag{15.12}$$

2. *There exists a positive sequence $\epsilon_n \rightarrow 0$ such that*

$$\lim_{n\rightarrow\infty} \sup_{x \in \Xi_n \cap \Xi} \lambda_n(x) \int_{|y-x|>\epsilon} |y-x|\mu_n(x,dy) \quad = \quad 0 \tag{15.13}$$

3. *The worst-case difference between $F(x)$ and the time-averaged rates of change goes to zero:*

$$\lim_{n\rightarrow\infty} \sup_{x \in \Xi_n \cap \Xi} \left| F(x) - \lambda_n(x) \int (y-x)\mu_n(x,dy) \right| \quad = \quad 0 \tag{15.14}$$

Let $X(s, x_0)$ be the solution to the initial-value problem where the differential is given by F, i.e., for each $0 \leq s \leq t$,

$$\frac{\partial}{\partial s} X(s,x_0) \quad = \quad F(X(s,x_0)) \tag{15.15}$$

$$X(0,x_0) \quad = \quad x_0 \tag{15.16}$$

and suppose there exists an $\eta > 0$ such that, for all n,

$$\Xi_n \cap \left\{ y \in \mathbb{R}^k : \inf_{0 \leq s \leq t} |y - X(s,x_0)| \leq \eta \right\} \quad \subseteq \quad \Xi \tag{15.17}$$

Then $\lim X_n(0) = x_0$ implies that, for every $\delta > 0$,

$$\lim_{n\rightarrow\infty} \mathbb{P}\left(\sup_{0 \leq s \leq t} |X_n(s) - X(s,x_0)| > \delta \right) \quad = \quad 0 \tag{15.18}$$

The first conditions on the X_n basically make sure that they are Feller processes. The subsequent ones make sure that the mean time-averaged rate of change of the jump processes converges on the instantaneous rate of change of the differential equation, and that, if we're sufficiently close to the solution of the differential equation in $\mathbb{R}^k$, we're not in some weird way outside the relevant domains of definition. Even though Theorem 170 is about weak convergence, converging in distribution on a non-random object is the same as converging in probability, which is how we get uniform-in-time convergence in probability for a conclusion.

There are, broadly speaking, two kinds of uses for this result. One kind is practical, and has to do with justifying convenient approximations. If n is large, we can get away with using an ODE instead of the noisy stochastic scheme, or alternately we can use stochastic simulation to approximate the solutions of ugly ODEs. The other kind is theoretical, about showing that the large-population limit behaves deterministically, even when the individual behavior is stochastic *and strongly dependent over time.*

Chapter 16

Convergence of Random Walks

This lecture examines the convergence of random walks to the Wiener process. This is very important both physically and statistically, and illustrates the utility of the theory of Feller processes.

Section 16.1 finds the semi-group of the Wiener process, shows it satisfies the Feller properties, and finds its generator.

Section 16.2 turns random walks into cadlag processes, and gives a fairly easy proof that they converge on the Wiener process.

16.1 The Wiener Process is Feller

Recall that the Wiener process $W(t)$ is defined by starting at the origin, by independent increments over non-overlapping intervals, by the Gaussian distribution of increments, and by continuity of sample paths (Examples 38 and 78). The process is homogeneous, and the transition kernels are (Section 11.1)

$$\mu_t(w_1, B) = \int_B dw_2 \frac{1}{\sqrt{2\pi t}} e^{-\frac{(w_2-w_1)^2}{2t}} \tag{16.1}$$

$$\frac{d\mu_t(w_1, w_2)}{d\lambda} = \frac{1}{\sqrt{2\pi t}} e^{-\frac{(w_2-w_1)^2}{2t}} \tag{16.2}$$

where the second line gives the density of the transition kernel with respect to Lebesgue measure.

Since the kernels are known, we can write down the corresponding evolution operators:

$$K_t f(w_1) = \int dw_2 f(w_2) \frac{1}{\sqrt{2\pi t}} e^{-\frac{(w_2-w_1)^2}{2t}} \tag{16.3}$$

We saw in Section 11.1 that the kernels have the semi-group property, so the evolution operators do too.

Let's check that $\{K_t\}, t \geq 0$ is a Feller semi-group. The first Feller property is easier to check in its probabilistic form, that, for all t, $y \rightarrow x$ implies $W_y(t) \xrightarrow{d} W_x(t)$. The distribution of $W_x(t)$ is just $\mathcal{N}(x,t)$, and it is indeed true that $y \rightarrow x$ implies $\mathcal{N}(y,t) \rightarrow \mathcal{N}(x,t)$. The second Feller property can be checked in its semi-group form: as $t \rightarrow 0$, $\mu_t(w_1, B)$ approaches $\delta(w - w_1)$, so $\lim_{t \rightarrow 0} K_t f(x) = f(x)$. Thus, the Wiener process is a Feller process. This implies that it has cadlag sample paths (Theorem 158), but we already knew that, since we know it's continuous. What we did not know was that the Wiener process is not just Markov but strong Markov, which follows from Theorem 159.

It's easier to find the generator of $\{K_t\}, t \geq 0$, it will help to re-write it in an equivalent form, as

$$K_t f(w) = \mathbf{E}\left[f(w + Z\sqrt{t})\right] \tag{16.4}$$

where Z is an independent $\mathcal{N}(0,1)$ random variable. (You should convince yourself that this is equivalent.) Now let's pick an $f \in C_0$ which is also twice continuously differentiable, i.e., $f \in C_0 \cap C^2$. Look at $K_t f(w) - f(w)$, and apply Taylor's theorem, expanding around w:

$$K_t f(w) - f(w) = \mathbf{E}\left[f(w + Z\sqrt{t})\right] - f(w) \tag{16.5}$$

$$= \mathbf{E}\left[f(w + Z\sqrt{t}) - f(w)\right] \tag{16.6}$$

$$= \mathbf{E}\left[Z\sqrt{t}f'(w) + \frac{1}{2}tZ^2 f''(w) + R(Z\sqrt{t})\right] \tag{16.7}$$

$$= \sqrt{t}f'(w)\mathbf{E}\left[Z\right] + t\frac{f''(w)}{2}\mathbf{E}\left[Z^2\right] + \mathbf{E}\left[R(Z\sqrt{t})\right] \tag{16.8}$$

$$\lim_{t \downarrow 0} \frac{K_t f(w) - f(w)}{t} = \frac{1}{2}f''(w) + \lim_{t \downarrow 0} \frac{\mathbf{E}\left[R(Z\sqrt{t})\right]}{t} \tag{16.9}$$

So, we need to investigate the behavior of the remainder term $R(Z\sqrt{t})$. We know from Taylor's theorem that

$$R(Z\sqrt{t}) = \frac{tZ^2}{2}\int_0^1 du\ f''(w + uZ\sqrt{t}) - f''(w) \tag{16.10}$$

$$\tag{16.11}$$

Since $f \in C_0 \cap C^2$, we know that $f'' \in C_0$. Therefore, f'' is uniformly continuous, and has a modulus of continuity,

$$m(f'', h) = \sup_{x,y:\ |x-y| \leq h} |f''(x) - f'\prime(y)| \tag{16.12}$$

which goes to 0 as $h \downarrow 0$. Thus

$$\left|R(Z\sqrt{t})\right| \leq \frac{tZ^2}{2}m(f'', Z\sqrt{t}) \tag{16.13}$$

$$\lim_{t \rightarrow 0} \frac{\left|R(Z\sqrt{t})\right|}{t} \leq \lim_{t \rightarrow 0} \frac{Z^2 m(f'', Z\sqrt{t})}{2} \tag{16.14}$$

$$= 0 \tag{16.15}$$

Plugging back in to Equation 16.9,

$$\begin{aligned} Gf(w) &= \frac{1}{2}f''(w) + \lim_{t\downarrow 0}\frac{\mathbf{E}\left[R(Z\sqrt{t})\right]}{t} && (16.16) \\ &= \frac{1}{2}f''(w) && (16.17) \end{aligned}$$

That is, $G = \frac{1}{2}\frac{d^2}{dw^2}$, one half of the Laplacian. We have shown this only for $C_0 \cap C^2$, but this is clearly a linear subspace of C_0, and, since C^2 is dense in C, it is dense in C_0, i.e., this is a core for the generator. Hence the generator is really the extension of $\frac{1}{2}\frac{d^2}{dw^2}$ to the whole of C_0, but this is too cumbersome to repeat all the time, so we just say it's the Laplacian.

16.2 Convergence of Random Walks

Let $X_1, X_2, \ldots$ be a sequence of IID variables with mean 0 and variance 1. The random walk process S_n is then just $\sum_{i=1}^{n} X_i$. It is a discrete-time Markov process, and consequently also a strong Markov process. Imagine each step of the walk takes some time h, and imagine this time interval becoming smaller and smaller. Then, between any two times t_1 and t_2, the number of steps of the random walk will be about $\frac{t_2-t_1}{h}$, which will go to infinity. The displacement of the random walk between t_1 and t_2 will then be a sum of an increasingly large number of IID random variables, and by the central limit theorem will approach a Gaussian distribution. Moreover, if we look at the interval of time from t_2 to t_3, we will see another Gaussian, but all of the random-walk steps going into it will be independent of those going into our first interval. So, we expect that the random walk will in some sense come to look like the Wiener process, no matter what the exact distribution of the X_1. Let's consider this in more detail.

Define $Y_n(t) = n^{-1/2}\sum_{i=0}^{[nt]} X_i = n^{-1/2}S_{[nt]}$, where $X_0 = 0$ and $[nt]$ is the integer part of the real number nt. You should convince yourself that this is a Markov process, with cadlag sample paths.

We want to consider the limiting distribution of Y_n as $n \to \infty$. First of all, we should convince ourselves that a limit distribution exists. But this is not too hard. For any fixed t, $Y_n(t)$ approaches a Gaussian distribution by the central limit theorem. For any fixed finite collection of times $t_1 \leq t_2 \ldots \leq t_k$, $Y_n(t_1), Y_n(t_2), \ldots Y_n(t_k)$ approaches a limiting distribution if $Y_n(t_1), Y_n(t_2) - Y_n(t_1), \ldots Y_n(t_k) - Y_n(t_{k-1})$ does, but that again will be true by the (multivariate) central limit theorem. Since the limiting finite-dimensional distributions exist, some limiting distribution exists (via Theorem 23). It remains to identify it.

Lemma 173 $Y_n \overset{fd}{\to} W$.

PROOF: For all n, $Y_n(0) = 0 = W(0)$. For any $t_2 > t_1$,

$$\begin{aligned}
\mathcal{L}\left(Y_n(t_2) - Y_n(t_1)\right) &= \mathcal{L}\left(\frac{1}{\sqrt{n}} \sum_{i=[nt_1]}^{[nt_2]} X_i\right) && (16.18)\\
&\xrightarrow{d} \mathcal{N}(0, t_2 - t_1) && (16.19)\\
&= \mathcal{L}\left(W(t_2) - W(t_1)\right) && (16.20)
\end{aligned}$$

Finally, for any three times $t_1 < t_2 < t_3$, $Y_n(t_3) - Y_n(t_2)$ and $Y_n(t_2) - Y_n(t_1)$ are independent for sufficiently large n, because they become sums of disjoint collections of independent random variables. Thus, the limiting distribution of Y_n starts at the origin and has independent Gaussian increments. Since these properties determine the finite-dimensional distributions of the Wiener process, $Y_n \xrightarrow{fd} W$. □

Theorem 174 $Y_n \xrightarrow{d} W$.

PROOF: By Theorem 165, it is enough to show that $Y_n \xrightarrow{fd} W$, and that any of the properties in Theorem 166 hold. The lemma took care of the finite-dimensional convergence, so we can turn to the second part. A sufficient condition is property (1) inn the latter theorem, that $|Y_n(\tau_n + h_n) - Y_n(\tau_n)| \xrightarrow{P} 0$ for all finite optional times τ_n and any sequence of positive constants $h_n \rightarrow 0$.

$$\begin{aligned}
|Y_n(\tau_n + h_n) - Y_n(\tau_n)| &= n^{-1/2}\left|S_{[n\tau_n + nh_n]} - S_{[n\tau_n]}\right| && (16.21)\\
&\overset{d}{=} n^{-1/2}\left|S_{[nh_n]} - S_0\right| && (16.22)\\
&= n^{-1/2}\left|S_{[nh_n]}\right| && (16.23)\\
&= n^{-1/2}\left|\sum_{i=0}^{[nh_n]} X_i\right| && (16.24)
\end{aligned}$$

To see that this converges in probability to zero, we will appeal to Chebyshev's inequality: if Z_i have common mean 0 and variance σ^2, then, for every positive ϵ,

$$\mathbb{P}\left(\left|\sum_{i=1}^{m} Z_i\right| > \epsilon\right) \leq \frac{m\sigma^2}{\epsilon^2} \qquad (16.25)$$

Here we have $Z_i = X_i/\sqrt{n}$, so $\sigma^2 = 1/n$, and $m = [nh_n]$. Thus

$$\mathbb{P}\left(n^{-1/2}\left|S_{[nh_n]}\right| > \epsilon\right) \leq \frac{[nh_n]}{n\epsilon^2} \qquad (16.26)$$

As $0 \leq [nh_n]/n \leq h_n$, and $h_n \rightarrow 0$, the bounding probability must go to zero for every fixed ϵ. Hence $n^{-1/2}\left|S_{[nh_n]}\right| \xrightarrow{P} 0$. □

Corollary 175 (The Invariance Principle) *Let $X_1, X_2, \ldots$ be IID random variables with mean μ and variance σ^2. Then*

$$Y_n(t) \equiv \frac{1}{\sqrt{n}} \sum_{i=0}^{[nt]} \frac{X_i - \mu}{\sigma} \quad \xrightarrow{d} \quad W(t) \tag{16.27}$$

PROOF: $(X_i - \mu)/\sigma$ has mean 0 and variance 1, so Theorem 174 applies. □

This result is called "the invariance principle", because it says that the limiting distribution of the sequences of sums depends only on the mean and variance of the individual terms, and is consequently *invariant* under changes which leave those alone. Both this result and the previous one are known as the "functional central limit theorem", because convergence in distribution is the same as convergence of all bounded continuous *functionals* of the sample path. Another name is "Donsker's Theorem", which is sometimes associated however with the following corollary of Theorem 174.

Corollary 176 (Donsker's Theorem) *Let $Y_n(t)$ and $W(t)$ be as before, but restrict the index set T to the unit interval $[0, 1]$. Let f be any function from $\mathbf{D}([0, 1])$ to $\mathbb{R}$ which is measurable and a.s. continuous at W. Then $f(Y_n) \xrightarrow{d} f(W)$.*

PROOF: Exercise. □

This version is especially important for statistical purposes, as we'll see a bit later.

16.3 Exercises

Exercise 16.1 *Go through all the details of Example 138.*

a Show that $\mathcal{F}_t^X \subseteq \mathcal{F}_t^W$ for all t, and that $\mathcal{F}^X \subset \mathcal{F}^W$.

b Show that $\tau = \inf_t X(t) = (0, 0)$ is a $\mathcal{F}^X$-optional time, and that it is finite with probability 1.

c Show that X is Markov with respect to both its natural filtration and the natural filtration of the driving Wiener process.

d Show that X is not strongly Markov at τ.

e Which, if any, of the Feller properties does X have?

Exercise 16.2 *Consider a d-dimensional Wiener process, i.e., an $\mathbb{R}^d$-valued process where each coordinate is an independent Wiener process. Find the generator.*

Exercise 16.3 *Prove Donsker's Theorem (Corollary 176).*

Exercise 16.4 (Diffusion equation) *As mentioned in class, the partial differential equation*

$$\frac{1}{2}\frac{\partial^2 f(x,t)}{\partial x^2} = \frac{\partial f(x,t)}{\partial t}$$

is called the diffusion equation*. From our discussion of initial value problems in Chapter 12 (Corollary 126 and related material), it is clear that the function $f(x,t)$ solves the diffusion equation with initial condition $f(x,0)$ if and only if $f(x,t) = K_t f(x,0)$, where K_t is the evolution operator of the Wiener process.*

a *Take $f(x,0) = \left(2\pi 10^{-4}\right)^{-1/2} e^{-\frac{x^2}{2\cdot 10^{-4}}}$. $f(x,t)$ can be found analytically; do so.*

b *Estimate $f(x,10)$ over the interval $[-5,5]$ stochastically. Use the fact that $K_t f(x) = \mathbf{E}\left[f(W(t))|W(0)=x\right]$, and that random walks converge on the Wiener process. (Be careful that you scale your random walks the right way!) Give an indication of the error in this estimate.*

c *Can you find an analytical form for $f(x,t)$ if $f(x,0) = \mathbf{1}_{[-0.5,0.5]}(x)$?*

d *Find $f(x,10)$, with the new initial conditions, by numerical integration on the domain $[-10,10]$, and compare it to a stochastic estimate.*

Chapter 17

Diffusions and the Wiener Process

Section 17.1 introduces the ideas which will occupy us for the next few lectures, the continuous Markov processes known as diffusions, and their description in terms of stochastic calculus.

Section 17.2 collects some useful properties of the most important diffusion, the Wiener process.

Section 17.3 shows, first heuristically and then more rigorously, that almost all sample paths of the Wiener process don't have derivatives.

17.1 Diffusions and Stochastic Calculus

So far, we have looked at Markov processes in general, and then paid particular attention to Feller processes, because the Feller properties are very natural continuity assumptions to make about stochastic models and have very important consequences, especially the strong Markov property and cadlag sample paths. The natural next step is to go to Markov processes with continuous sample paths. The most important case, overwhelmingly dominating the literature, is that of *diffusions*.

Definition 177 (Diffusion) *A stochastic process X adapted to a filtration $\mathcal{F}$ is a* diffusion *when it is a strong Markov process with respect to $\mathcal{F}$, homogeneous in time, and has continuous sample paths.*[1]

Diffusions matter to us for several reasons. First, they are very natural models of many important systems — the motion of physical particles (the

[1] Having said that, I should confess that some authors don't insist that diffusions be homogeneous, and some even don't insist that they be *strong* Markov processes. But this is the general sense in which the term is used.

source of the term "diffusion"), fluid flows, noise in communication systems, financial time series, etc. Probabilistic and statistical studies of time-series data thus need to understand diffusions. Second, many discrete Markov models have large-scale limits which are diffusion processes: these are important in physics and chemistry, population genetics, queueing and network theory, certain aspects of learning theory[2], etc. These limits are often more tractable than more exact finite-size models. (We saw a hint of this in Section 15.3.) Third, many statistical-inferential problems can be described in terms of diffusions, most prominently ones which concern goodness of fit, the convergence of empirical distributions to true probabilities, and nonparametric estimation problems of many kinds.

The easiest way to get at diffusions is to through the theory of stochastic differential equations; the most important diffusions can be thought of as, roughly speaking, the result of adding a noise term to the right-hand side of a differential equation. A more exact statement is that, just as an autonomous ordinary differential equation

$$\frac{dx}{dt} = f(x),\ x(t_0) = x_0 \tag{17.1}$$

has the solution

$$x(t) = \int_{t_0}^{t} f(x)ds + x_0 \tag{17.2}$$

a stochastic differential equation

$$\frac{dX}{dt} = f(X) + g(X)\frac{dY}{dt},\ X(t_0) = x_0 \text{ a.s.} \tag{17.3}$$

where X and Y are stochastic processes, is solved by

$$X(t) = \int f(X)ds + \int g(X)dY + x_0 \tag{17.4}$$

where $\int g(X,t)dY$ is a *stochastic integral*. It turns out that, properly constructed, this sort of integral, and so this sort of stochastic differential equation, makes sense even when dY/dt does not make sense as any sort of ordinary derivative, so that the more usual way of writing an SDE is

$$dX = f(X)dt + g(X)dY,\ X(t_0) = x_0 \text{ a.s.} \tag{17.5}$$

even though this seems to invoke infinitessimals, which don't exist.[3]

[2] Specifically, discrete-time reinforcement learning converges to the continuous-time replicator equation of evolutionary theory.

[3] Some people, like Ethier and Kurtz (1986), prefer to talk about stochastic *integral* equations, rather than stochastic *differential* equations, because things like 17.5 are really shorthands for "find an X such that Eq. 17.4 holds", and objects like dX don't really make much sense on their own. There's a certain logic to this, but custom is overwhelmingly against them.

The fully general theory of stochastic calculus considers integration with respect to a very broad range of stochastic processes, but the original case, which is still the most important, is integration with respect to the Wiener process, which corresponds to driving a system with white noise. In addition to its many applications in all the areas which use diffusions, the theory of integration against the Wiener process occupies a central place in modern probability theory; I simply would not be doing my job if this course did not cover it. We therefore begin our study of diffusions and stochastic calculus by reviewing some of the properties of the Wiener process — which is also the most important diffusion process.

17.2 Once More with the Wiener Process and Its Properties

To review, the standard Wiener process $W(t)$ is defined by (i) $W(0) = 0$, (ii) centered Gaussian increments with linearly-growing variance, $\mathcal{L}(W(t_2) - W(t_1)) = \mathcal{N}(0, t_2 - t_1)$, (iii) independent increments and (iv) continuity of sample paths. We have seen that it is a homogeneous Markov process (Section 11.1), and in fact (Section 16.1) a Feller process (and therefore a strong Markov process), whose generator is $\frac{1}{2}\nabla^2$. By Definition 177, W is a diffusion.

This section proves a few more useful properties.

Proposition 178 *The Wiener process is a martingale with respect to its natural filtration.*

Proof: This follows directly from the Gaussian increment property:

$$\begin{aligned}
\mathbf{E}\left[W(t+h)|\mathcal{F}_t^X\right] &= \mathbf{E}\left[W(t+h)|W(t)\right] && (17.6)\\
&= \mathbf{E}\left[W(t+h) - W(t) + W(t)|W(t)\right] && (17.7)\\
&= \mathbf{E}\left[W(t+h) - W(t)|W(t)\right] + W(t) && (17.8)\\
&= 0 + W(t) = W(t) && (17.9)
\end{aligned}$$

where the first line uses the Markov property of W, and the last line the Gaussian increments property. □

Definition 179 *If $W(t,\omega)$ is adapted to a filtration $\mathcal{F}$ and is an $\mathcal{F}$-filtration, it is an $\mathcal{F}$* Wiener process *or $\mathcal{F}$* Brownian motion*.*

It seems natural to speak of the Wiener process as a Gaussian process. This motivates the following definition.

Definition 180 (Gaussian Process) *A real-valued stochastic process is Gaussian when all its finite-dimensional distributions are multivariate Gaussian distributions.*

Proposition 181 *The Wiener process is a Gaussian process.*

Proof: Pick any k times $t_1 < t_2 < \ldots < t_k$. Then the increments $W(t_1) - W(0)$, $W(t_2) - W(t_1)$, $W(t_3) - W(t_2)$, $\ldots$ $W(t_k) - W(t_{k-1})$ are independent Gaussian random variables. If X and Y are independent Gaussians, then $X, X+Y$ is a multivariate Gausssian, so (recursively) $W(t_1) - W(0), W(t_2) - W(0), \ldots W(t_k) - W(0)$ has a multivariate Gaussian distribution. Since $W(0) = 0$, the Gaussian distribution property follows. Since $t_1, \ldots t_k$ were arbitrary, as was k, all the finite-dimensional distributions are Gaussian. □

Just as the distribution of a Gaussian random variable is determined by its mean and covariance, the distribution of a Gaussian process is determined by its mean over time, $\mathbf{E}[X(t)]$, and its covariance function, $\mathrm{cov}(X(s), X(t))$. (You might find it instructive to prove this *without* looking at Lemma 13.1 in Kallenberg.) Clearly, $\mathbf{E}[W(t)] = 0$, and, taking $s \leq t$ without loss of generality,

$$\begin{aligned}
\mathrm{cov}(W(s), W(t)) &= \mathbf{E}[W(s)W(t)] - \mathbf{E}[W(s)]\mathbf{E}[W(t)] && (17.10)\\
&= \mathbf{E}[(W(t) - W(s) + W(s))W(s)] && (17.11)\\
&= \mathbf{E}[(W(t) - W(s))W(s)] + \mathbf{E}[W(s)W(s)] && (17.12)\\
&= \mathbf{E}[W(t) - W(s)]\mathbf{E}[W(s)] + s && (17.13)\\
&= s && (17.14)
\end{aligned}$$

17.3 Wiener Measure; Most Continuous Curves Are Not Differentiable

We can regard the Wiener process as establishing a measure on the space $\mathbf{C}(\mathbb{R}^+)$ of continuous real-valued functions; this is one of the considerations which led Wiener to it (Wiener, 1958)[4]. This will be important when we want to do statistical inference for stochastic processes. All Bayesian methods, and most frequentist ones, will require us to have a likelihood for the model θ given data x, $f_\theta(x)$, but likelihoods are really Radon-Nikodym derivatives, $f_\theta(x) = \frac{d\nu_\theta}{d\mu}(x)$ with respect to some reference measure μ. When our sample space is $\mathbb{R}^d$, we generally use Lebesgue measure as our reference measure, since its support is the whole space, it treats all points uniformly, and it's reasonably normalizable. Wiener measure will turn out to play a similar role when our sample space is $\mathbf{C}$.

A mathematically important question, which will also turn out to matter to us very greatly when we try to set up stochastic differential equations, is whether, under this *Wiener measure*, most curves are differentiable. If, say, almost all curves were differentiable, then it would be easy to define dW/dt. Unfortunately, this is not the case; almost all curves are nowhere differentiable.

There is an easy heuristic argument to this conclusion. $W(t)$ is a Gaussian,

[4] The early chapters of this book form a wonderfully clear introduction to Wiener measure, starting from prescriptions on the measures of finite-dimensional cylinders and building from there, deriving the incremental properties we've started with as consequences.

whose variance is t. If we look at the ratio in a derivative

$$\frac{W(t+h) - W(t)}{(t+h) - t}$$

the numerator has variance h and the denominator is the constant h, so the ratio has variance $1/h$, which goes to infinity as $h \rightarrow 0$. In other words, as we look at the curve of $W(t)$ on smaller and smaller scales, it becomes more and more erratic, and the slope finally blows up into a completely unpredictable quantity. This is basically the shape of the more rigorous argument as well.

Theorem 182 *With probability 1, $W(t)$ is nowhere-differentiable.*

Proof: Assume, by way of contradiction, that $W(t)$ is differentiable at t_0. Then

$$\lim_{t \rightarrow t_0} \frac{W(t,\omega) - W(t_0,\omega)}{t - t_0} \tag{17.15}$$

must exist, for some set of ω of positive measure. Call its supposed value $W'(t_0,\omega)$. That is, for every $\epsilon > 0$, we must have some δ such that $t_0 - \delta \leq t \leq t_0 + \delta$ implies

$$\left| \frac{W(t,\omega) - W(t_0,\omega)}{t - t_0} - W'(t_0,\omega) \right| \leq \epsilon \tag{17.16}$$

Without loss of generality, take $t > t_0$. Then $W(t,\omega) - W(t_0,\omega)$ is independent of $W(t_0,\omega)$ and has a Gaussian distribution with mean zero and variance $t - t_0$. Therefore the differential ratio is $\mathcal{N}(0, \frac{1}{t-t_0})$. The quantity inside the absolute value sign in Eq. 17.16 is thus Gaussian with distribution $\mathcal{N}(-W'(t_0), \frac{1}{t-t_0})$. The probability that it exceeds any ϵ is therefore always positive, and in fact can be made arbitrarily large by taking t sufficiently close to t_0. Hence, with probability 1, there is no point of differentiability. □

Continuous curves which are nowhere differentiable are odd-looking beasts, but we've just established that such "pathological" cases are in fact typical, and non-pathological ones vanishingly rare in $\mathbf{C}$. What's worse, in the functional central limit theorem (174), we obtained W as the limit of piecewise constant, and so piecewise differentiable, random functions. We could even have linearly interpolated between the points of the random walk, and those random functions would also have converged in distribution on W. The continuous, almost-everywhere-differentiable curves form a subset of $\mathbf{C}$, and now we have a sequence of measures which give them probability 1, converging on Wiener measure, which gives them probability 0. This sounds like trouble, especially if we want to use Wiener measure as a reference measure in likelihoods, because it sounds like lots of interesting measures, which *do* produce differentiable curves, will not be absolutely continuous...

The trick here is to consider carefully our σ-algebra. Wiener measure is a probability measure on $\mathbf{R}^{\mathbf{R}^+}, \mathcal{B}^{\mathbf{R}^+} \cap \mathbf{C}(\mathbb{R}^+)$. The differentiability of a function in the vicinity of a point depends on its value at uncountably many coordinates. Hence (Exercise 1.1) it is not a member of the σ-field.

Chapter 18

Stochastic Integrals with the Wiener Process

Section 18.1 addresses an issue which came up in the last lecture, namely the martingale characterization of the Wiener process.

Section 18.2 gives a heuristic introduction to stochastic integrals, via Euler's method for approximating ordinary integrals.

Section 18.3 gives a rigorous construction for the integral of a function with respect to a Wiener process.

18.1 Martingale Characterization of the Wiener Process

Last time in lecture, I mentioned (without remembering much of the details) that there is a way of characterizing the Wiener process in terms of some martingale properties. Here it is.

Theorem 183 *If $M(t)$ is a continuous martingale, and $M^2(t) - t$ is also a martingale, then $M(t)$ is a Wiener process.*

There are some very clean proofs of this theorem[1] — but they require us to use stochastic calculus! Doob (1953, pp. 384ff) gives a proof which does not, however. The details of his proof are messy, but the basic idea is to get the central limit theorem to apply, using the martingale property of $M^2(t) - t$ to get the variance to grow linearly with time and to get independent increments, and then seeing that between any two times t_1 and t_2, we can fit arbitrarily many little increments so we can use the CLT.

We will return to this result as an illustration of the stochastic calculus.

[1] See especially Ethier and Kurtz (1986, Theorem 5.2.12, p. 290).

18.2 A Heuristic Introduction to Stochastic Integrals

Euler's method is perhaps the most basic method for numerically approximating integrals. If we want to evaluate $I(x) \equiv \int_a^b x(t)dt$, then we pick n intervals of time, with boundaries $a = t_0 < t_1 < \ldots t_n = b$, and set

$$I_n(x) = \sum_{i=1}^{n} x\left(t_{i-1}\right)\left(t_i - t_{i-1}\right)$$

Then $I_n(x) \rightarrow I(x)$, if x is well-behaved and the length of the largest interval $\rightarrow 0$. If we want to evaluate $\int_{t=a}^{t=b} x(t)dw$, where w is another function of t, the natural thing to do is to get the derivative of w, w', replace the integrand by $x(t)w'(t)$, and perform the integral with respect to t. The approximating sums are then

$$\sum_{i=1}^{n} x\left(t_{i-1}\right) w'\left(t_{i-1}\right)\left(t_i - t_{i-1}\right) \tag{18.1}$$

Alternately, we could, if $w(t)$ is nice enough, approximate the integral by

$$\sum_{i=1}^{n} x\left(t_{i-1}\right)\left(w\left(t_i\right) - w\left(t_{i-1}\right)\right) \tag{18.2}$$

(You may be more familiar with using Euler's method to solve ODEs, $dx/dt = f(x)$. Then one generally picks a Δt, and iterates

$$x(t + \Delta t) = x(t) + f(x)\Delta t \tag{18.3}$$

from the initial condition $x(t_0) = x_0$, and uses linear interpolation to get a continuous, almost-everywhere-differentiable curve. Remarkably enough, this converges on the actual solution as Δt shrinks (Arnol'd, 1973).)

Let's try to carry all this over to random functions of time $X(t)$ and $W(t)$. The integral $\int X(t)dt$ is generally not a problem — we just find a version of X with measurable sample paths (Section 8.2). $\int X(t)dW$ is also comprehensible if dW/dt exists (almost surely). Unfortunately, we've seen that this is not the case for the Wiener process, which (as you can tell from the W) is what we'd really like to use here. So we can't approximate the integral with a sum like Eq. 18.1. But there's nothing preventing us from using one like Eq. 18.2, since that only demands increments of W. So what we would like to say is that

$$\int_{t=a}^{t=b} X(t)dW \equiv \lim_{n\rightarrow\infty} \sum_{i=1}^{n} X\left(t_{i-1}\right)\left(W\left(t_i\right) - W\left(t_{i-1}\right)\right) \tag{18.4}$$

This is a crude-but-workable approach to numerically evaluating stochastic integrals, and apparently how the first stochastic integrals were defined, back in the 1920s. Notice that it is going to make the integral a *random variable*, i.e., a measurable function of ω. Notice also that I haven't said anything yet which should lead you to believe that the limit on the right-hand side exists, in any sense, or that it is independent of the choice of partitions $a = t_0 < t_1 < \ldots t_n\ b$. The next section will attempt to rectify this.

(When it comes to the SDE $dX = f(X)dt + g(X)dW$, the counterpart of Eq. 18.3 is

$$X(t + \Delta t) = X(t) + f(X(t))\Delta t + g(X(t))\Delta W \tag{18.5}$$

where $\Delta W = W(t+\Delta t)-W(t)$, and again we use linear interpolation in between the points, starting from $X(t_0) = x_0$.)

18.3 Integrals with Respect to the Wiener Process

The drill by now should be familiar: first we define integrals of step functions, then we approximate more general classes of functions by these elementary functions. We need some preliminary technicalities.

Definition 184 (Progressive Process) *A continuous-parameter stochastic process X adapted to a filtration $\mathcal{G}$ is* progressively measurable *or* progressive *when $X(s, \omega)$, $0 \leq s \leq t$, is always measurable with respect to $\mathcal{B}_t \times \mathcal{G}_t$, where $\mathcal{B}_t$ is the Borel σ-field on $[0, t]$.*

If X has continuous sample paths, for instance, then it is progressive.

Definition 185 (Non-anticipating filtrations, processes) *Let W be a standard Wiener process, $\{\mathcal{F}_t\}$ the right-continuous completion of the natural filtration of W, and $\mathcal{G}$ any σ-field independent of $\{\mathcal{F}_t\}$. Then the* non-anticipating filtrations *are the ones of the form $\sigma(\mathcal{F}_t \cap \mathcal{G})$, $0 \leq t < \infty$. A stochastic process X is* non-anticipating *if it is adapted to some non-anticipating filtration.*

The idea of the definition is that if X is non-anticipating, we allow it to depend on the history of W, and possibly some extra, independent random stuff, but none of that extra information is of any use in predicting the future development of W, since it's independent.

Definition 186 (Elementary non-anticipating process) *A progressive, non-anticipating process X is* elementary *if there exist an increasing sequence of times t_i, starting at zero and tending to infinity, such that $X(t) = X(t_n)$ if $t \in [t_n, t_{n+1})$, i.e., if X is a step-function of time.*

Definition 187 (Square-integrable in the mean) *A random process X is* square-integrable from a to b *if $\mathbf{E}\left[\int_a^b X^2(t)dt\right]$ is finite.*

Notice that if X is bounded on $[a, b]$, in the sense that $|X(t)| \leq M$ with probability 1 for all $a \leq t \leq b$, then X is square-integrable from a to b.

Definition 188 (Itô integral of an elementary process) *If X is an elementary, progressive, non-anticipative process, square-integrable from a to b, then its* Itô integral from a to b *is*

$$\int_a^b X(t)dW \quad \equiv \quad \sum_{i \geq 0} X(t_i)\left(W\left(t_{i+1}\right) - W\left(t_i\right)\right) \tag{18.6}$$

where the t_i are as in Definition 186, truncated below by a and above by b.

Notice that this is basically a Riemann-Stieltjes integral. It's a random variable, but we don't have to worry about the existence of a limit. Now we set about approximating more general sorts of processes by elementary processes.

Lemma 189 *Suppose X is progressive, non-anticipative, bounded on $[a, b]$, and has continuous sample paths. Then there exist bounded elementary processes X_n, Itô-integrable on $[a, b]$, such that*

$$\lim_{n\to\infty} \mathbf{E}\left[\int_a^b (X - X_n)^2 dt\right] \quad = \quad 0 \tag{18.7}$$

Proof: Set

$$X_n(t) \equiv \sum_{i=0}^{\infty} X(t_i)\mathbf{1}_{[i/2^n, (i+1)/2^n)}(t) \tag{18.8}$$

This is clearly elementary, bounded and square-integrable on $[a, b]$. Moreover, for fixed ω, $\int_a^b (X(t,\omega) - X_n(t,\omega))^2 dt \to 0$, since $X(t,\omega)$ is continuous. So the expectation of the time-integral goes to zero by bounded convergence. □

Lemma 190 *Suppose X is progressive, non-anticipative, and bounded on $[a, b]$. Then there exist progressive, non-anticipative processes X_n which are bounded and continuous on $[a, b]$ such that*

$$\lim_{n\to\infty} \mathbf{E}\left[\int_a^b (X - X_n)^2 dt\right] \quad = \quad 0 \tag{18.9}$$

Proof: Let M be the bound on the absolute value of X. For each n, pick a probability density $f_n(t)$ on $\mathbb{R}$ whose support is confined to the interval $(-1/n, 0)$. Set

$$X_n(t) \equiv \int_0^t f_n(s-t)X(s)ds \tag{18.10}$$

$X_n(t)$ is then a sort of moving average of X, over the interval $(t-1/n, t)$. Clearly, it's continuous, bounded, progressively measurable, and non-anticipative. Moreover, for each ω,

$$\lim_{n\to\infty} \int_a^b (X_n(t,\omega) - X(t,\omega))^2 dt = 0 \tag{18.11}$$

because of the way we've set up f_n and X_n. By bounded convergence, Eq. 18.9 follows. □

Lemma 191 *Suppose X is progressive, non-anticipative, and square-integrable on $[a,b]$. Then there exist a sequence of random processes X_n which are progressive, non-anticipative and bounded on $[a,b]$, such that*

$$\lim_{n\rightarrow\infty} \mathbf{E}\left[\int_a^b (X - X_n)^2 dt\right] = 0 \tag{18.12}$$

PROOF: Set $X_n(t) = (-n \vee X(t)) \wedge n$. This has the desired properties, and the result follows from dominated (not bounded!) convergence. □

Lemma 192 *Suppose X is progressive, non-anticipative, and square-integrable on $[a,b]$. Then there exist a sequence of bounded elementary processes X_n such that*

$$\lim_{n\rightarrow\infty} \mathbf{E}\left[\int_a^b (X - X_n)^2 dt\right] = 0 \tag{18.13}$$

PROOF: Combine the preceding three lemmas. □

This lemma gets its force from the following result.

Lemma 193 *Suppose X is as in Definition 188, and in addition bounded on $[a,b]$. Then*

$$\mathbf{E}\left[\left(\int_a^b X(t)dW\right)^2\right] = \mathbf{E}\left[\int_a^b X^2(t)dt\right] \tag{18.14}$$

PROOF: Set $\Delta W_i = W(t_{i+1}) - W(t_i)$. Notice that ΔW_j is independent of $X(t_i)X(t_j)\Delta W_i$ if $i < j$, because of the non-anticipation properties of X. On the other hand, $\mathbf{E}\left[(\Delta W_i)^2\right] = t_{i+1} - t_i$, by the linear variance of the increments of W. So

$$\mathbf{E}\left[X(t_i)X(t_j)\Delta W_j \Delta W_i\right] = \mathbf{E}\left[X^2(t_i)\right](t_{i+1} - t_i)\delta_{ij} \tag{18.15}$$

Substituting Eq. 18.6 into the left-hand side of Eq. 18.14,

$$\begin{aligned}
\mathbf{E}\left[\left(\int_a^b X(t)dW\right)^2\right] &= \mathbf{E}\left[\sum_{i,j} X(t_i)X(t_j)\Delta W_j \Delta W_i\right] && (18.16)\\
&= \sum_{i,j} \mathbf{E}\left[X(t_i)X(t_j)\Delta W_j \Delta W_i\right] && (18.17)\\
&= \sum_i \mathbf{E}\left[X^2(t_i)\right](t_{i+1} - t_i) && (18.18)\\
&= \mathbf{E}\left[\sum_i X^2(t_i)(t_{i+1} - t_i)\right] && (18.19)\\
&= \mathbf{E}\left[\int_a^b X^2(t)dt\right] && (18.20)
\end{aligned}$$

where the last step uses the fact that X^2 must also be elementary. □

Theorem 194 *Let X and X_n be as in Lemma 192. Then the sequence $I_n(X) \equiv$*

$$\int_a^b X_n(t)dW \tag{18.21}$$

has a limit in L_2. Moreover, this limit is the same for any such approximating sequence X_n.

Proof: For each X_n, $I_n(X(\omega))$ is an L_2 function of ω, by the fact that X_n is square-integrable and Lemma 193. Now, the X_n are converging on X, in the sense that

$$\mathbf{E}\left[\int_a^b (X(t) - X_n(t))^2 dt\right] \rightarrow 0$$

i.e., in an L_2 sense, but on the interval $[a, b]$ of the real line, and not on Ω. Nonetheless, because this is a convergent sequence, it must also be a Cauchy sequence, so, for every $\epsilon > 0$, there exists an n such that

$$\mathbf{E}\left[\int_a^b (X_{n+k}(t) - X_n(t))^2 dt\right] < \epsilon$$

for every positive k. Since X_n and X_{n+k} are both elementary processes, their difference is also elementary, and we can apply Lemma 193 to it. That is, for every $\epsilon > 0$, there is an n such that

$$\mathbf{E}\left[\left(\int_a^b (X_{n+k}(t) - X_n(t))dW\right)^2\right] < \epsilon$$

for all k. But this is to say that $I_n(X)$ is a Cauchy sequence in $L_2(\Omega)$, therefore it has a limit, which is also in $L_2(\Omega)$. If Y_n is another sequence of approximations of X by elementary processes, it is also a Cauchy sequence, and so must have the same limit. □

Definition 195 *Let X be progressive, non-anticipative and square-integrable on $[a, b]$. Then its* Itô integral *is*

$$\int_a^b X(t)dW \equiv \lim_n \int_a^b X_n(t)dW \tag{18.22}$$

taking the limit in L_2, with X_n as in Lemma 192. We will say that X is Itô-integrable on $[a, b]$.

Corollary 196 (The Itô isometry) *Eq. 18.14 holds for all Itô-integrable X.*

Proof: Obvious from the approximation by elementary processes and Lemma 193.

18.4 Exercises

Exercise 18.1 (Basic Properties of the Itô Integral) *Prove the following, first for elementary Itô-integrable processes, and then in general.*

a

$$\int_a^c X(t)dW = \int_a^b X(t)dW + \int_b^c X(t)dW$$

almost surely.

b If c is any real constant, then, almost surely,

$$\int_a^b (cX(t) + Y(t))dW = c\int_a^b XdW + \int_a^b Y(t)dW$$

Exercise 18.2 (Martingale Properties of the Itô Integral) *Suppose X is Itô-integrable on $[a, b]$. Show that*

$$I_x(t) \equiv \int_a^t X(s)dW$$

$a \leq t \leq b$, is a martingale. What is $E[I_x(t)]$?

Exercise 18.3 (Continuity of the Itô Integral) *Show that $I_x(t)$ has continuous sample paths.*

Chapter 19

Stochastic Differential Equations

Section 19.1 gives two easy examples of Itô integrals. The second one shows that there's something funny about change of variables, or if you like about the chain rule.

Section 19.2 explains how to do change of variables in a stochastic integral, also known as "Itô's formula".

Section 19.3 defines stochastic differential equations.

Section 19.4 sets up a more realistic model of Brownian motion, leading to an SDE called the Langevin equation, and solves it to get Ornstein-Uhlenbeck processes.

19.1 Some Easy Stochastic Integrals, with a Moral

19.1.1 $\int dW$

Let's start with the easiest possible stochastic integral:

$$\int_a^b dW \tag{19.1}$$

i.e., the Itô integral of the function which is always 1, $\mathbf{1}_{\mathbb{R}^+}(t)$. If this is any kind of integral at all, it should be W — more exactly, because this is a definite integral, we want $\int_a^b dW = W(b) - W(a)$. Mercifully, this works. Pick any set of time-points t_i we like, and treat 1 as an elementary function with those times as its break-points. Then, using our definition of the Itô integral for elementary functions,

$$\int_a^b dW = \sum_{t_i} W(t_{i+1}) - W(t_i) \tag{19.2}$$

$$= W(b) - W(a) \tag{19.3}$$

as required. (This would be a good time to convince yourself that adding extra break-points to an elementary function doesn't change its integral.)

19.1.2 $\int W dW$

Tradition dictates that the next example be $\int W dW$. First, we should convince ourselves that $W(t)$ is Itô-integrable: it's clearly measurable and non-anticipative, but is it square-integrable? Yes; by Fubini's theorem,

$$\mathbf{E}\left[\int_0^t W^2(s)ds\right] = \int_0^t \mathbf{E}\left[W^2(s)\right] ds \tag{19.4}$$

$$= \int_0^t s ds \tag{19.5}$$

which is clearly finite on finite intervals $[0,t]$. So, this integral should exist. Now, if the ordinary rules for change of variables held — equivalent, if the chain-rule worked the usual way — we could say that $W dW = \frac{1}{2}d(W^2)$, so $\int W dW = \frac{1}{2}\int dW^2$, and we'd expect $\int_0^t W dW = \frac{1}{2}W^2(t)$. But, alas, this can't be right. To see why, take the expectation: it'd be $\frac{1}{2}t$. But we know that it has to be zero, and it has to be a martingale in t, and this is neither. A bone-head would try to fix this by subtracting off the non-martingale part, i.e., a bone-head would guess that $\int_0^t W dW = \frac{1}{2}W^2(t) - t/2$. Annoyingly, in this case the bone-head is correct. The demonstration is fundamentally straightforward, if somewhat long-winded.

To begin, we need to approximate W by elementary functions. For each n, let $t_i = i\frac{t}{2^n}$, $0 \leq i \leq 2^n - 1$. Set $\phi_n(t) = \sum_{i=0}^{2^n-1} W(t_i)\mathbf{1}_{[t_i,t_{i+1})}$. Let's check that this converges to $W(t)$ as $n \to \infty$:

$$\begin{aligned}
\mathbf{E}\left[\int_0^t (\phi_n(s) - W(s))^2 ds\right] &= \mathbf{E}\left[\sum_{i=0}^{2^n-1} \int_{t_i}^{t_{i+1}} (B(t_i) - B(s))^2 ds\right] && (19.6)\\
&= \sum_{i=0}^{2^n-1} \mathbf{E}\left[\int_{t_i}^{t_{i+1}} (B(t_i) - B(s))^2 ds\right] && (19.7)\\
&= \sum_{i=0}^{2^n-1} \int_{t_i}^{t_{i+1}} \mathbf{E}\left[(B(t_i) - B(s))^2\right] ds && (19.8)\\
&= \sum_{i=0}^{2^n-1} \int_{t_i}^{t_{i+1}} (s - t_i) ds && (19.9)\\
&= \sum_{i=0}^{2^n-1} \int_0^{2^{-n}} s ds && (19.10)\\
&= \sum_{i=0}^{2^n-1} \left[\frac{t^2}{2}\right]_0^{2^{-n}} && (19.11)\\
&= \sum_{i=0}^{2^n-1} 2^{-2n-1} && (19.12)\\
&= 2^{-n-1} && (19.13)
\end{aligned}$$

which $\to 0$ as $n \to \infty$. Hence

$$\begin{aligned}
\int_0^t W(s) dW &= \lim_n \int_0^t \phi_n(s) dW && (19.14)\\
&= \lim_n \sum_{i=0}^{2^n-1} W(t_i)(W(t_{i+1}) - W(t_i)) && (19.15)\\
&= \lim_n \sum_{i=0}^{2^n-1} W(t_i) \Delta W(t_i) && (19.16)
\end{aligned}$$

where $\Delta W(t_i) \equiv W(t_{i+1}) - W(t_i)$, because I'm getting tired of writing both subscripts. Define $\Delta W^2(t_i)$ similarly. Since $W(0) = 0 = W^2(0)$, we have that

$$\begin{aligned}
W(t) &= \sum_i \Delta W(t_i) && (19.17)\\
W^2(t) &= \sum_i \Delta W^2(t_i) && (19.18)
\end{aligned}$$

Now let's re-write ΔW^2 in such a way that we get a $W \Delta W$ term, which is what we want to evaluate our integral.

$$\begin{aligned}
\Delta W^2(t_i) &= W^2(t_{i+1}) - W^2(t_i) && (19.19)\\
&= (\Delta W(t_i) + W(t_i))^2 - W^2(t_i) && (19.20)\\
&= (\Delta W(t_i))^2 + 2W(t_i)\Delta W(t_i) + W^2(t_i) - W^2(t_i) && (19.21)\\
&= (\Delta W(t_i))^2 + 2W(t_i)\Delta W(t_i) && (19.22)
\end{aligned}$$

This looks promising, because it's got $W\Delta W$ in it. Plugging in to Eq. 19.18,

$$\begin{aligned}
W^2(t) &= \sum_i (\Delta W(t_i))^2 + 2W(t_i)\Delta W(t_i) && (19.23)\\
\sum_i W(t_i)\Delta W(t_i) &= \frac{1}{2}W^2(t) - \frac{1}{2}\sum_i (\Delta W(t_i))^2 && (19.24)
\end{aligned}$$

Now, it is possible to show (Exercise 19.1) that

$$\lim_n \sum_{i=0}^{2^n-1} (\Delta W(t_i))^2 = t \qquad (19.25)$$

in L^2, so we have that

$$\begin{aligned}
\int_0^t W(s)dW &= \lim_n \sum_{i=0}^{2^n-1} W(t_i)\Delta W(t_i) && (19.26)\\
&= \frac{1}{2}W^2(t) - \lim_n \sum_{i=0}^{2^n-1} (\Delta W(t_i))^2 && (19.27)\\
&= \frac{1}{2}W^2(t) - \frac{t}{2} && (19.28)
\end{aligned}$$

as required.

Clearly, something weird is going on here, and it would be good to get to the bottom of this. At the very least, we'd like to be able to use change of variables, so that we can find functions of stochastic integrals.

19.2 Itô's Formula

Integrating $\int WdW$ has taught us two things: first, we want to avoid evaluating Itô integrals directly from the definition; and, second, there's something funny about change of variables in Itô integrals. A central result of stochastic calculus, known as *Itô's formula*, gets us around both difficulties, by showing how to write functions of stochastic integrals as, themselves, stochastic integrals.

Definition 197 (Itô Process) *If A is a non-anticipating measurable process, B is Itô-integrable, and X_0 is an L_2 random variable independent of W, then $X(t) = X_0 + \int_0^t A(s)ds + \int_0^t B(s)dW$ is an* Itô process*. This is equivalently written $dX = Adt + BdW$.*

Lemma 198 *Every Itô process is non-anticipating.*

Proof: Clearly, the non-anticipating processes are closed under linear operations, so it's enough to show that the three components of any Itô process are non-anticipating. But a process which is always equal to $X_0 \perp\!\!\!\perp W(t)$ is clearly non-anticipating. Similarly, since $A(t)$ is non-anticipating, $\int A(s)ds$ is too: its natural filtration is smaller than that of A, so it cannot provide more information about $W(t)$, and A is, by assumption, non-anticipating. Finally, Itô integrals are always non-anticipating, so $\int B(s)dW$ is non-anticipating. □

Theorem 199 (Itô's Formula (One-Dimension)) *Suppose X is an Itô process with $dX = Adt+BdW$. Let $f(t,x) : \mathbb{R}^+ \times \mathbb{R} \mapsto \mathbb{R}$ be a function with continuous partial time derivative $\frac{\partial f}{\partial t}$, and continuous second partial space derivative, $\frac{\partial^2 f}{\partial x^2}$. Then $F(t) = f(t, X(t))$ is an Itô process, and*

$$dF = \frac{\partial f}{\partial t}(t, X(t))dt + \frac{\partial f}{\partial x}(t, X(t))dX + \frac{1}{2}B^2(t)\frac{\partial^2 f}{dx^2}(t, X(t))dt \tag{19.29}$$

That is,

$$F(t) - F(0) = \tag{19.30}$$
$$\int_0^t \left[\frac{\partial f}{\partial t}(s, X(s)) + A(s)\frac{\partial f}{\partial x}(s, X(s)) + \frac{1}{2}B^2(s)\frac{\partial^2 f}{\partial x^2}(s, X(s))\right] dt + \int_0^t B(s)\frac{\partial f}{\partial x}(s, X(s))dW$$

Proof: I will suppose first of all that f, and its partial derivatives appearing in Eq. 19.29, are all bounded. (You can show that the general case of C^2 functions can be uniformly approximated by functions with bounded derivatives.) I will further suppose that A and B are elementary processes, since in the last chapter we saw how to use them to approximate general Itô-integrable functions. (If you are worried about the interaction of all these approximations and simplifications, I commend your caution, and suggest you step through the proof in the general case.)

For each n, let $t_i = i\frac{t}{2^n}$, as in the last section. Define $\Delta t_i \equiv t_{i+1} - t_i$, $\Delta X(t_i) = X(t_{i+1}) - X(t_i)$, etc. Thus

$$F(t) = f(t, X(t)) \quad = \quad f(0, X(0)) + \sum_{i=0}^{2^n-1} \Delta f(t_i, X(t_i)) \tag{19.31}$$

Now we'll approximate the increments of F by a Taylor expansion:

$$\begin{aligned}
F(t) &= f(0,X(0)) + \sum_{i=0}^{2^n-1} \frac{\partial f}{\partial t}\Delta t_i \qquad (19.32)\\
&+ \sum_{i=0}^{2^n-1} \frac{\partial f}{\partial x}\Delta X(t_i)\\
&+ \frac{1}{2}\sum_{i=0}^{2^n-1} \frac{\partial^2 f}{\partial t^2}(\Delta t_i)^2\\
&+ \sum_{i=0}^{2^n-1} \frac{\partial^2 f}{\partial t \partial x}\Delta t_i \Delta X(t_i)\\
&+ \frac{1}{2}\sum_{i=0}^{2^n-1} \frac{\partial^2 f}{\partial x^2}(\Delta X(t_i))^2\\
&+ \sum_{i=0}^{2^n-1} R_i
\end{aligned}$$

Because the derivatives are bounded, all the remainder terms R_i are $o((\Delta t_i)^2 + (\Delta X(t_i))^2)$. We will come back to showing that the remainders are harmless, but for now let's concentrate on the leading-order components of the Taylor expansion.

First, as $n \rightarrow \infty$,

$$\begin{aligned}
\sum_{i=0}^{2^n-1} \frac{\partial f}{\partial t}\Delta t_i &\rightarrow \int_0^t \frac{\partial f}{\partial t} ds \qquad (19.33)\\
\sum_{i=0}^{2^n-1} \frac{\partial f}{\partial x}\Delta X(t_i) &\rightarrow \int_0^t \frac{\partial f}{\partial x} dX \qquad (19.34)\\
&\equiv \int_0^t \frac{\partial f}{\partial x} A(s)dt + \int_0^t \frac{\partial f}{\partial x} B(s)dW \qquad (19.35)
\end{aligned}$$

[You can use the definition in the last line to build up a theory of stochastic integrals with respect to arbitrary Itô processes, not just Wiener processes.]

$$\sum_{i=0}^{2^n-1} \frac{\partial^2 f}{\partial t^2}(\Delta t_i)^2 \rightarrow 0 \int_0^t \frac{\partial^2 f}{\partial t^2} ds = 0 \qquad (19.36)$$

Next, since A and B are (by assumption) elementary,

$$\begin{aligned}\sum_{i=0}^{2^n-1} \frac{\partial^2 f}{\partial x^2}(\Delta X(t_i))^2 &= \sum_{i=0}^{2^n-1} \frac{\partial^2 f}{\partial x^2} A^2(t_i)\left(\Delta t_i\right)^2 \\ &+2\sum_{i=0}^{2^n-1} \frac{\partial^2 f}{\partial x^2} A(t_i)B(t_i)\Delta t_i \Delta W(t_i) \\ &+\sum_{i=0}^{2^n-1} \frac{\partial^2 f}{\partial x^2} B^2(t_i)(\Delta W(t_i))^2\end{aligned} \tag{19.37}$$

The first term on the right-hand side, in $(\Delta t)^2$, goes to zero as n increases. Since A is square-integrable and $\frac{\partial^2 f}{\partial x^2}$ is bounded, $\sum \frac{\partial^2 f}{\partial x^2} A^2(t_i)\Delta t_i$ converges to a finite value as $\Delta t \rightarrow 0$, so multiplying by another factor Δt, as $n \rightarrow \infty$, gives zero. (This is the same argument as the one for Eq. 19.36.) Similarly for the second term, in $\Delta t \Delta X$:

$$\lim_n \sum_{i=0}^{2^n-1} \frac{\partial^2 f}{\partial x^2} A(t_i)B(t_i)\Delta t_i \Delta W(t_i) = \lim_n \frac{t}{2^n}\int_0^t \frac{\partial^2 f}{\partial x^2} A(s)B(s)dW \tag{19.38}$$

because A and B are elementary and the partial derivative is bounded. Now apply the Itô isometry:

$$\mathbf{E}\left[\left(\frac{t}{2^n}\int_0^t \frac{\partial^2 f}{\partial x^2} A(s)B(s)dW\right)^2\right] = \frac{t^2}{2^{2n}}\mathbf{E}\left[\int_0^t \left(\frac{\partial^2 f}{\partial x^2}\right)^2 A^2(s)B^2(s)ds\right]$$

The time-integral on the right-hand side is finite, since A and B are square-integrable and the partial derivative is bounded, and so, as n grows, both sides go to zero. But this means that, in L_2,

$$\sum_{i=0}^{2^n-1} \frac{\partial^2 f}{\partial x^2} A(t_i)B(t_i)\Delta t_i \Delta W(t_i) \quad \rightarrow \quad 0 \tag{19.39}$$

The third term, in $(\Delta X)^2$, does *not* vanish, but rather converges in L_2 to a time integral:

$$\sum_{i=0}^{2^n-1} \frac{\partial^2 f}{\partial x^2} B^2(t_i)(\Delta W(t_i))^2 \quad \rightarrow \quad \int_0^t \frac{\partial^2 f}{\partial x^2} B^2(s)ds \tag{19.40}$$

You will prove this in part b of Exercise 19.1.

The mixed partial derivative term has no counterpart in Itô's formula, so it needs to go away.

$$\sum_{i=0}^{2^n-1} \frac{\partial^2 f}{\partial t \partial x}\Delta t_i \Delta X(t_i) = \sum_{i=0}^{2^n-1} \frac{\partial^2 f}{\partial t \partial x}\left[A(t_i)(\Delta t_i)^2 + B(t_i)\Delta t_i \Delta W(t_i)\right] \tag{19.41}$$

$$\begin{aligned}\sum_{i=0}^{2^n-1} \frac{\partial^2 f}{\partial t \partial x} A(t_i) (\Delta t_i)^2 &\rightarrow 0 && (19.42)\\ \sum_{i=0}^{2^n-1} \frac{\partial^2 f}{\partial t \partial x} B(t_i) \Delta t_i \Delta W(t_i) &\rightarrow 0 && (19.43)\end{aligned}$$

where the argument for Eq. 19.43 is the same as that for Eq. 19.36, while that for Eq. 19.43 follows the pattern of Eq. 19.39.

Let us, as promised, dispose of the remainder term. Clearly,

$$\begin{aligned}(\Delta X)^2 &= A^2 (\Delta t)^2 + 2AB\Delta t \Delta W + B^2 (\Delta W)^2 && (19.44)\\ &= A^2 (\Delta t)^2 + 2AB\Delta t \Delta W + B^2 \Delta t && (19.45)\end{aligned}$$

so, from the foregoing, it is clear that this goes to zero as $\Delta t \rightarrow 0$. Hence the remainder term will vanish as n increases.

Putting everything together, we have that

$$F(t) - F(0) = \int_0^t \left[\frac{\partial f}{\partial t} + \frac{\partial f}{\partial x} A + \frac{1}{2} B^2 \frac{\partial^2 f}{\partial x^2} \right] dt + \int_0^t \frac{\partial f}{\partial x} B dW \quad (19.46)$$

exactly as required. This completes the proof, under the stated restrictions on f, A and B; approximation arguments extend the result to the general case. □

Remark 1. Our manipulations in the course of the proof are often summarized in the following multiplication rules for differentials: $dtdt = 0$, $dWdt = 0$, $dtdW = 0$, and, most important of all,

$$dWdW = dt$$

This last is of course related to the linear growth of the variance of the increments of the Wiener process.

Remark 2. Re-arranging Itô's formula a little yields

$$dF = \frac{\partial f}{\partial t} dt + \frac{\partial f}{\partial x} dX + \frac{1}{2} \frac{\partial^2 f}{\partial x^2} dt \quad (19.47)$$

The first two terms are what we expect from the ordinary rules of calculus; it's the third term which is new and strange. Notice that it disappears if $\frac{\partial^2 f}{\partial x^2} = 0$. When we come to stochastic differential equations, this will correspond to state-independent noise.

Remark 3. One implication of Itô's formula is that *Itô processes are closed under the application of C^2 mappings.*

Example 200 *The integral $\int W dW$ is now trivial. Let $X(t) = W(t)$ (by setting $A = 0$, $B = 1$ in the definition of an Itô process), and $f(t, x) = x^2/2$. Applying*

Itô's formula,

$$\begin{aligned} dF &= \frac{\partial f}{\partial t}dt + \frac{\partial f}{\partial x}dW + \frac{1}{2}\frac{\partial^2 f}{\partial x^2}dt && (19.48) \\ \frac{1}{2}dW^2 &= WdW + \frac{1}{2}dt && (19.49) \\ \frac{1}{2}\int dW^2 &= \int WdW + \frac{1}{2}\int dt && (19.50) \\ \int_0^t W(s)dW &= \frac{1}{2}W^2(t) - \frac{t}{2} && (19.51) \end{aligned}$$

All of this extends naturally to higher dimensions.

Definition 201 (Multidimensional Itô Process) *Let A by an n-dimensional vector of non-anticipating processes, B an $n \times m$ matrix of Itô-integrable processes, and W an m-dimensional Wiener process. Then*

$$\begin{aligned} X(t) &= X(0) + \int_0^t A(s)ds + \int_0^t B(s)dW && (19.52) \\ dX &= A(t)dt + B(t)dW && (19.53) \end{aligned}$$

*is an n-*dimensional Itô process*.*

Theorem 202 (Itô's Formula (Multidimensional)) *Let $X(t)$ be an n-dimensional Itô process, and let $f(t,x) : \mathbb{R}^+ \times \mathbb{R}^n \mapsto \mathbb{R}^m$ have a continuous partial time derivative and continuous second partial space derivatives. Then $F(t) = f(t, X(t))$ is an m-dimensional Itô process, whose k^{th} component F_k is given by*

$$dF_k = \frac{\partial g_k}{\partial t}dt + \frac{\partial g_k}{\partial x_i}dX_i + \frac{1}{2}\frac{\partial^2 g_k}{\partial X_i \partial X_j}dX_i dX_j \qquad (19.54)$$

summing over repeated indices, with the understanding that $dW_i dW_j = \delta_{ij}dt$, $dW_i dt = dt dW_i = dt dt = 0$.

PROOF: Entirely parallel to the one-dimensional case, only with even more algebra. □

It is also possible to define Wiener processes and stochastic integrals on arbitrary curved manifolds, but this would take us way, way too far afield.

19.2.1 Stratonovich Integrals

It is possible to make the extra term in Eq. 19.47 go away, and have stochastic differentials which work just like the ordinary ones. This corresponds to making stochastic integrals limits of sums of the form

$$\sum_i X\left(\frac{t_{i+1} + t_i}{2}\right)\Delta W(t_i)$$

rather than the Itô sums we are using,

$$\sum_i X(t_i)\Delta W(t_i)$$

That is, we could evade the Itô formula if we evaluated our test function in the middle of intervals, rather than at their beginnning. This leads to what are called *Stratonovich integrals*. However, while Stratonovich integrals give simpler change-of-variable formulas, they have many other inconveniences: they are not martingales, for instance, and the nice connections between the form of an SDE and its generator, which we will see and use in the next chapter, go away. Fortunately, every Stratonovich SDE can be converted into an Itô SDE, and vice versa, by adding or subtracting the appropriate noise term.

19.2.2 Martingale Representation

One property of the Itô integral is that it is always a square-integrable martingale. Remarkably enough, the converse is also true. In the interest of time, I omit the proof of the following theorem; there is one using only tools we've seen so far in Øksendal (1995, ch. 4).

Theorem 203 *Let $M(t)$ be a martingale, with $\mathbf{E}\left[M^2(t)\right] < \infty$ for all $t \geq 0$. Then there exists a unique process $M'(t)$, Itô-integrable for all finite positive t, such that*

$$M(t) = \mathbf{E}\left[M(0)\right] + \int_0^t M'(t)dW \ a.s. \tag{19.55}$$

19.3 Stochastic Differential Equations

Definition 204 (Stochastic Differential Equation, Solutions) *Let $a(x)$: $\mathbb{R}^n \mapsto \mathbb{R}^n$ and $b(x) : \mathbb{R}^n \mapsto \mathbb{R}^{nm}$ be measurable functions (vector and matrix valued, respectively), W an m-dimensional Wiener process, and X_0 an L_2 random variable in $\mathbb{R}^n$, independent of W. Then an $\mathbb{R}^n$-valued stochastic process X on $\mathbb{R}^+$ is a solution to the* autonomous stochastic differential equation

$$dX = a(X)dt + b(X)dW, \ X(0) = X_0 \tag{19.56}$$

when, with probability 1, it is equal to the corresponding Itô process,

$$X(t) \quad = \quad X_0 + \int_0^t a(X(s))ds + \int_0^s b(X(s))dW \ a.s. \tag{19.57}$$

The a term is called the drift*, and the b term the* diffusion.

Remark 1: A given process X can fail to be a solution either because it happens not to agree with Eq. 19.57, or, perhaps more seriously, because the integrals on the right-hand side don't even exist. This can, in particular, happen if $b(X(t))$ is anticipating. For a *fixed* choice of Wiener process, there are circumstances where otherwise reasonable SDEs have no solution, for basically this reason — the Wiener process is constructed in such a way that the class of Itô processes is impoverished. This leads to the idea of a *weak solution* to Eq. 19.56, which is a *pair* X, W such that W is a Wiener process, with respect to the appropriate filtration, and X then is given by Eq. 19.57. I will avoid weak solutions in what follows.

Remark 2: In a non-autonomous SDE, the coefficients would be explicit functions of time, $a(t,X)dt + b(t,X)dW$. The usual trick for dealing with non-autonomous n-dimensional ODEs is turn them into autonomous $n+1$-dimensional ODEs, making $x_{n+1} = t$ by decreeing that $x_{n+1}(t_0) = t_0$, $x'_{n+1} = 1$ (Arnol'd, 1973). This works for SDEs, too: add time as an extra variable with constant drift 1 and constant diffusion 0. Without loss of generality, therefore, I'll only consider autonomous SDEs.

Let's now prove the existence of unique solutions to SDEs. First, recall how we do this for ordinary differential equations. There are several approaches, most of which carry over to SDEs, but one of the most elegant is the "method of successive approximations", or "Picard's method" (Arnol'd, 1973, SS30–31)). To construct a solution to $dx/dt = f(x)$, $x(0) = x_0$, this approach uses functions $x_n(t)$, with $x_{n+1}(t) = x_0 + \int_0^t f(x_n(s)ds$, starting with $x_0(t) = x_0$. That is, there is an operator P such that $x_{n+1} = Px_n$, and x solves the ODE iff it is a fixed point of the operator. Step 1 is to show that the sequence x_n is Cauchy on finite intervals $[0,T]$. Step 2 uses the fact that the space of continuous functions is complete, with the topology of uniform convergence of compact sets — which, for $\mathbb{R}^+$, is the same as uniform convergence on finite intervals. So, x_n has a limit. Step 3 is to show that the limit point must be a fixed point of P, that is, a solution. Uniqueness is proved by showing that there cannot be more than one fixed point.

Before plunging in to the proof, we need some lemmas: an algebraic triviality, a maximal inequality for martingales, a consequent maximal inequality for Itô processes, and an inequality from real analysis about integral equations.

Lemma 205 *For any real numbers a and b, $(a+b)^2 \leq 2a^2 + 2b^2$.*

Proof: No matter what a and b are, a^2, b^2, and $(a-b)^2$ are non-negative, so

$$(a-b)^2 \geq 0 \quad (19.58)$$

$$a^2 + b^2 - 2ab \geq 0 \quad (19.59)$$

$$a^2 + b^2 \geq 2ab \quad (19.60)$$

$$2a^2 + 2b^2 \geq a^2 + 2ab + b^2 = (a+b)^2 \quad (19.61)$$

□

Definition 206 (Maximum Process) *Given a stochastic process $X(t)$, we define its* maximum process $X^*(t)$ *as* $\sup_{0 \leq s \leq t} |X(s)|$.

Remark: Example 78 was of course designed with malice aforethought.

Definition 207 *Let $\mathcal{QM}(T)$, $T > 0$, be the space of all non-anticipating processes, square-integrable on $[0, T]$, with norm $\|X\|_{\mathcal{QM}(T)} \equiv \|X^*(T)\|_2$.*

(Technically, this is only a norm on equivalence classes of processes, where the equivalence relation is "is a version of". You may make that amendment mentally as you read what follows.)

Lemma 208 *$\mathcal{QM}(T)$ is a complete normed space for each T.*

Proof: Identical to the usual proof that L_p spaces are complete.

Lemma 209 (Doob's Martingale Inequalities) *If $M(t)$ is a continuous martingale, then, for all $p \geq 1$, $t \geq 0$ and $\epsilon > 0$,*

$$\begin{aligned} \mathbb{P}\left(M^*(t) \geq \epsilon\right) &\leq \frac{\mathbf{E}\left[|M(t)|^p\right]}{\epsilon^p} && (19.62) \\ \|M^*(t)\|_p &\leq q\|M(t)\|_p && (19.63) \end{aligned}$$

where $q^{-1} + p^{-1} = 1$. In particular, for $p = q = 2$,

$$\mathbf{E}\left[\left(M^*(t)\right)^2\right] \leq 4\mathbf{E}\left[M^2(t)\right]$$

Proof: See Propositions 7.15 and 7.16 in Kallenberg (pp. 128 and 129). □

These can be thought of as versions of the Markov inequality, only for martingales. They accordingly get used *all the time*.

Lemma 210 *Let $X(t)$ be an Itô process, $X(t) = X_0 + \int_0^t A(s)ds + \int_0^t B(s)dW$. Then there exists a constant C, depending only on T, such that, for all $t \in [0, T]$,*

$$\|X\|^2_{\mathcal{QM}(t)} \leq C\left(\mathbf{E}\left[X_0^2\right] + \mathbf{E}\left[\int_0^t A^2(s) + B^2(s)ds\right]\right) \qquad (19.64)$$

Proof: Clearly,

$$\begin{aligned} X^*(t) &\leq |X_0| + \int_0^t |A(s)|ds + \sup_{0 \leq s \leq t}\left|\int_0^s B(s)dW\right| && (19.65) \\ \left(X^*(t)\right)^2 &\leq 2X_0^2 + 2\left(\int_0^t |A(s)|ds\right)^2 + 2\left(\sup_{0 \leq s \leq t}\left|\int_0^s B(s')dW\right|\right)^2 && (19.66) \end{aligned}$$

by Lemma 205. By Jensen's inequality[1],

$$\left(\int_0^t |A(s)|ds\right)^2 \quad \leq \quad t\int_0^t A^2(s)ds \tag{19.67}$$

Writing $I(t)$ for $\int_0^t B(s)dW$, and noticing that it is a martingale, we have, from Doob's inequality (Lemma 209), $\mathbf{E}\left[(I^*(t))^2\right] \leq 4\mathbf{E}\left[I^2(t)\right]$. But, from Itô's isometry (Corollary 196), $\mathbf{E}\left[I^2(t)\right] = \mathbf{E}\left[\int_0^t B^2(s)ds\right]$. Putting all the parts together, then,

$$\mathbf{E}\left[(X^*(t))^2\right] \quad \leq \quad 2\mathbf{E}\left[X_0^2\right] + 2\mathbf{E}\left[t\int_0^t A^2(s)ds + \int_0^t B^2(s)ds\right] \tag{19.68}$$

and the conclusion follows, since $t \leq T$. □

Remark: The lemma also holds for multidimensional Itô processes, and for powers greater than two (though then the Doob inequality needs to be replaced by a different one: see Rogers and Williams (2000, Ch. V, Lemma 11.5, p. 129)).

Definition 211 *Given an SDE $dX = a(X)dt + b(X)dW$ with initial condition X_0, the corresponding integral operator $P_{X_0,a,b}$ is defined for all Itô processes Y as*

$$P_{X_0,a,b}Y(t) = X_0 + \int_0^t a(Y(s))ds + \int_0^t b(Y(s))dW \tag{19.69}$$

Lemma 212 *Y is a solution of $dX = a(X)dt + b(X)dW$, $X(0) = X_0$, if and only if $P_{X_0,a,b}Y = Y$ a.s.*

Proof: Obvious from the definitions. □

Lemma 213 *If a and b are uniformly Lipschitz continuous, with constants K_a and K_B, then, for some positive D depending only on T, K_a and K_b,*

$$\|P_{X_0,a,b}X - P_{X_0,a,b}Y\|^2_{\mathcal{QM}(t)} \leq D\int_0^t \|X-Y\|_{\mathcal{QM}(s)}ds \tag{19.70}$$

Proof: Since the SDE is understood to be fixed, abbreviate $P_{X_0,a,b}$ by P. Let X and Y be any two Itô processes. We want to find the $\mathcal{QM}(t)$ norm of

[1] Remember that Lebesgue measure isn't a probability measure on $[0,t]$, but $\frac{1}{t}ds$ is a probability measure, so we can apply Jensen's inequality to that. This is where the t on the right-hand side will come from.

$PX - PY$.

$$\begin{aligned}
& |PX(t) - PY(t)| && (19.71)\\
= & \left| \int_0^t a(X(s)) - a(Y(s)) dt + \int_0^t b(X(s)) - b(Y(s)) dW \right| && \\
\leq & \int_0^t \left|a(X(s)) - a(Y(s))\right| ds + \int_0^t \left|b(X(s)) - b(Y(s))\right| dW && (19.72)\\
\leq & \int_0^t K_a \left|X(s) - Y(s)\right| ds + \int_0^t K_b \left|X(s) - Y(s)\right| dW && (19.73)\\
& \|PX - PY\|^2_{\mathcal{QM}(t)} && (19.74)\\
\leq & C(K_a^2 + K_b^2)\mathbf{E}\left[\int_0^t |X(s) - Y(s)|^2 ds\right] && \\
\leq & C(K_a^2 + K_b^2)t \int_0^t \|X - Y\|^2_{\mathcal{QM}(s)} ds && (19.75)
\end{aligned}$$

which, as $t \leq T$, completes the proof. □

Lemma 214 (Gronwall's Inequality) *If f is continuous function on $[0, T]$ such that $f(t) \leq c_1 + c_2 \int_0^t f(s)ds$, then $f(t) \leq c_1 e^{c_2 t}$.*

PROOF: See Kallenberg, Lemma 21.4, p. 415. □

Theorem 215 (Existence and Uniqueness of Solutions to SDEs in One Dimension) *Let X_0, a, b and W be as in Definition 204, and let a and b be uniformly Lipschitz continuous. Then there exists a square-integrable, non-anticipating $X(t)$ which solves $dX = a(X)dt + b(X)dW$ with initial condition X_0, and this solution is unique (almost surely).*

PROOF: I'll first prove existence, along with square-integrability, and then uniqueness. That X is non-anticipating follows from the fact that it is an Itô process (Lemma 198). For concision, abbreviate $P_{X_0,a,b}$ by P.

As with ODEs, iteratively construct approximate solutions. Fix a $T > 0$, and, for $t \in [0, T]$, set

$$\begin{aligned}
X_0(t) &= X_0 && (19.76)\\
X_{n+1}(t) &= PX_n(t) && (19.77)
\end{aligned}$$

The first step is showing that X_n is Cauchy in $\mathcal{QM}(T)$. Define $\phi_n(t) \equiv \|X_{n+1}(t) - X_n(t)\|^2_{\mathcal{QM}(t)}$. Notice that $\phi_n(t) = \|PX_n(t) - PX_{n-1}(t)\|^2_{\mathcal{QM}(t)}$, and that, for each n, $\phi_n(t)$ is non-decreasing in t (because of the supremum embedded in its definition). So, using Lemma 213,

$$\begin{align}
\phi_n(t) &\leq D\int_0^t \|X_n - X_{n-1}\|^2_{\mathcal{QM}(s)} ds \tag{19.78}\\
&\leq D\int_0^t \phi_{n-1}(s) ds \tag{19.79}\\
&\leq D\int_0^t \phi_{n-1}(t) ds \tag{19.80}\\
&= Dt\phi_{n-1}(0) \tag{19.81}\\
&\leq \frac{D^n t^n}{n!}\phi_0(t) \tag{19.82}\\
&\leq \frac{D^n t^n}{n!}\phi_0(T) \tag{19.83}
\end{align}$$

Since, for any constant c, $c^n/n! \rightarrow 0$, to get the successive approximations to be Cauchy, we just need to show that $\phi_0(T)$ is finite, using Lemma 210.

$$\begin{align}
\phi_0(T) &= \|P_{X_0,a,b,}X_0 - X_0\|^2_{\mathcal{QM}(T)} \tag{19.84}\\
&= \left\|\int_0^t a(X_0)ds + \int_0^t b(X_0)dW\right\|^2_{\mathcal{QM}(T)} \tag{19.85}\\
&\leq C\mathbf{E}\left[\int_0^T a^2(X_0) + b^2(X_0) ds\right] \tag{19.86}\\
&\leq CT\mathbf{E}\left[a^2(X_0) + b^2(X_0)\right] \tag{19.87}
\end{align}$$

Because a and b are Lipschitz, this will be finite if X_0 has a finite second moment, which, by assumption, it does. So X_n is a Cauchy sequence in $\mathcal{QM}(T)$, which is a complete space, so X_n has a limit in $\mathcal{QM}(T)$, call it X.

The next step is to show that X is a fixed point of the operator P. This is true because PX is also a limit of the sequence X_n.

$$\begin{align}
\|PX - X_{n+1}\|^2_{\mathcal{QM}(T)} &= \|PX - PX_n\|^2_{\mathcal{QM}(T)} \tag{19.88}\\
&\leq DT\|X - X_n\|^2_{\mathcal{QM}(T)} \tag{19.89}
\end{align}$$

which $\rightarrow 0$ as $n \rightarrow \infty$. So PX is the limit of X_{n+1}, which means it is the limit of X_n, and, since X is also a limit of X_n and limits are unique, $PX = X$. Thus, by Lemma 212, X is a solution.

To prove uniqueness, suppose that there were another solution, Y. By Lemma 212, $PY = Y$ as well. So, with Lemma 213,

$$\begin{align}
\|X - Y\|^2_{\mathcal{QM}(t)} &= \|PX - PY\|^2_{\mathcal{QM}(t)} \tag{19.90}\\
&\leq D\int_0^t \|X - Y\|^2_{\mathcal{QM}(s)} ds \tag{19.91}
\end{align}$$

So, from Gronwall's inequality (Lemma 214), we have that $\|X - Y\|_{\mathcal{QM}(t)} \leq 0$ for all t, implying that $X(t) = Y(t)$ a.s. □

Remark: For an alternative approach, based on Euler's method (rather than Picard's), see Fristedt and Gray (1997, §33.4). It has a certain appeal, but it also involves some uglier calculations. For a side-by-side comparison of the two methods, see Lasota and Mackey (1994).

Theorem 216 *Theorem 215 also holds for multi-dimensional stochastic differential equations, provided a and b are uniformly Lipschitz in the appropriate Euclidean norms.*

PROOF: Entirely parallel to the one-dimensional case, only with more algebra. □

The conditions on the coefficients can be reduced to something like "locally Lipschitz up to a stopping time", but it does not seem profitable to pursue this here. See Rogers and Williams (2000, Ch. V, Sec. 12).

19.4 Brownian Motion, the Langevin Equation, and Ornstein-Uhlenbeck Processes

The Wiener process is not a realistic model of Brownian motion, because it implies that Brownian particles do not have well-defined velocities, which is absurd. Setting up a (somewhat) more realistic model will eliminate this absurdity, and illustrate how SDEs can be used as models. I will first need to summarize classical mechanics in one paragraph.

Classical mechanics starts with Newton's laws of motion. The zeroth law, implicit in everything, is that the laws of nature are differential equations in position variables with respect to time. The first law says that they are not first-order differential equations. The second law says that they are second-order differential equations. The usual trick for higher-order differential equations is to introduce supplementary variables, so that we have a higher-dimensional system of first-order differential equations. The supplementary variable here is momentum. Thus, for particle i, with mass m_i,

$$\frac{d\vec{x}_i}{dt} = \frac{\vec{p}_i}{m_i} \tag{19.92}$$

$$\frac{d\vec{p}_i}{dt} = \frac{F(\mathbf{x},\mathbf{p},t)}{m_i} \tag{19.93}$$

constitute the laws of motion. All the physical content comes from specifying the force function $F(\mathbf{x},\mathbf{p},t)$. We will consider only autonomous systems, so we do not need to deal with forces which are explicit functions of time. Newton's third law says that total momentum is conserved, when all bodies are taken into account.

Consider a large particle of (without loss of generality) mass 1, such as a pollen grain, sitting in a still fluid at thermal equilibrium. What forces act on it? One is drag. At a molecular level, this is due to the particle colliding with the molecules (mass m) of the fluid, whose average momentum is zero. This typically results in momentum being transferred from the pollen to the fluid molecules, and the amount of momentum lost by the pollen is proportional to what it had, i.e., one term in $d\vec{p}/dt$ is $-\gamma\vec{p}$. In addition, however, there will be fluctuations, which will be due to the fact that the fluid molecules are not all at rest. In fact, because the fluid is at equilibrium, the momenta of the molecules will follow a Maxwell-Boltzmann distribution,

$$f(\vec{p}_{\mathrm{molec}}) = (2\pi m k_B T)^{-3/2} e^{-\frac{1}{2}\frac{p^2_{\mathrm{molec}}}{m k_B T}}$$

where which is a zero-mean Gaussian with variance $m k_B T$. Tracing this through, we expect that, over short time intervals in which the pollen grain nonetheless collides with a large number of molecules, there will be a random impulse (i.e., random change in momentum) which is Gaussian, but uncorrelated over shorter sub-intervals (by the functional CLT). That is, we would like to write

$$d\vec{p} = -\gamma\vec{p}dt + DIdW \tag{19.94}$$

where D is the *diffusion constant*, I is the 3×3 identity matrix, and W of course is the standard three-dimensional Wiener process. This is known as the *Langevin equation* in the physics literature, as this model was introduced by Langevin in 1907 as a correction to Einstein's 1905 model of Brownian motion. (Of course, Langevin didn't use Wiener processes and Itô integrals, which came much later, but the spirit was the same.) If you like time-series models, you might recognize this as a continuous-time version of an mean-reverting AR(1) model, which explains why it also shows up as an interest rate model in financial theory.

We can consider each component of the Langevin equation separately, because they decouple, and solve them easily with Itô's formula:

$$d(e^{\gamma t}p) = De^{\gamma t}dW \tag{19.95}$$

$$e^{\gamma t}p(t) = p_0 + D\int_0^t e^{\gamma s}dW \tag{19.96}$$

$$p(t) = p_0 e^{-\gamma t} + D\int_0^t e^{-\gamma(t-s)}dW \tag{19.97}$$

We will see in the next chapter a general method of proving that solutions of equations like 19.94 are Markov processes; for now, you can either take that on faith, or try to prove it yourself.

Assuming p_0 is itself Gaussian, with mean 0 and variance σ^2, then (using Exercise 19.2), $\vec{p}$ always has mean zero, and the covariance is

$$\mathrm{cov}\left(\vec{p}(t), \vec{p}(s)\right) = \sigma^2 e^{-\gamma(s+t)} + \frac{D^2}{2\gamma}\left(e^{-\gamma|s-t|} - e^{-\gamma(s+t)}\right) \tag{19.98}$$

If $\sigma^2 = D^2/2\gamma$, then the covariance is a function of $|s-t|$ alone, and the process is weakly stationary. Such a solution of Eq. 19.94 is known as a *stationary*

Ornstein-Uhlenbeck process. (Ornstein and Uhlenbeck provided the Wiener processes and Itô integrals.)

Weak stationarity, and the fact that the Ornstein-Uhlenbeck process is Markovian, allow us to say that the distribution $\mathcal{N}(0, D^2/2\gamma)$ is invariant. Now, if the Brownian particle began in equilibrium, we expect its energy to have a Maxwell-Boltzmann distribution, which means that its momentum has a Gaussian distribution, and the variance is (as with the fluid molecules) $k_B T$. Thus, $k_B T = D^2/2\gamma$, or $D^2 = 2\gamma k_b T$. This is an example of what the physics literature calls a *fluctuation-dissipation relation*, since one side of the equation involves the magnitude of fluctuations (the diffusion coefficient D) and the other the response to fluctuations (the frictional damping coefficient γ). Such relationships turn out to hold quite generally at or near equilibrium, and are often summarized by the saying that "systems respond to forcing just like fluctuations". (Cf. 19.97.)

Oh, and that story I told you before about Brownian particles following Wiener processes? It's something of a lie told to children, or at least to probability theorists, but see Exercise 19.5.

For more on the physical picture of Brownian motion, fluctuation-dissipation relations, and connections to more general thermodynamic processes in and out of equilibrium, see Keizer (1987).[2]

19.5 Exercises

Exercise 19.1 *Use the notation of Section 19.1 here.*

a *Show that $\sum_i (\Delta W(t_i))^2$ converges on t (in L_2) as n grows.* Hint: *Show that the terms in the sum are IID, and that their variance shrinks sufficiently fast as n grows. (You will need the fourth moment of a Gaussian distribution.)*

b *If $X(t)$ is measurable and non-anticipating, show that*

$$\lim_n \sum_{i=0}^{2^n-1} X(t_i)(\Delta W(t_i))^2 = \int_0^t X(s)ds$$

in L_2.

Exercise 19.2 *For any fixed, non-random cadlag function f on $\mathbb{R}^+$, let $I_f(t) = \int_0^t f(s)dW$.*

a *Show that* $\mathbf{E}[I_f(t)] = 0$ *for all t.*

b *Show* $\text{cov}(I_f(t), I_f(s)) = \int_0^{t \wedge s} f^2(u)du$.

[2] Be warned that he perversely writes the probability of event A conditional on event B as $\mathbb{P}(B|A)$, not $\mathbb{P}(A|B)$.

c *Show that $I_f(t)$ is a Gaussian process.*

Exercise 19.3 *Consider*

$$dX = \frac{1}{2}X dt + \sqrt{1+X^2} dW \tag{19.99}$$

a *Show that there is a unique solution for every initial value $X(0) = x_0$.*

b *It happens (you do not have to show this) that, for fixed x_0, the the solution has the form $X(t) = \phi(W(t))$, where ϕ is a C^2 function. Use Itô's formula to find the first two derivatives of ϕ, and then solve the resulting second-order ODE to get ϕ.*

c *Verify that, with the ϕ you found in the previous part, $\phi(W(t))$ solves Eq. 19.99 with initial condition $X(0) = x_0$.*

Exercise 19.4 *Let X be an Itô process given by $dX = Adt + BdW$. Use Itô's formula to prove that*

$$f(X(t)) - f(X(0)) - \int_0^t \left[A\frac{\partial f}{\partial x} + \frac{1}{2}B^2\frac{\partial^2 f}{\partial x^2}\right] dt$$

where f is an C^2 function, is a martingale.

Exercise 19.5 (Brownian Motion and the Ornstein-Uhlenbeck Process) *Consider a Brownian particle whose momentum follows a stationary Ornstein-Uhlenbeck process, in one spatial dimension (for simplicity). Assume that its initial position $x(0)$ is fixed at the origin, and then $x(t) = \int_0^t p(t)dt$. Show that as $D \rightarrow \infty$ and $D/\gamma \rightarrow 1$, the distribution of $x(t)$ converges to a standard Wiener process. Explain why this limit is a physically reasonable one.*

Chapter 20

More on Stochastic Differential Equations

> Section 20.1 shows that the solutions of SDEs are diffusions, and how to find their generators. Our previous work on Feller processes and martingale problems pays off here. Some other basic properties of solutions are sketched, too.
>
> Section 20.2 explains the "forward" and "backward" equations associated with a diffusion (or other Feller process). We get our first taste of finding invariant distributions by looking for stationary solutions of the forward equation.
>
> Section 20.3 makes sense of the idea of white noise. This topic will be continued in the next lecture, forming one of the bridges to ergodic theory.

For the rest of this lecture, whenever I say "an SDE", I mean "an SDE satisfying the requirements of the existence and uniqueness theorem", either Theorem 215 (in one dimension) or Theorem 216 (in multiple dimensions). And when I say "a solution", I mean "a strong solution". If you are really curious about what has to be changed to accommodate weak solutions, see Rogers and Williams (2000, ch. V, sec. 16–18).

20.1 Solutions of SDEs are Diffusions

Solutions of SDEs are diffusions: i.e., continuous, homogeneous strong Markov processes.

Theorem 217 *The solution of an SDE is non-anticipating, and has a version with continuous sample paths. If $X(0) = x$ is fixed, then $X(t)$ is $\mathcal{F}_t^W$-adapted.*

PROOF: Every solution is an Itô process, so it is non-anticipating by Lemma 198. The adaptation for non-random initial conditions follows similarly. (Informally: there's nothing else for it to depend on.) In the proof of the existence of solutions, each of the successive approximations is continuous, and we bound the maximum deviation over time, so the solution must be continuous too. □

Theorem 218 *Let X_x be the process solving a one-dimensional SDE with non-random initial condition $X(0) = x$. Then X_x forms a homogeneous strong Markov family.*

PROOF: By Exercise 19.4, for every C^2 function f,

$$f(X(t)) - f(X(0)) - \int_0^t \left[a(X(s)) \frac{\partial f}{\partial x}(X(s)) + \frac{1}{2} b^2(X(s)) \frac{\partial^2 f}{\partial x^2}(X(s)) \right] ds \quad (20.1)$$

is a martingale. Hence, for every x_0, there is a unique, continuous solution to the martingale problem with operator $G = a(x)\frac{\partial}{\partial x} + \frac{1}{2}b^2(x)\frac{\partial^2}{\partial x^2}$ and function class $\mathcal{D} = C^2$. Since the process is continuous, it is also cadlag. Therefore, by Theorem 137, X is a homogeneous strong Markov family, whose generator equals G on C^2. □

Similarly, for a multi-dimensional SDE, where a is a vector and b is a matrix, the generator extends[1] $a_i(x)\partial_i + \frac{1}{2}(bb^T)_{ij}(x)\partial^2_{ij}$. Notice that the coefficients are *outside* the differential operators.

Corollary 219 *Solutions of SDEs are diffusions.*

PROOF: Obvious from Theorem 218, continuity, and Definition 177. □

Remark: To see what it is like to try to prove this without using our more general approach, read pp. 103–114 in Øksendal (1995).

Theorem 220 *Solutions of SDEs are Feller processes.*

PROOF: We need to show that (i) for every $t > 0$, $X_y(t) \xrightarrow{d} X_x(t)$ as $y \rightarrow x$, and (ii) $X_x(t) \xrightarrow{P} x$ as $t \rightarrow 0$. But, since solutions are a.s. continuous, $X_x(t) \rightarrow x$ with probability 1, automatically implying convergence in probability, so (ii) is automatic.

[1] Here, and elsewhere, I am going to freely use the Einstein conventions for vector calculus: repeated indices in a term indicate that you should sum over those indices, ∂_i abbreviates $\frac{\partial}{\partial x_i}$, ∂^2_{ij} means $\frac{\partial^2}{\partial x_i \partial x_j}$, etc. Also, $\partial_t \equiv \frac{\partial}{\partial t}$.

To get (i), prove convergence in mean square (i.e. in L_2), which implies convergence in distribution.

$$\begin{aligned}
&\mathbf{E}\left[|X_x(t) - X_y(t)|^2\right] & (20.2)\\
= \quad & \mathbf{E}\left[\left|x - y + \int_0^t a(X_x(s)) - a(X_y(s))ds + \int_0^t b(X_x(s)) - b(X_y(s))dW\right|^2\right] \\
\leq \quad & |x-y|^2 + \mathbf{E}\left[\left|\int_0^t a(X_x(s)) - a(X_y(s))ds\right|^2\right] & (20.3)\\
& + \mathbf{E}\left[\left|\int_0^t b(X_x(s)) - b(X_y(s))dW\right|^2\right] \\
= \quad & |x-y|^2 + \mathbf{E}\left[\left|\int_0^t a(X_x(s)) - a(X_y(s))ds\right|^2\right] & (20.4)\\
& + \int_0^t \mathbf{E}\left[|b(X_x(s)) - b(X_y(s))|^2\right] ds \\
\leq \quad & |x-y|^2 + K\int_0^t \mathbf{E}\left[|X_x(s) - X_y(s)|^2\right] ds & (20.5)
\end{aligned}$$

for some $K \geq 0$, using the Lipschitz properties of a and b. So, by Gronwall's Inequality (Lemma 214),

$$\mathbf{E}\left[|X_x(t) - X_y(t)|^2\right] \leq |x-y|^2 e^{Kt} \qquad (20.6)$$

This clearly goes to zero as $y \rightarrow x$, so $X_y(t) \rightarrow X_x(t)$ in L_2, which implies convergence in distribution. □

Corollary 221 *For a given SDE, convergence in distribution of the initial condition implies convergence in distribution of the trajectories: if* $Y \xrightarrow{d} X_0$, *then* $X_Y \xrightarrow{d} X_{X_0}$.

Proof: For every initial condition, the generator of the semi-group is the same (Theorem 218, proof). Since the process is Feller for every initial condition (Theorem 220), and a Feller semi-group is determined by its generator (Theorem 153), the process has the same evolution operator for every initial condition. Hence, condition (ii) of Theorem 170 holds. This implies condition (iv) of that theorem, which is the stated convergence. □

20.2 Forward and Backward Equations

You will often seen probabilists, and applied stochastics people, write about "forward" and "backward" equations for Markov processes, sometimes with the eponym "Kolmogorov" attached. We have already seen a version of the "backward" equation for Markov processes, with semi-group K_t and generator G, in Theorem 125:

$$\partial_t K_t f(x) = G K_t f(x) \tag{20.7}$$

Let's unpack this a little, which will help see where the "backwards" comes from. First, remember that the operators K_t are really just conditional expectation:

$$\partial_t \mathbf{E}\left[f(X_t)|X_0 = x\right] = G\mathbf{E}\left[f(X_t)|X_0 = x\right] \tag{20.8}$$

Next, turn the expectations into integrals with respect to the transition probability kernels:

$$\partial_t \int \mu_t(x,dy) f(y) = G \int \mu_t(x,dy) f(y) \tag{20.9}$$

Finally, assume that there is some reference measure $\lambda \gg \mu_t(x,\cdot)$, for all $t \in T$ and $x \in \Xi$. Denote the correspond transition densities by $\kappa_t(x,y)$.

$$\begin{aligned}
\partial_t \int d\lambda \kappa_t(x,y) f(y) &= G \int d\lambda \kappa_t(x,y) f(y) && (20.10)\\
\int d\lambda f(y) \partial_t \kappa_t(x,y) &= \int d\lambda f(y) G\kappa_t(x,y) && (20.11)\\
\int d\lambda f(y) \left[\partial_t \kappa_t(x,y) - G\kappa_t(x,y)\right] &= 0 && (20.12)
\end{aligned}$$

Since this holds for arbitrary nice test functions f,

$$\partial_t \kappa_t(x,y) = G\kappa_t(x,y) \tag{20.13}$$

The operator G alters the way a function depends on x, the *initial* state. That is, this equation is about how the transition density κ depends on the starting point, "backwards" in time. Generally, we're in a position to know $\kappa_0(x,y) = \delta(x-y)$; what we want, rather, is $\kappa_t(x,y)$ for some positive value of t. To get this, we need the "forward" equation.

We obtain this from Lemma 122, which asserts that $GK_t = K_t G$.

$$\begin{aligned}
\partial_t \int d\lambda \kappa_t(x,y) f(y) &= K_t G f(x) && (20.14)\\
&= \int d\lambda \kappa_t(x,y) G f(y) && (20.15)
\end{aligned}$$

Notice that here, G is altering the dependence on the y coordinate, i.e. the state at time t, not the initial state at time 0. Writing the adjoint[2] operator as $G^\dagger$,

$$\begin{aligned}
\partial_t \int d\lambda \kappa_t(x,y) f(y) &= \int d\lambda G^\dagger \kappa_t(x,y) f(y) && (20.16)\\
\partial_t \kappa_t(x,y) &= G^\dagger \kappa_t(x,y) && (20.17)
\end{aligned}$$

[2] Recall that, in a vector space with an inner product, such as L_2, the adjoint of an operator A is another operator, defined through $\langle f, Ag\rangle = \langle A^\dagger f, g\rangle$. Further recall that L_2 is an inner-product space, where $\langle f, g\rangle = \mathbf{E}\left[f(X)g(X)\right]$.

N.B., $G^\dagger$ is acting on the y-dependence of the transition density, i.e., it says how the probability density is going to change *going forward from t*.

In the physics literature, this is called the Fokker-Planck equation, because Fokker and Planck (independently, so far as I know) discovered it, at least in the special case of Langevin-type equations, in 1913, about 20 years before Kolmogorov's work on Markov processes. Notice that, writing ν_t for the distribution of X_t, $\nu_t = \nu_0 \mu_t$. Assuming ν_t has density ρ_t w.r.t. λ, one can get, by integrating the forward equation over space,

$$\partial_t \rho_t(x) = G^\dagger \rho_t(x) \tag{20.18}$$

and this, too, is sometimes called the "Fokker-Planck equation".

We saw, in the last section, that a diffusion process solving an equation with drift terms $a_i(x)$ and diffusion terms $b_{ij}(x)$ has the generator

$$Gf(x) = a_i(x)\partial_i f(x) + \frac{1}{2}(bb^T)_{ij}(x)\partial^2_{ij} f(x) \tag{20.19}$$

You can show — it's an exercise in vector calculus, integration by parts, etc. — that the adjoint to G is the differential operator

$$G^\dagger f(x) = -\partial_i a_i(x) f(x) + \frac{1}{2}\partial^2_{ij}(bb^T)_{ij}(x) f(x) \tag{20.20}$$

Notice that the space-dependence of the SDE's coefficients now appears *inside* the derivatives. Of course, if a and b are independent of x, then they simply pull outside the derivatives, giving us, in that special case,

$$G^\dagger f(x) = -a_i \partial_i f(x) + \frac{1}{2}(bb^T)_{ij}\partial^2_{ij} f(x) \tag{20.21}$$

Let's interpret this physically, imagining a large population of independent tracer particles wandering around the state space Ξ, following independent copies of the diffusion process. The second derivative term is easy: diffusion tends to smooth out the probability density, taking probability mass away from maxima (where $f'' < 0$) and adding it to minima. (Remember that bb^T is positive semi-definite.) If a_i is positive, then particles tend to move in the positive direction along the i^{th} axis. If $\partial_i \rho$ is also positive, this means that, on average, the point x sends more particles up along the axis than wander down, against the gradient, so the density at x will tend to decline.

Example 222 (Wiener process, heat equation) *Notice that (for diffusions produced by SDEs) $G^\dagger = G$ when $a = 0$ and b is constant over the state space. This is the case with Wiener processes, where $G = G^\dagger = \frac{1}{2}\nabla^2$. Thus, the heat equation holds both for the evolution of observable functions of the Wiener process, and for the evolution of the Wiener process's density. You should convince yourself that there is no non-negative integrable ρ such that $G\rho(x) = 0$.*

Example 223 (Ornstein-Uhlenbeck process) *For the one-dimensional Ornstein-Uhlenbeck process, the generator may be read off from the Langevin equation,*

$$Gf(p) = -\gamma p \partial_p f(p) + \frac{1}{2} D^2 \partial^2_{pp} f(p)$$

and the Fokker-Planck equation becomes

$$\partial_t \rho(p) = \gamma \partial_p (p\rho(p)) + D^2 \frac{1}{2} \partial^2_{pp} f(p)$$

It's easily checked that $\rho(p) = \mathcal{N}(0, D^2/2\gamma)$ gives $\partial_t \rho = 0$. That is, the long-run invariant distribution can be found as a stationary solution of the Fokker-Planck equation. See also Exercise 20.1.

20.3 White Noise

Scientists and engineers are often uncomfortable with the SDEs in the way probabilists write them, because they want to divide through by dt and have the result mean something. The trouble, of course, is that dW/dt does not, in any ordinary sense, exist. They, however, are often happier ignoring this inconvenient fact, and talking about "white noise" as what dW/dt ought to be. This is not totally crazy. Rather, one can define $\xi \equiv dW/dt$ as a *generalized* derivative, one whose value at any given time is a random real linear functional, rather than a random real number. Consequently, it only really makes sense in integral expressions (like the solutions of SDEs!), but it can, in many ways, be formally manipulated like an ordinary function.

One way to begin to make sense of this is to start with a standard Wiener process $W(t)$, and a C^1 non-random function $u(t)$, and to use integration by parts:

$$\begin{aligned}
\frac{d}{dt}(uW) &= u\frac{dW}{dt} + \frac{du}{dt}W && (20.22)\\
&= u(t)\xi(t) + \dot{u}(t)W(t) && (20.23)\\
\int_0^t \frac{d}{dt}(uW)ds &= \int_0^t \dot{u}(s)W(s) + u(s)\xi(s)ds && (20.24)\\
u(t)W(t) - u(0)W(0) &= \int_0^t \dot{u}(s)W(s)ds + \int_0^t u(s)\xi(s)ds && (20.25)\\
\int_0^t u(s)\xi(s)ds &\equiv u(t)W(t) - \int_0^t \dot{u}(s)W(s)ds && (20.26)
\end{aligned}$$

We can take the last line to *define* ξ, and time-integrals within which it appears. Notice that the terms on the RHS are well-defined *without* the Itô calculus: one is just a product of two measurable random variables, and the other is the time-integral of a continuous random function. With this definition, we can establish some properties of ξ.

Proposition 224 *$\xi(t)$ is a linear functional:*

$$\int_0^t (a_1 u_1(s) + a_2 u_2(s))\xi(s)ds = a_1 \int_0^t u_1(s)\xi(s)ds + a_2 \int_0^t u_2(s)\xi(s)ds \quad (20.27)$$

PROOF:

$$\begin{aligned} & \int_0^t (a_1 u_1(s) + a_2 u_2(s))\xi(s)ds & (20.28) \\ = & \ (a_1 u_1(t) + a_2 u_2(t))W(t) - \int_0^t (a_1 \dot{u}_1(s) + a_2 \dot{u}_2(s))W(s)ds & \\ = & \ a_1 \int_0^t u_1(s)\xi(s)ds + a_2 \int_0^t u_2(s)\xi(s)ds & (20.29) \end{aligned}$$

□

Proposition 225 *For all t, $\mathbf{E}[\xi(t)] = 0$*

PROOF:

$$\begin{aligned} \int_0^t u(s)\mathbf{E}[\xi(s)]\,ds &= \mathbf{E}\left[\int_0^t u(s)\xi(s)ds\right] & (20.30) \\ &= \mathbf{E}\left[u(t)W(t) - \int_0^t \dot{u}(s)W(s)ds\right] & (20.31) \\ &= \mathbf{E}[u(t)W(t)] - \int_0^t \dot{u}(s)\mathbf{E}[W(t)]\,ds & (20.32) \\ &= 0 - 0 = 0 & (20.33) \end{aligned}$$

Proposition 226 *For all $u \in C^1$, $\int_0^t u(s)\xi(s)ds = \int_0^t u(s)dW$.*

PROOF: Apply Itô's formula to the function $f(t, W) = u(t)W(t)$:

$$\begin{aligned} d(uW) &= W(t)\dot{u}(t)dt + u(t)dW & (20.34) \\ u(t)W(t) &= \int_0^t \dot{u}(s)W(s)ds + \int_0^t u(t)dW & (20.35) \\ \int_0^t u(t)dW &= u(t)W(t) - \int_0^t \dot{u}(s)W(s)ds & (20.36) \\ &= \int_0^t u(s)\xi(s)ds & (20.37) \end{aligned}$$

□

This can be used to extend the definition of white-noise integrals to any Itô-integrable process.

Proposition 227 *ξ has delta-function covariance:* $\mathrm{cov}(\xi(t_1), \xi(t_2)) = \delta(t_1 - t_2)$.

PROOF: Since $\mathbf{E}\left[\xi(t)\right] = 0$, we just need to show that $\mathbf{E}\left[\xi(t_1)\xi(t_2)\right] = \delta(t_1 - t_2)$. Remember (Eq. 17.14 on p. 95) that $\mathbf{E}\left[W(t_1)W(t_2)\right] = t_1 \wedge t_2$.

$$\begin{aligned}
& \int_0^t \int_0^t u(t_1)u(t_2)\mathbf{E}\left[\xi(t_1)\xi(t_2)\right] dt_1 dt_2 && (20.38)\\
= & \ \mathbf{E}\left[\int_0^t u(t_1)\xi(t_1)dt_1 \int_0^t u(t_2)\xi(t_2)dt_2\right] && (20.39)\\
= & \ \mathbf{E}\left[\left(\int_0^t u(t_1)\xi(t_1)dt_1\right)^2\right] && (20.40)\\
= & \int_0^t \mathbf{E}\left[u^2(t_1)\right] dt_1 = \int_0^t u^2(t_1)dt_1 && (20.41)
\end{aligned}$$

using the preceding proposition, the Itô isometry, and the fact that u is non-random. But

$$\int_0^t \int_0^t u(t_1)u(t_2)\delta(t_1 - t_2)dt_1 dt_2 = \int_0^t u^2(t_1)dt_1 \tag{20.42}$$

so $\delta(t_1 - t_2) = \mathbf{E}\left[\xi(t_1)\xi(t_2)\right] = \operatorname{cov}\left(\xi(t_1), \xi(t_2)\right)$. □

Proposition 228 *ξ is weakly stationary.*

PROOF: Its mean is independent of time, and its covariance depends only on $|t_1 - t_2|$, so it satisfies Definition 50. □

Proposition 229 *ξ is Gaussian, and hence strongly stationary.*

PROOF: To show that it is Gaussian, use Exercise 19.2. Strong stationarity follows from weak stationarity (Proposition 228) and the fact that it is Gaussian. □

20.4 Exercises

Exercise 20.1 *A conservative force is one derived from an external potential, i.e., there is a function $\phi(x)$ giving energy, and $F(x) = -d\phi/dx$. The equations of motion for a body subject to a conservative force, drag, and noise read*

$$\begin{aligned}
dx &= \frac{p}{m}dt && (20.43)\\
dp &= -\gamma p dt + F(x)dt + \sigma dW && (20.44)
\end{aligned}$$

a *Find the corresponding forward (Fokker-Planck) equation.*

b *Find a stationary density for this equation, at least up to normalization constants. Hint: use separation of variables, i.e., $\rho(x,p) = u(x)v(p)$. You should be able to find the normalizing constant for the momentum density $v(p)$, but not for the position density $u(x)$. (Its general form should however be familiar from theoretical statistics: what is it?)*

c *Show that your stationary solution reduces to that of the Ornstein-Uhlenbeck process, if* $F(x) = 0$.

Chapter 21

Spectral Analysis and L_2 Ergodicity

Section 21.1 introduces the spectral representation of weakly stationary processes, and the central Wiener-Khinchin theorem connecting autocovariance to the power spectrum. Subsection 21.1.1 explains why white noise is "white".

Section 21.2 gives our first classical ergodic result, the "mean square" (L_2) ergodic theorem for weakly stationary processes. Subsection 21.2.1 gives an easy proof of a sufficient condition, just using the autocovariance. Subsection 21.2.2 gives a necessary and sufficient condition, using the spectral representation.

Any reasonable real-valued function $x(t)$ of time, $t \in \mathbb{R}$, has a Fourier transform, that is, we can write

$$\tilde{x}(\nu) = \frac{1}{2\pi} \int_{-\infty}^{\infty} dt e^{i\nu t} x(t)$$

which can usually be inverted to recover the original function,

$$x(t) = \int_{-\infty}^{\infty} d\nu e^{-i\nu t} \tilde{x}(\nu)$$

This one example of an "analysis", in the original sense of resolving into parts, of a function into a collection of orthogonal basis functions. (You can find the details in any book on Fourier analysis, as well as the varying conventions on where the 2π goes, the constraints on $\tilde{x}$ which arise from the fact that x is real, etc.)

There are various reasons to prefer the trigonometric basis functions $e^{i\nu t}$ over other possible choices. One is that they are invariant under translation in time, which just changes phases[1]. This suggests that the Fourier basis will

[1] If $t \mapsto t + \tau$, then $\tilde{x}(\nu) \mapsto e^{i\nu\tau}\tilde{x}(\nu)$.

be particularly useful when dealing with time-invariant systems. For stochastic processes, however, time-invariance is stationarity. This suggests that there should be some useful way of doing Fourier analysis on stationary random functions. In fact, it turns out that stationary and even weakly-stationary processes *can* be productively Fourier-transformed. This is potentially a huge topic, especially when it's expanded to include representing random functions in terms of (countable) series of orthogonal functions. The spectral theory of random functions connects Fourier analysis, disintegration of measures, Hilbert spaces and ergodicity. This lecture will do no more than scratch the surface, covering, in succession, the basics of the spectral representation of weakly-stationary random functions and the fundamental Wiener-Khinchin theorem linking covariance functions to power spectra, why white noise is called "white", and the mean-square ergodic theorem.

Good sources, if you want to go further, are the books of Bartlett (1955, ch. 6) (from whom I've stolen shamelessly), the historically important and inspiring Wiener (1949, 1961), and of course Doob (1953). Loève (1955, ch. X) is highly edifying, particular his discussion of Karhunen-Loève transforms, and the associated construction of the Wiener process as a Fourier series with random phases.

21.1 Spectral Representation of Weakly Stationary Procesess

This section will only handle spectral representations of real-valued one-parameter processes in continuous time. Generalizations to vector-valued and multi-parameter processes are straightforward; handling discrete time is actually in some ways more irritating, because of limitations on allowable frequencies of Fourier components (to the range from $-\pi$ to π).

Definition 230 (Autocovariance Function) *Suppose that, for all $t \in T$, X is real and $\mathbf{E}\left[X^2(t)\right]$ is finite. Then $\Gamma(t_1, t_2) \equiv \mathbf{E}\left[X(t_1)X(t_2)\right]$ is the* autocovariance function *of the process. If the process is weakly stationary, so that $\Gamma(t, t+\tau) = \Gamma(0, \tau)$ for all t, τ, write $\Gamma(\tau)$. If $X(t) \in \mathbb{C}$, then $\Gamma(t_1, t_2) \equiv \mathbf{E}\left[X^\dagger(t_1)X(t_2)\right]$, where $\dagger$ is complex conjugation.*

Proposition 231 *If X is real and weakly stationary, then $\Gamma(\tau) = \Gamma(-\tau)$; if X is complex and weakly stationary, then $\Gamma(\tau) = \Gamma^\dagger(-\tau)$.*

PROOF: Direct substitution into the definitions. □

Remarks on terminology. It is common, when only dealing with one stochastic process, to drop the qualifying "auto" and just speak of the covariance function; I probably will myself. It is also common (especially in the time series literature) to switch to the (auto)correlation function, i.e., to normalize by the standard deviations. Finally, be warned that the statistical physics literature (e.g. Forster, 1975) uses "correlation function" to mean $\mathbf{E}\left[X(t_1)X(t_2)\right]$, i.e., the *uncentered* mixed second moment. This is a matter of tradition, not (despite appearances) ignorance.

Definition 232 (Second-Order Process) *A real-valued process X is* second order *when $\mathbf{E}\left[X^2(t)\right] < \infty$ for all t.*

Definition 233 (Spectral Representation, Power Spectrum) *A real-valued process X on T has a complex-valued* spectral process *$\tilde{X}$, if it has a* spectral representation*:*

$$X(t) \equiv \int_{-\infty}^{\infty} e^{-i\nu t} d\tilde{X}_\nu \tag{21.1}$$

The power spectrum *$V(\nu) \equiv \mathbf{E}\left[\left|\tilde{X}(\nu)\right|^2\right]$.*

Remark. The name "power spectrum" arises because this is proportional to the amount of power (energy per unit time) carried by oscillations of frequency $\leq \nu$, at least in a linear system.

Notice that if a process has a spectral representation, then, roughly speaking, for a fixed ω the amplitudes of the different Fourier components in $X(t, \omega)$ are fixed, and shifting forward in time just involves changing their phases. (Making this simple is why we have to allow $\tilde{X}$ to have complex values.)

Proposition 234 *When it exists, $\tilde{X}(\nu)$ has right and left limits at every point ν, and limits as $\nu \rightarrow \pm\infty$.*

Proof: See Loève (1955, §34.4). You can prove this yourself, however, using the material on characteristic functions in 36-752. □

Definition 235 *The* jump of the spectral process at *ν, $\Delta\tilde{X}(\nu) \equiv \tilde{X}(\nu + 0) - \tilde{X}(\nu - 0)$.*

Remark 1: As usual, $\tilde{X}(\nu+0) \equiv \lim_{h\downarrow 0} \tilde{X}(\nu + h)$, and $\tilde{X}(\nu-0) \equiv \lim_{h\downarrow 0} \tilde{X}(\nu - h)$. The jump at ν is the difference between the right and left-hand limits at ν.

Remark 2: Some people call the set of points at which the jump is non-zero the "spectrum". This usage comes from functional analysis, but seems needlessly confusing in the present context.

Proposition 236 *Every weakly-stationary process has a spectral representation.*

Proof: See Loève (1955, §34.4), or Bartlett (1955, §6.2). □

The spectral representation is another stochastic integral, and it can be made sense of in the same way that we made sense of integrals with respect to the Wiener process, by starting with elementary functions and building up from there. Crucial in this development is the following property.

Definition 237 (Orthogonal Increments) *A one-parameter random function (real or complex) has* orthogonal increments *if, for $t_1 \leq t_2 \leq t_3 \leq t_4 \in T$, the covariance of the increment from t_1 to t_2 and the increment from t_3 to t_4 is always zero:*

$$\mathbf{E}\left[\left(\tilde{X}(\nu_4) - \tilde{X}(\nu_3)\right)\left(\tilde{X}(\nu_2) - \tilde{X}(\nu_1)\right)^{\dagger}\right] = 0 \tag{21.2}$$

Proposition 238 *The spectral process of a second-order process has orthogonal increments if and only if the process is weakly stationary.*

Sketch Proof: Assume, without loss of generality, that $\mathbf{E}[X(t)] = 0$, so $\mathbf{E}\left[\tilde{X}(\nu)\right] = 0$. "If": We can write, using the fact that $X(t) = X^{\dagger}(t)$ for real-valued processes,

$$\begin{aligned}
\Gamma(\tau) &= \Gamma(t, t+\tau) && (21.3)\\
&= \mathbf{E}\left[X^{\dagger}(t)X(t+\tau)\right] && (21.4)\\
&= \mathbf{E}\left[\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{i\nu_1 t}e^{-i\nu_2 t+\tau} d\tilde{X}^{\dagger}_{\nu_1} d\tilde{X}_{\nu_2}\right] && (21.5)\\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{i(\nu_1-\nu_2)t}e^{-i\nu_2\tau}\mathbf{E}\left[d\tilde{X}^{\dagger}_{\nu_1} d\tilde{X}_{\nu_2}\right] && (21.6)
\end{aligned}$$

Since t is arbitrary, every term on the right must be independent of t, which implies the orthogonality of the increments of $\tilde{X}$. "Only if": if the increments are orthogonal, then clearly the steps of the argument can be reversed to conclude that $\Gamma(t_1, t_2)$ depends only on $t_2 - t_1$. □

Definition 239 (Spectral Function, Spectral Density) *The spectral function of a weakly stationary process is the function $S(\nu)$ appearing in the spectral representation of its autocovariance:*

$$\Gamma(\tau) = \int_{-\infty}^{\infty} e^{-i\nu\tau} dS_{\nu} \tag{21.7}$$

Remark. Some people prefer to talk about the spectral function as the Fourier transform of the autocorrelation function, rather than of the autocovariance. This has the advantage that the spectral function turns out to be a normalized cumulative distribution function (see Theorem 240 immediately below), but is otherwise inconsequential.

Theorem 240 *The spectral function exists for every weakly stationary process, if $\Gamma(\tau)$ is continuous. Moreover, $S(\nu) \geq 0$, S is non-decreasing, $S(-\infty) = 0$, $S(\infty) = \Gamma(0)$, and $lim_{h\downarrow 0}S(\nu + h)$ and $\lim_{h\downarrow 0} S(\nu - h)$ exist for every ν.*

Proof: Usually, by an otherwise-obscure result in Fourier analysis called Bochner's theorem. A more direct proof is due to Loève. Assume, without loss of generality, that $\mathbf{E}[X(t)] = 0$.

Start by defining

$$H_T(\nu) \equiv \frac{1}{\sqrt{T}} \int_{-T/2}^{T/2} e^{i\nu t} X(t) dt \tag{21.8}$$

and define $f_T(\nu)$ through H:

$$\begin{aligned}
2\pi f_T(\nu) &\equiv \mathbf{E}\left[H_T(\nu) H_T^\dagger(\nu)\right] &(21.9)\\
&= \mathbf{E}\left[\frac{1}{T}\int_{-T/2}^{T/2}\int_{-T/2}^{T/2} e^{i\nu t_1} X(t_1) e^{-i\nu t_2} X^\dagger(t_2) dt_1 dt_2\right] &(21.10)\\
&= \frac{1}{T}\int_{-T/2}^{T/2}\int_{-T/2}^{T/2} e^{i\nu(t_1-t_2)} \mathbf{E}\left[X(t_1)X(t_2)\right] dt_1 dt_2 &(21.11)\\
&= \frac{1}{T}\int_{-T/2}^{T/2}\int_{-T/2}^{T/2} e^{i\nu(t_1-t_2)} \Gamma(t_1-t_2) dt_1 dt_2 &(21.12)\\
&= \int_{-T}^{T} \left(1 - \frac{|\tau|}{T}\right) \Gamma(\tau) e^{i\nu\tau} d\tau &(21.13)
\end{aligned}$$

Recall that $\Gamma(\tau)$ defines a non-negative quadratic form, meaning that

$$\sum_{s,t} a_s^\dagger a_t \Gamma(t-s) \geq 0$$

for any sets of times and any complex numbers a_t. This will in particular work if the complex numbers lie on the unit circle and can be written $e^{i\nu t}$. This means that integrals

$$\int\int e^{i\nu(t_1-t_2)} \Gamma(t_1-t_2) dt_1 dt_2 \geq 0 \tag{21.14}$$

so $f_T(\nu) \geq 0$.

Define $\phi_T(\tau)$ as the integrand in Eq. 21.13, so that

$$f_T(\nu) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \phi_T(\tau) e^{i\nu\tau} d\tau \tag{21.15}$$

which is recognizable as a proper Fourier transform. Now pick some $N > 0$ and massage the equation so it starts to look like an inverse transform.

$$\begin{aligned}
f_T(\nu) e^{-i\nu t} &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \phi_T(\tau) e^{i\nu\tau} e^{-i\nu t} d\tau &(21.16)\\
\left(1 - \frac{|\nu|}{N}\right) f_T(\nu) e^{-i\nu t} &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \phi_T(\tau) e^{i\nu\tau} e^{-i\nu t} \left(1 - \frac{|\nu|}{N}\right) d\tau &(21.17)
\end{aligned}$$

Integrating over frequencies,

$$\int_{-N}^{N}\left(1-\frac{|\nu|}{N}\right)f_T(\nu)e^{-i\nu t}d\nu \tag{21.18}$$

$$= \int_{-N}^{N}\frac{1}{2\pi}\int_{-\infty}^{\infty}\phi_T(\tau)e^{i\nu\tau}e^{-i\nu t}\left(1-\frac{|\nu|}{N}\right)d\tau d\nu$$

$$= \frac{1}{2\pi}\int_{-\infty}^{\infty}\phi_T(\tau)\left(\frac{\sin N(\tau-t)/2}{N(\tau-t)/2}\right)^2 N d\tau \tag{21.19}$$

$\left(1-\frac{|\nu|}{N}\right)f_T(\nu)\geq 0$, so the left-hand side of the final equation is like a characteristic function of a distribution, up to, perhaps, an over-all normalizing factor, which will be $\phi_T(0)=\Gamma(0)>0$. Since $\Gamma(\tau)$ is continuous, $\phi_T(\tau)$ is too, and so, as $N\rightarrow\infty$, the right-hand side converges uniformly on $\phi_T(t)$, but a uniform limit of characteristic functions is still a characteristic function. Thus $\phi_T(t)$, too, can be obtained from a characteristic function. Finally, since $\Gamma(t)$ is the uniform limit of $\phi_T(t)$ on every bounded interval, $\Gamma(t)$ has a characteristic-function representation of the stated form. This allows us to further conclude that $S(\nu)$ is real-valued, non-decreasing, $S(-\infty)=0$ and $S(\infty)=\Gamma(0)$, and has both right and left limits everywhere. □

There is a converse, with a cute constructive proof.

Theorem 241 *Let $S(\nu)$ be any function with the properties described at the end of Theorem 240. Then there is a weakly stationary process whose autocovariance is of the form given in Eq. 21.7.*

PROOF: Define $\sigma^2=\Gamma(0)$, $F(\nu)=S(\nu)/\sigma^2$. Now $F(\nu)$ is a properly normalized cumulative distribution function. Let N be a random variable distributed according to F, and $\Phi\sim U(0,2\pi)$ be independent of A. Set $X(t)\equiv\sigma e^{i(\Phi-Nt)}$. Then $\mathbf{E}\left[X(t)\right]=\sigma\mathbf{E}\left[e^{i\Phi}\right]\mathbf{E}\left[e^{-iNt}\right]=0$. Moreover,

$$\mathbf{E}\left[X^{\dagger}(t_1)X(t_2)\right] = \sigma^2\mathbf{E}\left[e^{-i(\Phi-Nt_1)}e^{i(\Phi-Nt_2)}\right] \tag{21.20}$$

$$= \sigma^2\mathbf{E}\left[e^{-iN(t_1-t_2)}\right] \tag{21.21}$$

$$= \sigma^2\int_{-\infty}^{\infty}e^{-i\nu(t_1-t_2)}dF_{nu} \tag{21.22}$$

$$= \Gamma(t_1-t_2) \tag{21.23}$$

□

Definition 242 *The* jump of the spectral function at ν, $\Delta S(\nu)$, *is* $S(\nu+0)-S(\nu-0)$.

Proposition 243 $\Delta S(\nu)\geq 0$.

PROOF: Obvious from the fact that $S(\nu)$ is non-decreasing. □

Theorem 244 (Wiener-Khinchin Theorem) *If X is a weakly stationary process, then its power spectrum is equal to its spectral function.*

$$V(\nu) \equiv \mathbf{E}\left[\left|\tilde{X}(\nu)\right|^2\right] = S(\nu) \tag{21.24}$$

Proof: Assume, without loss of generality, that $\mathbf{E}[X(t)] = 0$. Substitute the spectral representation of X into the autocovariance, using Fubini's theorem to turn a product of integrals into a double integral.

$$\begin{aligned}
\Gamma(\tau) &= \mathbf{E}[X(t)X(t+\tau)] && (21.25)\\
&= \mathbf{E}\left[X^{\dagger}(t)X(t+\tau)\right] && (21.26)\\
&= \mathbf{E}\left[\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-i(t+\tau)\nu_1}e^{it\nu_2}d\tilde{X}_{\nu_1}d\tilde{X}_{\nu_2}\right] && (21.27)\\
&= \mathbf{E}\left[\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-it(\nu_1-\nu_2)}e^{-i\tau\nu_2}d\tilde{X}_{\nu_1}d\tilde{X}_{\nu_2}\right] && (21.28)\\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-it(\nu_1-\nu_2)}e^{-i\tau\nu_2}\mathbf{E}\left[d\tilde{X}_{\nu_1}d\tilde{X}_{\nu_2}\right] && (21.29)
\end{aligned}$$

using the fact that integration and expectation commute to (formally) bring the expectation inside the integral. Since $\tilde{X}$ has orthogonal increments, $\mathbf{E}\left[d\tilde{X}^{\dagger}_{\nu_1}d\tilde{X}_{\nu_2}\right] = 0$ unless $\nu_1 = \nu_2$. This turns the double integral into a single integral, and kills the $e^{-it(\nu_1-\nu_2)}$ factor, which had to go away because t was arbitrary.

$$\begin{aligned}
\Gamma(\tau) &= \int_{-\infty}^{\infty} e^{-i\tau\nu}\mathbf{E}\left[d(\tilde{X}^{\dagger}_{\nu}\tilde{X}_{\nu})\right] && (21.30)\\
&= \int_{-\infty}^{\infty} e^{-i\tau\nu}dV_{\nu} && (21.31)
\end{aligned}$$

using the definition of the power spectrum. Since $\Gamma(\tau) = \int_{-\infty}^{\infty} e^{-i\tau\nu}dV_{nu}$, it follows that S_ν and V_ν differ by a constant, namely the value of $V(-\infty)$, which can be chosen to be zero without affecting the spectral representation of X. □

21.1.1 How the White Noise Lost Its Color

Why is white noise, as defined in Section 20.3, called "white"? The answer is easy, given the Wiener-Khinchin relation in Theorem 244.

Recall from Proposition 227 that the autocovariance function of white noise is $\delta(t_1 - t_2)$. Recall from general analysis that one representation of the delta function is the following Fourier integral:

$$\delta(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} d\nu e^{i\nu t}$$

(This can be "derived" from inserting the definition of the Fourier transform into the inverse Fourier transform, among other, more respectable routes.) Appealing then to the theorem, $S(\nu) = \frac{1}{2\pi}$ for all ν. That is, there is equal power at all frequencies, just as white light is composed of light of all colors (frequencies), mixed with equal intensity.

Relying on this analogy, there is an elaborate taxonomy red, pink, black, brown, and other variously-colored noises, depending on the shape of their power spectra. The value of this terminology has honestly never been very clear to me, but the curious reader is referred to the (very fun) book of Schroeder (1991) and references therein.

21.2 The Mean-Square Ergodic Theorem

Ergodic theorems relate functionals calculated along individual sample paths (say, the time average, $T^{-1}\int_0^T dtX(t)$, or the maximum attained value) to functionals calculated over the whole distribution (say, the expectation, $\mathbf{E}\left[X(t)\right]$, or the expected maximum). The basic idea is that the two should be close, and they should get closer the longer the trajectory we use, because in some sense any one sample path, carried far enough, is representative of the whole distribution. Since there are many different kinds of functionals, and many different modes of stochastic convergence, there are many different kinds of ergodic theorem. The classical ergodic theorems say that time averages converge on expectations[2], either in L_p or a.s. (both implying convergence in distribution or in probability). The traditional centerpiece of ergodic theorem is Birkhoff's "individual" ergodic theorem, asserting a.s. convergence. We will see its proof, but it will need a lot of preparatory work, and it requires strict stationarity. By contrast, the L_2, or "mean square", ergodic theorem, attributed to von Neumann[3] is already in our grasp, and holds for weakly stationary processes.

We will actually prove it twice, once with a fairly transparent sufficient condition, and then again with a more complicated necessary-and-sufficient condition. The more complicated proof will wait until next lecture.

21.2.1 Mean-Square Ergodicity Based on the Autocovariance

First, the easy version, which gives an estimate of the rate of convergence. (What I say here is ripped off from the illuminating discussion in (Frisch, 1995, sec. 4.4, especially pp. 49–50).)

Definition 245 (Time Averages) *When X is a one-sided, continuous-parameter random process, we say that its* time average *between times T_1 and T_2 is* $\overline{X}(T_1, T_2) \equiv$

[2] Proverbially: "time averages converge on space averages", the space in question being the state space Ξ; or "converge on phase averages", since physicists call certain kinds of state space "phase space".

[3] See von Plato (1994, ch. 3) for a fascinating history of the development of ergodic theory through the 1930s, and its place in the history of mathematical probability.

$(T_2 - T_1)^{-1} \int_{T_1}^{T_2} dt X(t)$. When we only mention one time argument, by default the time average is from 0 to T, $\overline{X}(T) \equiv \overline{X}(0,T)$.

(Only considering time averages starting from zero involves no loss of generality for weakly stationary processes: why?)

Theorem 246 *Let $X(t)$ be a weakly stationary process, $\mathbf{E}[X(t)] = 0$. If $\int_0^\infty d\tau |\Gamma(\tau)| < \infty$, then $\overline{X}(T) \overset{L_2}{\rightarrow} 0$ as $T \rightarrow \infty$.*

Proof: Use Fubini's theorem to to the square of the integral into a double integral, and then bring the expectation inside it:

$$\begin{aligned}
\mathbf{E}\left[\left(\frac{1}{T}\int_0^T dt X(t)\right)^2\right] &= \mathbf{E}\left[\frac{1}{T^2}\int_0^T\int_0^T dt_1 dt_2 X(t_1)X(t_2)\right] && (21.32)\\
&= \frac{1}{T^2}\int_0^T\int_0^T dt_1 dt_2 \mathbf{E}\left[X(t_1)X(t_2)\right] && (21.33)\\
&= \frac{1}{T^2}\int_0^T\int_0^T dt_1 dt_2 \Gamma(t_1 - t_2) && (21.34)\\
&= \frac{2}{T^2}\int_0^T dt_1 \int_0^{t_1} d\tau \Gamma(\tau) && (21.35)\\
&\leq \frac{2}{T^2}\int_0^T dt_1 \int_0^{\infty} d\tau |\Gamma(\tau)| && (21.36)\\
&= \frac{2}{T}\int_0^{\infty} d\tau |\Gamma(\tau)| && (21.37)
\end{aligned}$$

As $T \rightarrow \infty$, this $\rightarrow 0$. □

Remark. From the proof, we can see that the rate of convergence of the mean-square of $\left\|\overline{X}(T)\right\|_2^2$ is (at least) $O(1/T)$. This would give a root-mean-square (rms) convergence rate of $O(1/\sqrt{T})$, which is what the naive statistician who ignored inter-temporal dependence would expect from the central limit theorem. (This ergodic theorem says *nothing* about the form of the distribution of $\overline{X}(T)$ for large T. We will see that, under some circumstances, it *is* Gaussian, but that needs stronger assumptions [forms of "mixing"] than we have imposed.) The naive statistician would expect that the mean-square time average would go like $\Gamma(0)/T$, since $\Gamma(0) = \mathbf{E}\left[X^2(t)\right] = \mathbf{Var}\left[X(t)\right]$. The proportionality constant is instead $\int_0^\infty d\tau |\Gamma(\tau)|$. This is equal to the naive guess for white noise, and for other collections of IID variables, but not in the general case. This leads to the following

Definition 247 (Integral Time Scale) *The* integral time scale *of a weakly-stationary random process, $\mathbf{E}[X(t)] = 0$, is*

$$\tau_{\text{int}} \equiv \frac{\int_0^\infty d\tau |\Gamma(\tau)|}{\Gamma(0)} \quad (21.38)$$

Notice that τ_{int} does, indeed, have units of time.

Corollary 248 *Under the conditions of Theorem 246,*

$$\mathbf{Var}\left[\overline{X}(T)\right] \leq 2\mathbf{Var}\left[X(0)\right]\frac{\tau_{\text{int}}}{T} \tag{21.39}$$

Proof: Since $X(t)$ is centered, $\mathbf{E}\left[\overline{X}(T)\right] = 0$, and $\left\|\overline{X}(T)\right\|_2^2 = \mathbf{Var}\left[\overline{X}(T)\right]$. Everything else follows from re-arranging the bound in the proof of Theorem 246, Definition 247, and the fact that $\Gamma(0) = \mathbf{Var}\left[X(0)\right]$. □

As a consequence of the corollary, if $T \gg \tau_{\text{int}}$, then the variance of the time average is negigible compared to the variance at any one time.

21.2.2 Mean-Square Ergodicity Based on the Spectrum

Let's warm up with some lemmas of a technical nature. The first relates the jumps of the spectral process $\tilde{X}(\nu)$ to the jumps of the spectral function $S(\nu)$.

Lemma 249 *For a weakly stationary process,* $\mathbf{E}\left[\left|\Delta\tilde{X}(\nu)\right|^2\right] = \Delta S(\nu)$.

Proof: This follows directly from the Wiener-Khinchin relation (Theorem 244). □

Lemma 250 *The jump of the spectral function at ν is given by*

$$\Delta S(\nu) = \lim_{T\rightarrow\infty}\frac{1}{T}\int_0^T \Gamma(\tau)e^{i\nu\tau}d\tau \tag{21.40}$$

Proof: This is a basic inversion result for characteristic functions. It should become plausible by thinking of this as getting the Fourier transform of Γ as T grows. □

Lemma 251 *If X is weakly stationary, then for any real f, $\overline{e^{ift}X}(T)$ converges in L_2 to $\Delta\tilde{X}(f)$.*

Proof: Start by looking at the squared modulus of the time average for finite time.

$$\left|\frac{1}{T}\int_0^T e^{ift}X(t)dt\right|^2 \tag{21.41}$$

$$= \frac{1}{T^2}\int_0^T\int_0^T e^{-if(t_1-t_2)}X^{\dagger}(t_1)X(t_2)dt_1dt_2$$

$$= \frac{1}{T^2}\int_0^T\int_0^T e^{-if(t_1-t_2)}\int_{-\infty}^{\infty}e^{i\nu_1 t_1}d\tilde{X}_{\nu_1}\int_{-\infty}^{\infty}e^{-i\nu_2 t_2}d\tilde{X}_{\nu_2} \tag{21.42}$$

$$= \frac{1}{T^2}\int_0^T\int_{-\infty}^{\infty}dt_1 d\tilde{X}_{\nu_1}e^{it_1(f-\nu_1)}\int_0^T\int_{-\infty}^{\infty}dt_2 d\tilde{X}_{\nu_2}e^{-it_2(f-\nu_2)} \tag{21.43}$$

As $T \to \infty$, these integrals pick out $\Delta\tilde{X}(f)$ and $\Delta\tilde{X}^{\dagger}(f)$. So, $\overline{e^{ift}X}(T) \overset{L_2}{\to} \Delta\tilde{X}(f)$. □

Notice that the limit provided by the lemma is a random quantity. What's really desired, in most applications, is convergence to a *deterministic* limit, which here would mean convergence (in L_2) to zero.

Theorem 252 (Mean-Square Ergodic Theorem) *If X is weakly stationary, and $\mathbf{E}[X(t)] = 0$, then $\overline{X}(t)$ converges in L_2 to 0 iff*

$$\lim T^{-1} \int_0^T d\tau \Gamma(\tau) = 0 \tag{21.44}$$

Proof: Taking $f = 0$ in Lemma 251, $\overline{X}(T) \overset{L_2}{\to} \Delta\tilde{X}(0)$, the jump in the spectral function at zero. Let's show that the (i) expectation of this jump is zero, and that (ii) its variance is given by the integral expression on the LHS of Eq. 21.44. For (i), because $\overline{X}(T) \overset{L_2}{\to} Y$, we know that $\mathbf{E}\left[\overline{X}(T)\right] \to \mathbf{E}[Y]$. But $\mathbf{E}\left[\overline{X}(T)\right] = \overline{\mathbf{E}[X]}(T) = 0$. So $\mathbf{E}\left[\Delta\tilde{X}(0)\right] = 0$. For (ii), Lemma 249, plus the fact that $\mathbf{E}\left[\Delta\tilde{X}(0)\right] = 0$, shows that the variance is equal to the jump in the spectrum at 0. But, by Lemma 250 with $\nu = 0$, that jump is exactly the LHS of Eq. 21.44. □

Remark 1: Notice that if the integral time is finite, then the integral condition on the autocovariance is automatically satisfied, but not vice versa, so the hypotheses here are strictly weaker than in Theorem 246.

Remark 2: One interpretation of the theorem is that the time-average is converging on the zero-frequency component of the spectral process. If there is a jump at 0, then this has finite variance; if not, not.

Remark 3: Lemma 251 establishes the L_2 convergence of time-averages of the form

$$\frac{1}{T}\int_0^T e^{ift} X(t) dt$$

for any real f. Specifically, from Lemma 249, the mean-square of this variable is converging on the jump in the spectrum at f. While the ergodic theorem itself only needs the $f = 0$ case, this result is useful in connection with estimating spectra from time series (Doob, 1953, ch. X, §7).

21.3 Exercises

Exercise 21.1 *It is often convenient to have a mean-square ergodic theorem for discrete-time sequences rather than continuous-time processes. If the dt in the definition of $\overline{X}$ is re-interpreted as counting measure on $\mathbb{N}$, rather than Lebesgue measure on $\mathbb{R}^+$, does the proof of Theorem 246 remain valid? (If yes, say why; if no, explain where the argument fails.)*

Exercise 21.2 *State and prove a version of Theorem 246 which does not assume that* $\mathbf{E}[X(t)] = 0$.

Exercise 21.3 *Suppose* X *is a weakly stationary process, and* f *is a measurable function such that* $\|f(X_0)\|_2 < \infty$. *Is* $f(X)$ *a weakly stationary process? (If yes, prove it; if not, give a counter-example.)*

Exercise 21.4 *Suppose the Ornstein-Uhlenbeck process is has its invariant distribution as its initial distribution, and is therefore weakly stationary. Does Theorem 246 apply?*

Chapter 22

Large Deviations for Small-Noise Stochastic Differential Equations

This lecture is at once the end of our main consideration of diffusions and stochastic calculus, and a first taste of large deviations theory. Here we study the divergence between the trajectories produced by an ordinary differential equation, and the trajectories of the same system perturbed by a small amount of white noise.

Section 22.1 establishes that, in the small noise limit, the SDE's trajectories converge in probability on the ODE's trajectory. This uses Feller-process convergence.

Section 22.2 upper bounds the rate at which the probability of large deviations goes to zero as the noise vanishes. The methods are elementary, but illustrate deeper themes to which we will recur once we have the tools of ergodic and information theory.

In this chapter, we will use the results we have already obtained about SDEs to give a rough estimate of a basic problem, frequently arising in practice[1] namely taking a system governed by an ordinary differential equation and seeing how much effect injecting a small amount of white noise has. More exactly, we will put an upper bound on the probability that the perturbed trajectory goes very far from the unperturbed trajectory, and see the rate at which this probability goes to zero as the amplitude ϵ of the noise shrinks; this will be

[1] For applications in statistical physics and chemistry, see Keizer (1987). For applications in signal processing and systems theory, see Kushner (1984). For applications in nonparametric regression and estimation, and also radio engineering (!) see Ibragimov and Has'minskii (1979/1981). The last book is especially recommended for those who care about the connections between stochastic process theory and statistical inference, but unfortunately expounding the results, or even just the problems, would require a too-long detour through asymptotic statistical theory.

$O(e^{-C\epsilon^2})$. This will be our first illustration of a large deviations calculation. It will be crude, but it will also introduce some themes to which we will return (inshallah!) at greater length towards the end of the course. Then we will see that the major improvement of the more refined tools is to give a lower bound to match the upper bound we will calculate now, and see that we at least got the logarithmic rate right.

I should say before going any further that this example is shamelessly ripped off from Freidlin and Wentzell (1998, ch. 3, sec. 1, pp. 70–71), which is *the* book on the subject of large deviations for continuous-time processes.

22.1 Convergence in Probability of SDEs to ODEs

To begin with, consider an unperturbed ordinary differential equation:

$$\frac{d}{dt}x(t) = a(x(t)) \quad (22.1)$$

$$x(0) = x_0 \in \mathbb{R}^d \quad (22.2)$$

Assume that a is uniformly Lipschitz-continuous (as in the existence and uniqueness theorem for ODEs, and more to the point for SDEs). Then, for the given, non-random initial condition, there exists a unique continuous function x which solves the ODE.

Now, for $\epsilon > 0$, consider the SDE

$$dX_\epsilon = a(X_\epsilon)dt + \epsilon dW \quad (22.3)$$

where W is a standard d-dimensional Wiener process, with non-random initial condition $X_\epsilon(0) = x_0$. Theorem 216 clearly applies, and consequently so does Theorem 220, meaning X_ϵ is a Feller diffusion with generator $G_\epsilon f(x) = a_i(x)\partial_i f'(x) + \frac{\epsilon^2}{2}\nabla^2 f(x)$.

Write X_0 for the deterministic solution of the ODE.

Our first assertion is that $X_\epsilon \xrightarrow{d} X_0$ as $\epsilon \rightarrow 0$. Notice that X_0 is a Feller process[2], whose generator is $G_0 = a_i(x)\partial_i$. We can apply Theorem 170 on convergence of Feller processes. Take the class of functions with bounded second derivatives. This is clearly a core for G_0, and for every G_ϵ. For every function f in this class,

$$\|G_\epsilon f - G_0 f\|_\infty = \left\|a_i\partial_i f(x) + \frac{\epsilon^2}{2}\nabla^2 f(x) - a_i\partial_i f(x)\right\|_\infty \quad (22.4)$$

$$= \frac{\epsilon^2}{2}\left\|\nabla^2 f(x)\right\|_\infty \quad (22.5)$$

which goes to zero as $\epsilon \rightarrow 0$. But this is condition (i) of the convergence theorem, which is equivalent to condition (iv), that convergence in distribution of the

[2] You can amuse yourself by showing this. Remember that $X_y(t) \xrightarrow{d} X_x(t)$ is equivalent to $\mathbf{E}\left[f(X_t)|X_0 = y\right] \rightarrow \mathbf{E}\left[f(X_t)|X_0 = x\right]$ for all bounded *continuous* f, and the solution of an ODE depends continuously on its initial condition.

initial condition implies convergence in distribution of the whole trajectory. Since the initial condition is the same non-random point x_0 for all ϵ, we have $X_\epsilon \xrightarrow{d} X_0$ as $\epsilon \rightarrow 0$. In fact, since X_0 is non-random, we have that $X_\epsilon \xrightarrow{P} X_0$. That last assertion really needs some consideration of metrics on the space of continuous random functions to make sense (see Appendix A2 of Kallenberg), but once that's done, the upshot is

Theorem 253 *Let $\Delta_\epsilon(t) = |X_\epsilon(t) - X_0(t)|$. For every $T > 0$, $\delta > 0$,*

$$\lim_{\epsilon \rightarrow 0} \mathbb{P}\left(\sup_{0 \leq t \leq T} \Delta_\epsilon(t) > \delta \right) = 0 \tag{22.6}$$

Or, using the maximum-process notation, for every $T > 0$,

$$\Delta(T)^* \xrightarrow{P} 0 \tag{22.7}$$

PROOF: See above. □

This is a version of the weak law of large numbers, and nice enough in its own way. One crucial limitation, however, is that it tells us nothing about the *rate* of convergence. That is, it leaves us clueless about how big the noise can be, while still leaving us in the small-noise limit. If the rate of convergence were, say, $O(\epsilon^{1/100})$, then this would not be very useful. (In fact, if the convergence were that slow, we should be really suspicious of numerical solutions of the *unperturbed* ODE.)

22.2 Rate of Convergence; Probability of Large Deviations

Large deviations theory is essentially a study of rates of convergence in probabilistic limit theorems. Here, we will estimate the rate of convergence: our methods will be crude, but it will turn out that even more refined estimates won't change the rate, at least not by more than log factors.

Let's go back to the difference between the perturbed and unperturbed trajectories, going through our now-familiar procedure.

$$\begin{aligned}
X_\epsilon(t) - X_0(t) &= \int_0^t \left[a(X_\epsilon(s)) - a(X_0(s))\right] ds + \epsilon W(t) &(22.8)\\
\Delta_\epsilon(t) &\leq \int_0^t \left|a(X_\epsilon(s)) - a(X_0(s))\right| ds + \epsilon |W(t)| &(22.9)\\
&\leq K_a \int_0^t \Delta_\epsilon(s) ds + \epsilon |W(t)| &(22.10)\\
\Delta^*_\epsilon(T) &\leq \epsilon W^*(T) + K_a \int_0^t \Delta^*_\epsilon(s) ds &(22.11)
\end{aligned}$$

Applying Gronwall's Inequality (Lemma 214),

$$\Delta^*_\epsilon(T) \leq \epsilon W^*(T) e^{K_a T} \tag{22.12}$$

The only random component on the RHS is the supremum of the Wiener process, so we're in business, at least once we take on two standard results, one about the Wiener process itself, the other just about multivariate Gaussians.

Lemma 254 *For a standard Wiener process,* $\mathbb{P}\left(W^*(t) > a\right) = 2\mathbb{P}\left(|W(t)| > a\right)$*.*

Proof: Proposition 13.13 (pp. 256–257) in Kallenberg. □

Lemma 255 *If Z is a d-dimensional standard Gaussian (i.e., mean 0 and covariance matrix I), then*

$$\mathbb{P}\left(|Z| > z\right) \leq \frac{2z^{d-2}e^{-z^2/2}}{2^{d/2}\Gamma(d/2)} \tag{22.13}$$

for sufficiently large z.

Proof: Each component of Z, $Z_i \sim \mathcal{N}(0,1)$. So $|Z| = \sqrt{\sum_{i=1}^{d} Z_i^2}$ has the density function (see, e.g., (Cramér, 1945, sec. 18.1, p. 236))

$$f(z) = \frac{2}{2^{d/2}\sigma^d\Gamma(d/2)} z^{d-1} e^{-\frac{z^2}{2\sigma^2}}$$

This is the d-dimensional Maxwell-Boltzmann distribution, sometimes called the χ-distribution, because $|Z|^2$ is χ^2-distributed with d degrees of freedom. Notice that $\mathbb{P}\left(|Z| \geq z\right) = \mathbb{P}\left(|Z|^2 \geq z^2\right)$, so we will be able to solve this problem in terms of the χ^2 distribution. Specifically, $\mathbb{P}\left(|Z|^2 \geq z^2\right) = \Gamma(d/2, z^2/2)/\Gamma(d/2)$, where $\Gamma(r, a)$ is the upper incomplete gamma function. For said function, for every r, $\Gamma(r,a) \leq a^{r-1}e^{-a}$ for sufficiently large a (Abramowitz and Stegun, 1964, Eq. 6.5.32, p. 263). Hence (for sufficiently large z)

$$\begin{aligned}
\mathbb{P}\left(|Z| \geq z\right) &= \mathbb{P}\left(|Z|^2 \geq z^2\right) && (22.14)\\
&= \frac{\Gamma(d/2, z^2/2)}{\Gamma(d/2)} && (22.15)\\
&\leq \frac{\left(z^2\right)^{d/2-1} 2^{1-d/2} e^{-z^2/2}}{\Gamma(d/2)} && (22.16)\\
&= \frac{2z^{d-2}e^{-z^2/2}}{2^{d/2}\Gamma(d/2)} && (22.17)
\end{aligned}$$

□

Theorem 256 *In the limit as $\epsilon \to 0$, for every $\delta > 0$, $T > 0$,*

$$\log \mathbb{P}\left(\Delta_\epsilon^*(T) > \delta\right) \leq O(\epsilon^{-2}) \tag{22.18}$$

PROOF: Start by directly estimating the probability of the deviation, using preceding lemmas.

$$\begin{aligned}\mathbb{P}\left(\Delta^*_\epsilon(T) > \delta\right) &\leq \mathbb{P}\left(|W|^*(T) > \frac{\delta e^{-K_a T}}{\epsilon}\right) && (22.19)\\ &= 2\mathbb{P}\left(|W(T)| > \frac{\delta e^{-K_a T}}{\epsilon}\right) && (22.20)\\ &\leq \frac{4}{2^{d/2}\Gamma(d/2)}\left(\frac{\delta^2 e^{-2K_a T}}{\epsilon^2}\right)^{d/2-1} e^{-\frac{\delta^2 e^{-2K_a T}}{2\epsilon^2}} && (22.21)\end{aligned}$$

if ϵ is sufficiently small, so that ϵ^{-1} is sufficiently large to apply Lemma 255. Now take the log and multiply through by ϵ^2:

$$\begin{aligned}&\epsilon^2 \log \mathbb{P}\left(\Delta^*_\epsilon(T) > \delta\right) && (22.22)\\ &\leq \epsilon^2 \log \frac{4}{2^{d/2}\Gamma(d/2)} + \epsilon^2\left(\frac{d}{2} - 1\right)\left[\log \delta^2 e^{-2K_a T} - 2\log\epsilon\right] - \delta^2 e^{-2K_a T}\end{aligned}$$

$$\lim_{\epsilon \downarrow 0} \epsilon^2 \log \mathbb{P}\left(\Delta^*_\epsilon(T) > \delta\right) \leq -\delta^2 e^{-2K_a T} \qquad (22.23)$$

since $\epsilon^2 \log \epsilon \to 0$, and the conclusion follows. □

Notice several points.

1. Here ϵ gauges the size of the noise, and we take a small noise limit. In many forms of large deviations theory, we are concerned with large-sample ($N \to \infty$) or long-time ($T \to \infty$) limits. In every case, we will identify some asymptotic parameter, and obtain limits on the asymptotic probabilities. There are deviations inequalities which hold *non*-asymptotically, but they have a different flavor, and require different machinery. (Some people are made uncomfortable by an ϵ^2 rate, and prefer to write the SDE $dX = a(X)dt + \sqrt{\epsilon}dW$ so as to avoid it. I don't get this.)

2. The magnitude of the deviation δ does not change as the noise becomes small. This is basically what makes this a *large* deviations result. There is also a theory of *moderate* deviations, which with any luck we'll be able to at least touch on.

3. We only have an upper bound. This is enough to let us know that the probability of large deviations becomes exponentially small. But we might be wrong about the rate — it could be even faster than we've estimated. In this case, however, it'll turn out that we've got at least the order of magnitude correct.

4. We also don't have a *lower* bound on the probability, which is something that would be *very* useful in doing reliability analyses. It will turn out that, under many circumstances, one can obtain a lower bound on the probability of large deviations, which has the *same* asymptotic dependence on ϵ as the upper bound.

5. Suppose we're right about the rate (which, it will turn out, we are), and it holds both from above and below. It would be nice to be able to say something like

   $$\mathbb{P}\left(\Delta^*_\epsilon(T) > \delta\right) \rightarrow C_1(\delta, T) e^{-C_2(\delta,T)\epsilon^{-2}} \tag{22.24}$$

   rather than

   $$\epsilon^2 \log \mathbb{P}\left(\Delta^*_\epsilon(T) > \delta\right) \rightarrow -C_2(\delta, T) \tag{22.25}$$

   The difficulty with making an assertion like 22.24 is that the large deviation *probability* actually converges on *any* function which goes to asymptotically to zero! So, to extract the actual rate of dependence, we need to get a result like 22.25.

More generally, one consequence of Theorem 256 is that SDE trajectories which are far from the trajectory of the ODE have exponentially small probabilities. The vast majority of the probability will be concentrated around the unperturbed trajectory. Reasonable sample-path functionals can therefore be well-approximated by averaging their value over some small (δ) neighborhood of the unperturbed trajectory. This should sound very similar to Laplace's method for the evaluate of asymptotic integrals in Euclidean space, and in fact one of the key parts of large deviations theory is an extension of Laplace's method to infinite-dimensional function spaces.

In addition to this mathematical content, there is also a close connection to the principle of least action in physics. In classical mechanics, the system follows the trajectory of least action, the "action" of a trajectory being the integral of the kinetic minus potential energy along that path. In quantum mechanics, this is no longer an axiom but a *consequence* of the dynamics: the action-minimizing trajectory is the most probable one, and large deviations from it have exponentially small probability. Similarly, the theory of large deviations can be used to establish quite general *stochastic* principles of least action for Markovian systems.[3]

[3] For a fuller discussion, see Eyink (1996),Freidlin and Wentzell (1998, ch. 3).

Chapter 23

Ergodicity

Section 23.1 gives a general orientation to ergodic theory, which we will study in discrete time.

Section 23.2 introduces dynamical systems and their invariants, the setting in which we will prove our ergodic theorems.

Section 23.3 considers time averages, defines what we mean for a function to have an ergodic property (its time average converges), and derives some consequences.

Section 23.4 defines asymptotic mean stationarity, and shows that, with AMS dynamics, the limiting time average is equivalent to conditioning on the invariant sets.

23.1 General Remarks

To begin our study of ergodic theory, let us consider a famous[1] line from Gnedenko and Kolmogorov (1954, p. 1):

> In fact, all epistemological value of the theory of probability is based on this: that large-scale random phenomena in their collective action create strict, nonrandom regularity.

Now, this is how Gnedenko and Kolmogorov introduced their classic study of the limit laws for *independent* random variables, but most of the random phenomena we encounter around us are not independent. Ergodic theory is a study of how large-scale *dependent* random phenomena nonetheless create nonrandom regularity. The classical limit laws for IID variables $X_1, X_2, \ldots$ assert that, under the right conditions, sample averages converge on expectations,

$$\overline{X}_n \equiv \frac{1}{n} \sum_{i=1}^{n} X_i \rightarrow \mathbf{E}\left[X_i\right]$$

[1] Among mathematical scientists, anyway.

where the sense of convergence can be "almost sure" (strong law of large numbers), "L_p" (p^{th} mean), "in probability" (weak law), etc., depending on the hypotheses we put on the X_i. One meaning of this convergence is that sufficiently large random samples are representative of the entire population — that $\overline{X}_n$ makes a good estimate of $\mathbf{E}[X]$.

The ergodic theorems, likewise, assert that for *dependent* sequences $X_1, X_2, \ldots$, time averages converge on expectations

$$\overline{X}_t \equiv \frac{1}{t}\sum_{i=1}^{t} X_i \rightarrow \mathbf{E}[X_\infty]$$

where X_∞ is some limiting random variable, or in the most useful cases a *non*-random variable. Once again, the mode of convergence will depend on the kind of hypotheses we make about the random sequence X. Once again, the interpretation is that a *single* sample path is representative of the entire distribution over sample paths, *if* it goes on long enough.

Chapter 21 proved a mean-square (L_2) ergodic theorem for weakly stationary continuous-parameter processes. The next few chapters, by contrast, will develop ergodic theorems for non-stationary discrete-parameter processes.[2] This is a little unusual, compared to most probability books, so let me say a word or two about why. (1) Results we get will include stationary processes as special cases, but stationarity fails for many applications where ergodicity (in a suitable sense) holds. So this is more general and more broadly applicable. (2) Our results will all have continuous-time analogs, but the algebra is a lot cleaner in discrete time. (3) Some of the most important applications (for people like you!) are to statistical inference and learning with dependent samples, and to Markov chain Monte Carlo, and both of those are naturally discrete-parameter processes. We will, however, stick to continuous state spaces.

23.2 Dynamical Systems and Their Invariants

It is a very remarkable fact — but one with deep historical roots (von Plato, 1994, ch. 3) — that the way to get regular limits for stochastic processes is to first turn them into irregular deterministic dynamical systems, and then let averaging smooth away the irregularity. This section will begin by laying out dynamical systems, and their invariant sets and functions, which will be the foundation for what follows.

Definition 257 (Dynamical System) *A* dynamical system *consists of a measurable space Ξ, a σ-field $\mathcal{X}$ on Ξ, a probability measure μ defined on $\mathcal{X}$, and a mapping $T : \Xi \mapsto \Xi$ which is $\mathcal{X}/\mathcal{X}$-measurable.*

Remark: Measure-preserving transformations (Definition 53) are special cases of dynamical systems. Since (Theorem 52) every strongly stationary process can

[2] In doing so, I'm ripping off Gray (1988), especially chapters 6 and 7.

be represented by a measure-preserving transformation, namely the shift (Definition 48), the theory of ergodicity for dynamical systems which we'll develop is easily seen to include the usual ergodic theory of strictly-stationary processes as a special case. Thus, at the cost of going to the infinite-dimensional space of sample paths, we can always make it the case that the time-evolution is completely deterministic, and the only stochastic component to the process is its initial condition.

Lemma 258 (Dynamical Systems are Markov Processes) *Let* $\Xi, \mathcal{X}, \mu, T$ *be a dynamical system. Let* $\mathcal{L}(X_1) = \mu$*, and define* $X_t = T^{t-1}X_1$*. Then the* X_t *form a Markov process, with evolution operator* K *defined through* $Kf(x) = f(Tx)$*.*

PROOF: For every $x \in \Xi$ and $B \in \mathcal{X}$, define $\kappa(x, B) \equiv \mathbf{1}_B(Tx)$. For fixed x, this is clearly a probability measure (specifically, the δ measure at Tx). For fixed B, this is a measurable function of x, because T is a measurable mapping. Hence, $\kappa(x, B)$ is a probability kernel. So, by Theorem 103, the X_t form a Markov process. By definition, $\mathbf{E}\left[f(X_1)|X_0 = x\right] = Kf(x)$. But the expectation is in this case just $f(Tx)$. □

Notice that, as a consequence, there is a corresponding operator, call it U, which takes signed measures (defined over $\mathcal{X}$) to signed measures, and specifically takes probability measures to probability measures.

Definition 259 (Observable) *A function* $f : \Xi \mapsto \mathbb{R}$ *which is* $\mathbb{B}/\mathcal{X}$ *measurable is an* observable *of the dynamical system* $\Xi, \mathcal{X}, \mu, T$*.*

Pretty much all of what follows would work if the observables took values in any real or complex vector space, but that situation can be built up from this one.

Definition 260 (Invariant Function, Invariant Set, Invariant Measure) *A function is invariant, under the action of a dynamical system, if* $f(Tx) = f(x)$ *for all* $x \in \Xi$*, or equivalently if* $Kf = f$ *everywhere. An event* $B \in \mathcal{X}$ *is invariant if its indicator function is an invariant function. A measure* ν *is invariant if it is preserved by* T*, i.e. if* $\nu(C) = \nu(T^{-1}C)$ *for all* $C \in \mathcal{X}$*, equivalently if* $U\nu = \nu$*.*

Lemma 261 *The class* $\mathcal{I}$ *of all measurable invariant sets in* Ξ *is a* σ*-algebra.*

PROOF: Clearly, Ξ is invariant. The other properties of a σ-algebra follow because set-theoretic operations (union, complementation, etc.) commute with taking inverse images. □

Lemma 262 *An observable is invariant if and only if it is* $\mathcal{I}$*-measurable. Consequently,* $\mathcal{I}$ *is the* σ*-field generated by the invariant observables.*

Proof: "If": Pick any Borel set B. Since $f = f \circ T$, $f^{-1}(B) = (f \circ T)^{-1}(B) = T^{-1}f^{-1}B$. Hence $f^{-1}(B) \in \mathcal{I}$. Since the inverse image of every Borel set is in $\mathcal{I}$, f is $\mathcal{I}$-measurable. "Only if": Again, pick any Borel set B. By assumption, $f^{-1}(B) \in \mathcal{I}$, so $f^{-1}(B) = T^{-1}f^{-1}(B) = (f \circ T)^{-1}(B)$, so the inverse image of under Tf of any Borel set is an invariant set, implying that $f \circ T$ is $\mathcal{I}$-measurable. Since, for every B, $f^{-1}(B) = (f \circ T)^{-1}(B)$, we must have $f \circ T = f$. The consequence follows. □

Definition 263 (Infinitely Often, i.o.) *For any set $C \in \mathcal{X}$, the set C* infinitely often*, $C_{\mathrm{i.o.}}$, consists of all those points in Ξ whose trajectories visit C infinitely often, $C_{\mathrm{i.o.}} \equiv \limsup_t T^{-t}C$.*

Lemma 264 *For every $C \in \mathcal{X}$, $C_{\mathrm{i.o.}}$ is invariant.*

Proof: Exercise. □

Definition 265 (Invariance Almost Everywhere) *A measurable function is* invariant μ-a.e.*, or* almost invariant*, when*

$$\mu\{x \in \Xi | \forall n,\ f(x) = K^n f(x)\} = 1 \tag{23.1}$$

A measurable set is invariant μ-a.e.*, when its indicator function is almost invariant.*

Remark 1: Some of the older literature annoyingly calls these objects *totally invariant.*

Remark 2: Invariance implies invariance μ-almost everywhere, for any μ.

Lemma 266 *The almost-invariant sets form a σ-field, $\mathcal{I}'$, and an observable is almost invariant if and only if it is measurable with respect to this field.*

Proof: Entirely parallel to that for the strict invariants. □

Let's close this section with a simple lemma, which will however be useful in approximation-by-simple-function arguments in building up expectations.

Lemma 267 *A simple function, $f(x) = \sum_{k=1}^{m} a_m \mathbf{1}_{C_k}(x)$, is invariant if and only if all the sets $C_k \in \mathcal{I}$. Similarly, a simple function is almost invariant iff all the defining sets are almost invariant.*

Proof: Exercise. □

23.3 Time Averages and Ergodic Properties

For convenience, let's re-iterate the definition of a time average. (The notation differs here a little from that given earlier.)

Definition 268 (Time Average) *The time-average of an observable f is the real-valued function*

$$\overline{f}_t(x) \equiv \frac{1}{t}\sum_{i=0}^{t-1} f(T^i x) \tag{23.2}$$

The operator taking functions to their time-averages will be written $A_t f$:

$$A_t f(x) \equiv \overline{f}_t(x) \tag{23.3}$$

Lemma 269 *For every t, the time-average of an observable is an observable.*

Proof: The class of measurable functions is closed under finite iterations of arithmetic operations. □

Definition 270 (Ergodic Property) *An observable f* has the ergodic property *when $\overline{f}_t(x)$ converges as $t \to \infty$ for μ-almost-all x. An observable* has the mean ergodic property *when $\overline{f}_t(x)$ converges in $L_1(\mu)$, and similarly for the other L_p* ergodic properties*. If for some class of functions $\mathcal{D}$, every $f \in \mathcal{D}$ has an ergodic property, then the class $\mathcal{D}$ has that ergodic property.*

Remark. Notice that what is required for f to have the ergodic property is that almost every initial point has *some* limit for its time average,

$$\mu\left\{x \in \Xi \middle| \exists r \in \mathbb{R} : \lim_{t\to\infty} \overline{f}_t(x) = r\right\} = 1 \tag{23.4}$$

Not that there is some *common* limit for almost every initial point,

$$\exists r \in \mathbb{R} : \mu\left\{x \in \Xi \middle| \lim_{t\to\infty} \overline{f}_t(x) = r\right\} = 1 \tag{23.5}$$

Similarly, a class of functions has the ergodic property if all of their time averages converge; they do not have to converge uniformly.

Definition 271 *If an observable f has the ergodic property, define $\overline{f}(x)$ to be the limit of $\overline{f}_t(x)$ where that exists, and* 0 *elsewhere. The corresponding operator will be written A:*

$$Af(x) = \overline{f}(x) \tag{23.6}$$

The domain of A consists of all and only the functions with ergodic properties.

Observe that

$$Af(x) = \lim_{t\to\infty} \frac{1}{t}\sum_{n=0}^{t} K^n f(x) \tag{23.7}$$

That is, A is the limit of an arithmetic mean of conditional expectations. This suggests that it should itself have many of the properties of conditional expectations. In fact, under a reasonable condition, we will see that $Af = \mathbf{E}[f|\mathcal{I}]$, expectation conditional on the σ-algebra of invariant sets. We'll check first that A has the properties we'd want from a conditional expectation.

Lemma 272 *A is a linear operator, and its domain is a linear space.*

PROOF: If c is any real number, then $A_t cf(x) = cA_t f(x)$, and so clearly, if the limit exists, $Acf(x) = cAf(x)$. Similarly, $A_t(f+g)(x) = A_t f(x) + A_t g(x)$, so if f and g both have ergodic properties, then so does $f+g$, and $A(f+g)(x) = Af(x) + Ag(x)$. □

Lemma 273 *If $f \in \mathrm{Dom}A$, and, for all $n \geq 0$, $fT^n \geq 0$ a.e., then $Af(x) \geq 0$ a.e.*

PROOF: The event $Af(x) < 0$ is a sub-event of $\bigcup_n \{f(T^n(x)) < 0\}$. Since the union of a countable collection of measure zero events has measure zero, $Af(x) \geq 0$ almost everywhere. □

We can't just say $f \geq 0$ a.e., because the effect of the transformation T might be to map every point to the bad set of f; the lemma guards against that. Of course, if $f(x) \geq 0$ for all, and not just almost all, x, then the bad set is non-existent, and $Af \geq 0$ follows automatically.

Lemma 274 *The constant function 1 has the ergodic property. Consequently, so does every other constant function.*

PROOF: For every n, $1(T^n x) = 1$. Hence $A_t 1(x) = 1$ for all t, and so $A1(x) = 1$. Extension to other constants follows by linearity. □

Remember that for any Markov operator K, $K1 = 1$.

Lemma 275 *If $f \in \mathrm{Dom}(A)$, then, for all n, $f \circ T^n$ is too, and $Af(x) = Af \circ T^n(x)$. Or, $AK^n f(x) = Af(x)$.*

PROOF: Start with $n = 1$, and show that the discrepancy goes to zero.

$$AKf(x) - Af(x) = \lim_t \frac{1}{t} \sum_{i=0}^{t} \left(K^{i+1}f(x) - K^i f(x)\right) \tag{23.8}$$

$$= \lim_t \frac{1}{t} \left(K^t f(x) - f(x)\right) \tag{23.9}$$

Since $Af(x)$ exists a.e., we know that the series $t^{-1}\sum_{i=0}^{t-1} K^i f(x)$ converges a.e., implying that $(t+1)^{-1}K^t f(x) \to 0$ a.e.. But $t^{-1} = \frac{t+1}{t}(t+1)^{-1}$, and for large t, $t + 1/t < 2$. Hence $(t+1)^{-1}K^t f(x) \leq t^{-1}K^t f(x) \leq 2(t+1)^{-1}K^t f(x)$, implying that $t^{-1}K^t f(x)$ itself goes to zero (a.e.). Similarly, $t^{-1}f(x)$ must go to zero. Thus, overall, we have $AKf(x) = Af(x)$ a.e., and $Kf(x) \in \mathrm{Dom}(A)$. □

Lemma 276 *If $f \in \mathrm{Dom}(A)$, then Af is an invariant, and $\mathcal{I}$-measurable.*

PROOF: Af exists, so (previous lemma) AKf exists and is equal to Af (almost everywhere). But $AKf(x) = Af(Tx)$, by definition, hence Af is invariant, i.e., $KAf = AKf = Af$. Measurability follows from Lemma 262. □

Lemma 277 *If $f \in \mathrm{Dom}(A)$, and B is any set in $\mathcal{I}$, then $A(\mathbf{1}_B(x)f(x)) = \mathbf{1}_B(x)Af(x)$.*

PROOF: For every n, $\mathbf{1}_B(T^n x)f(T^n x) = \mathbf{1}_B(x)f(T^n x)$, since $x \in B$ iff $T^n x \in B$. So, for all finite t, $A_t(\mathbf{1}_B(x)f(x)) = \mathbf{1}_B(x)A_t f(x)$, and the lemma follows by taking the limit. □

Lemma 278 *All indicator functions of measurable sets have ergodic properties if and only if all bounded observables have ergodic properties.*

PROOF: A standard approximation-by-simple-functions argument, as in the construction of Lebesgue integrals. □

Lemma 279 *Let f be bounded and have the ergodic property. Then Af is μ-integrable, and $\mathbf{E}[Af(X)] = \mathbf{E}[f(X)]$, where $\mathcal{L}(X) = \mu$.*

PROOF: Since f is bounded, it is integrable. Hence $A_t f$ is bounded, too, for all t, and $A_t f(X)$ is an integrable random variable. A sequence of bounded, integrable random variables is uniformly integrable. Uniform integrability, plus the convergence $A_t f(x) \rightarrow Af(x)$ for μ-almost-all x, gives us that $\mathbf{E}[Af(X)]$ exists and is equal to $\lim \mathbf{E}[A_t f(X)]$ via Fatou's lemma. (See e.g., Theorem 117 in the notes to 36-752.)

Now use the invariance of Af, i.e., the fact that $Af(X) = Af(TX)$ μ-a.s.

$$\begin{aligned}
0 &= \mathbf{E}[Af(TX)] - \mathbf{E}[Af(X)] && (23.10)\\
&= \lim \frac{1}{t}\sum_{n=0}^{t-1} \mathbf{E}[K^n f(TX)] - \lim \frac{1}{t}\sum_{n=0}^{t-1} \mathbf{E}[K^n f(X)] && (23.11)\\
&= \lim \frac{1}{t}\sum_{n=0}^{t-1} \mathbf{E}[K^n f(TX)] - \mathbf{E}[K^n f(X)] && (23.12)\\
&= \lim \frac{1}{t}\sum_{n=0}^{t-1} \mathbf{E}\left[K^{n+1} f(X)\right] - \mathbf{E}[K^n f(X)] && (23.13)\\
&= \lim \frac{1}{t}\left(\mathbf{E}\left[K^t f(X)\right] - \mathbf{E}[f(X))\right] && (23.14)
\end{aligned}$$

Hence

$$\mathbf{E}[Af] = \lim \frac{1}{t}\sum_{n=0}^{t-1} \mathbf{E}[K^n f(X)] = \mathbf{E}[f(X)] \qquad (23.15)$$

as was to be shown. □

Lemma 280 *If f is as in Lemma 279, then $A_t f \rightarrow f$ in $L_1(\mu)$.*

PROOF: From Lemma 279, $\lim \mathbf{E}[A_t f(X)] = \mathbf{E}[f(X)]$. Since the variables $A_t f(X)$ are uniformly integrable (as we saw in the proof of that lemma), it follows (Proposition 4.12 in Kallenberg, p. 68) that they also converge in $L_1(\mu)$. □

Lemma 281 *Let f be as in Lemmas 279 and 280, and $B \in \mathcal{X}$ be an arbitrary measurable set. Then*

$$\lim_{t\rightarrow\infty} \frac{1}{t}\sum_{n=0}^{t-1} \mathbf{E}\left[\mathbf{1}_B(X) K^n f(X)\right] = \mathbf{E}\left[\mathbf{1}_B(X) f(X)\right] \tag{23.16}$$

where $\mathcal{L}\left(X\right) = \mu$.

Proof: Let's write out the expectations explicitly as integrals.

$$\begin{aligned}
& \left|\int_B f(x) d\mu - \frac{1}{t}\sum_{n=0}^{t-1}\int_B K^n f(x) d\mu\right| & (23.17)\\
= & \left|\int_B f(x) - \frac{1}{t}\sum_{n=0}^{t-1} K^n f(x) d\mu\right| & \\
= & \left|\int_B f(x) - A_t f(x) d\mu\right| & (23.18)\\
\leq & \int_B \left|f(x) - A_t f(x)\right| d\mu & (23.19)\\
\leq & \int \left|f(x) - A_t f(x)\right| d\mu & (23.20)\\
= & \left\|f - A_t f\right\|_{L_1(\mu)} & (23.21)
\end{aligned}$$

But (previous lemma) these functions converge in $L_1(\mu)$, so the limit of the norm of their difference is zero. □

Boundedness is not essential.

Corollary 282 *Lemmas 279, 280 and 281 hold for any integrable observable $f \in \mathrm{Dom}(A)$, bounded or not, provided that $A_t f$ is a uniformly integrable sequence.*

Proof: Examining the proofs shows that the boundedness of f was important only to establish the uniform integrability of $A_t f$. □

23.4 Asymptotic Mean Stationarity

Next, we come to an important concept which will prove to be necessary and sufficient for the most important ergodic properties to hold.

Definition 283 (Asymptotically Mean Stationary) *A dynamical system is* asymptotically mean stationary *when, for every $C \in \mathcal{X}$, the limit*

$$m(C) \equiv \lim_{t\rightarrow\infty}\frac{1}{t}\sum_{n=0}^{t-1}\mu(T^{-n}C) \tag{23.22}$$

exists, and the set function m is its stationary mean.

Remark 1: It might've been more logical to call this "asymptotically measure stationary", or something, but I didn't make up the names...

Remark 2: Symbolically, we can write

$$m = \lim_{t\rightarrow\infty} \frac{1}{t}\sum_{n=0}^{t-1} U^n \mu$$

where U is the operator taking measures to measures. This leads us to the next proposition.

Proposition 284 *If a dynamical system is stationary, i.e., T is preserves the measure μ, then it is asymptotically mean stationary, with stationary mean μ.*

Proof: If T preserves μ, then for every measurable set, $\mu(C) = \mu(T^{-1}C)$. Hence every term in the sum in Eq. 23.22 is $\mu(C)$, and consequently the limit exists and is equal to $\mu(C)$. □

Theorem 285 (Vitali-Hahn Theorem) *If m_t are a sequence of probability measures on a common σ-algebra $\mathcal{X}$, and $m(C)$ is a set function such that $\lim_t m_t(C) = m(C)$ for all $C \in \mathcal{X}$, then m is a probability measure on $\mathcal{X}$.*

Proof: This is a standard result from measure theory. □

Theorem 286 *If a dynamical system is asymptotically mean stationary, then its stationary mean is an invariant probability measure.*

Proof: For every t, let $m_t(C) = \frac{1}{t}\sum_{n=0}^{t-1}\mu(T^{-n}(C))$. Then m_t is a linear combination of probability measures, hence a probability measure itself. Since, for every $C \in \mathcal{X}$, $\lim m_t(C) = m(C)$, by Definition 283, Proposition 285 says that $m(C)$ is also a probability measure. It remains to check invariance.

$$\begin{aligned} & m(C) - m(T^{-1}C) & (23.23)\\ = & \lim \frac{1}{t}\sum_{n=0}^{t-1}\mu(T^{-n}(C)) - \lim \frac{1}{t}\sum_{n=0}^{t-1}\mu(T^{-n}(T^{-1}C)) & \\ = & \lim \frac{1}{t}\sum_{n=0}^{t-1}\mu(T^{-n-1}C) - \mu(T^{-n}C) & (23.24)\\ = & \lim \frac{1}{t}\left(\mu(T^{-t}C) - \mu(C)\right) & (23.25)\end{aligned}$$

Since the probability measure of any set is at most 1, the difference between two probabilities is at most 1, and so $m(C) = m(T^{-1}C)$, for all $C \in \mathcal{X}$. But this means that m is invariant under T (Definition 53). □

Remark: Returning to the symbolic manipulations, if μ is AMS with stationary mean m, then $Um = m$ (because m is invariant), and so we can write $\mu = m + (\mu - m)$, knowing that $\mu - m$ goes to zero under averaging. Speaking loosely (this can be made precise, at the cost of a fairly long excursion) m is an eigenvector of U (with eigenvalue 1), and $\mu - m$ lies in an orthogonal direction, along which U is contracting, so that, under averaging, it goes away, leaving only m, which is like the projection of the original measure μ on to the invariant manifold of U.

The relationship between an AMS measure μ and its stationary mean m is particularly simple on invariant sets: they are equal there. A slightly more general theorem is actually just as easy to prove, however, so we'll do that.

Lemma 287 *If μ is AMS with limit m, and f is an observable which is invariant μ-a.e., then $\mathbf{E}_{\mu}\left[f\right] = \mathbf{E}_{m}\left[f\right]$.*

Proof: Let C be any almost invariant set. Then, for any t, C and $T^{-t}C$ differ by, at most, a set of μ-measure 0, so that $\mu(C) = \mu(T^{-t}C)$. The definition of the stationary mean (Equation 23.22) then gives $\mu(C) = m(C)$, or $\mathbf{E}_{\mu}\left[\mathbf{1}_C\right] = \mathbf{E}_{m}\left[\mathbf{1}_C\right]$, i.e., the result holds for indicator functions. By Lemma 267, this then extends to simple functions. The usual arguments then take us to all functions which are measurable with respect to $\mathcal{I}'$, the σ-field of almost-invariant sets, but this (Lemma 266) is the class of all almost-invariant functions. □

Lemma 288 *If μ is AMS with stationary mean m, and f is a bounded observable,*

$$\lim_{t\rightarrow\infty} \mathbf{E}_{\mu}\left[A_t f\right] = \mathbf{E}_{m}\left[f\right] \tag{23.26}$$

Proof: By Eq. 23.22, this must hold when f is an indicator function. By the linearity of A_t and of expectations, it thus holds for simple functions, and so for general measurable functions, using boundedness to exchange limits and expectations where necessary. □

Lemma 289 *If f is a bounded observable in* $\mathrm{Dom}(A)$*, and μ is AMS with stationary mean m, then $\mathbf{E}_{\mu}\left[Af\right] = \mathbf{E}_{m}\left[f\right]$.*

Proof: From Lemma 281, $\mathbf{E}_{\mu}\left[Af\right] = \lim_{t\rightarrow\infty} \mathbf{E}_{\mu}\left[A_t f\right]$. From Lemma 288, the latter is $\mathbf{E}_{m}\left[f\right]$. □

Remark: Since Af is invariant, we've got $\mathbf{E}_{\mu}\left[Af\right] = \mathbf{E}_{m}\left[Af\right]$, from Lemma 287, but that's not the same.

Corollary 290 *Lemmas 288 and 289 continue to hold if f is not bounded, but $A_t f$ is uniformly integrable (μ).*

Proof: As in Corollary 282. □

Theorem 291 *If μ is AMS, with stationary mean m, and the dynamics have ergodic properties for all the indicator functions, then, for any measurable set C,*

$$A\mathbf{1}_C = m(C|\mathcal{I}) \tag{23.27}$$

with probability 1 under both μ and m.

Proof: By Lemma 276, $A\mathbf{1}_C$ is an invariant function. Pick any set $B \in \mathcal{I}$, so that $\mathbf{1}_B$ is also invariant. By Lemma 277, $A(\mathbf{1}_B\mathbf{1}_C) = \mathbf{1}_B A\mathbf{1}_C$, which is invariant (as a product of invariant functions). So Lemma 287 gives

$$\mathbf{E}_{\mu}\left[\mathbf{1}_B A\mathbf{1}_C\right] = \mathbf{E}_m\left[\mathbf{1}_B A\mathbf{1}_C\right] \tag{23.28}$$

while Lemma 289 says

$$\mathbf{E}_{\mu}\left[A(\mathbf{1}_B\mathbf{1}_C)\right] = \mathbf{E}_m\left[\mathbf{1}_B\mathbf{1}_C\right] \tag{23.29}$$

Since the left-hand sides are equal, the right-hand sides must be equal as well, so

$$\begin{aligned} m(B \cap C) &= \mathbf{E}_m\left[\mathbf{1}_B\mathbf{1}_C\right] && (23.30)\\ &= \mathbf{E}_m\left[\mathbf{1}_B A\mathbf{1}_C\right] && (23.31) \end{aligned}$$

Since this holds for all invariant sets $B \in \mathcal{I}$, we conclude that $A\mathbf{1}_C$ must be a version of the conditional probability $m(C|\mathcal{I})$. □

Corollary 292 *Under the assumptions of Theorem 291, for any bounded observable f,*

$$Af = \mathbf{E}_m\left[f|\mathcal{I}\right] \tag{23.32}$$

Proof: From Lemma 278, every bounded observable has the ergodic property. One can then imitate the proof of the theorem to obtain the desired result. □

Corollary 293 *Equation 23.32 continues to hold if $A_t f$ are uniformly μ-integrable, or f is m-integrable.*

Proof: Exercise. □

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,*

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

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

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

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

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.

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

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

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

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,

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

Since ϵ can be made arbitrarily small, we conclude that

$$\mathbf{E}\left[\overline{f}\right] \leq \mathbf{E}\left[f\right] \tag{24.32}$$

as was to be shown.

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

Since $\mathbf{E}\left[\underline{f}\right] \geq \mathbf{E}\left[f\right] \geq \mathbf{E}\left[\overline{f}\right]$, 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, $Af = \mathbf{E}_m\left[f|\mathcal{I}\right]$, 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 $\mathbf{E}\left[\overline{f}_M\right] \leq \mathbf{E}\left[f\right]$. Then let $M \rightarrow \infty$, and apply the monotone convergence theorem to conclude that $\mathbf{E}\left[\overline{f}\right] \leq \mathbf{E}\left[f\right]$; the rest of the proof goes through as before. □

Chapter 25

Ergodicity

This lecture explains what it means for a process to be ergodic or metrically transitive, gives a few characterizes of these properties (especially for AMS processes), and deduces some consequences. The most important one is that sample averages have deterministic limits.

25.1 Ergodicity and Metric Transitivity

Definition 300 *A dynamical system* $\Xi, \mathcal{X}, \mu, T$ *is* ergodic*, or* an ergodic system *or* an ergodic process *when* $\mu(C) = 0$ *or* $\mu(C) = 1$ *for every* T*-invariant set* C*.* μ *is called a* T-ergodic measure*, and* T *is called a* μ*-ergodic transformation, or just an* ergodic measure *and* ergodic transformation*, respectively.*

Remark: Most authorities require a μ-ergodic transformation to also be measure-preserving for μ. But (Corollary 54) measure-preserving transformations are necessarily stationary, and we want to minimize our stationarity assumptions. So what most books call "ergodic", we have to qualify as "stationary and ergodic". (Conversely, when other people talk about processes being "stationary and ergodic", they mean "stationary with only one ergodic component"; but of that, more later.

Definition 301 *A dynamical system is* metrically transitive*,* metrically indecomposable*, or* irreducible *when, for any two sets* $A, B \in \mathcal{X}$*, if* $\mu(A), \mu(B) > 0$*, there exists an* n *such that* $\mu(T^{-n}A \cap B) > 0$*.*

Remark: In dynamical systems theory, metric transitivity is contrasted with *topological* transitivity: T is topologically transitive on a domain D if for any two open sets $U, V \subseteq D$, the images of U and V remain in D, and there is an n such that $T^n U \cap V \neq \emptyset$. (See, e.g., Devaney (1992).) The "metric" in "metric transitivity" refers not to a distance function, but to the fact that a measure is involved. Under certain conditions, metric transitivity in fact implies topological transitivity: e.g., if D is a subset of a Euclidean space and μ has a positive density with respect to Lebesgue measure. The converse is not generally true, however: there are systems which are transitive topologically but not metrically.

A dynamical system is *chaotic* if it is topologically transitive, and it contains dense periodic orbits (Banks *et al.*, 1992). The two facts together imply that a trajectory can start out arbitrarily close to a periodic orbit, and so remain near it for some time, only to eventually find itself arbitrarily close to a *different* periodic orbit. This is the source of the fabled "sensitive dependence on initial conditions", which paradoxically manifests itself in the fact that all typical trajectories look pretty much the same, at least in the long run. Since metric transitivity generally implies topological transitivity, there is a close connection between ergodicity and chaos; in fact, most of the well-studied chaotic systems are also ergodic (Eckmann and Ruelle, 1985), including the logistic map. However, it is possible to be ergodic without being chaotic: the one-dimensional rotations with irrational shifts are, because there periodic orbits do not exist, and *a fortiori* are not dense.

Proposition 302 *A dynamical system is ergodic if it is metrically transitive.*

Proof: By contradiction. Suppose there was an invariant set A whose μ-measure was neither 0 nor 1; then A^c is also invariant, and has strictly positive measure. By metric transitivity, for some n, $\mu(T^{-n}A \cap A^c) > 0$. But $T^{-n}A = A$, and $\mu(A \cap A^c) = 0$. So metrically transitive systems are ergodic. $\square$

There is a partial converse.

Proposition 303 *If a dynamical systems is ergodic and stationary, then it is metrically transitive.*

Proof: Take any $\mu(A), \mu(B) > 0$. Let $A_{\mathrm{ever}} \equiv \bigcup_{n=0}^{\infty} T^{-n}A$ — the union of A with all its pre-images. This set contains its pre-images, $T^{-1}A_{\mathrm{ever}} \subseteq A_{\mathrm{ever}}$, since if $x \in T^{-n}A$, $T^{-1}x \in T^{-n-1}A$. The sequence of pre-images is thus non-increasing, and so tends to a limiting set, $\bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} T^{-k}A = A_{\mathrm{i.o.}}$, the set of points which not only visit A eventually, but visit A infinitely often. This is an invariant set (Lemma 264), so by ergodicity it has either measure 0 or measure 1. By the Poincaré recurrence theorem (Corollaries 66 and 67), since $\mu(A) > 0$, $\mu(A_{\mathrm{i.o.}}) = 1$. Hence, for any B, $\mu(A_{\mathrm{i.o.}} \cap B) = \mu(B)$. But this means that, for some n, $\mu(T^{-n}A \cap B) > 0$, and the process is metrically transitive. $\square$

Theorem 304 *A T transformation is μ-ergodic if and only if all T-invariant observables are constant μ-almost-everywhere.*

Proof: "Only if": Because invariant observables are $\mathcal{I}$-measurable (Lemma 262), the pre-image under an invariant observable f of any Borel set B is an invariant set. Since every invariant set has μ-probability 0 or 1, the probability that $f(x) \in B$ is either 0 or 1, hence f is constant with probability 1. "If": The indicator function of an invariant set is an invariant function. If all invariant functions are constant μ-a.s., then for any $A \in \mathcal{I}$, either $\mathbf{1}_A(x) = 0$ or $\mathbf{1}_A(x) = 1$ for μ-almost all x, which is the same as saying that either $\mu(A) = 0$ or $\mu(A) = 1$, as required. □

Lemma 305 *If μ is T-ergodic, and μ is AMS with stationary mean m, then*

$$\lim_{t \to \infty} \frac{1}{t} \sum_{n=0}^{t-1} \mu(B \cap T^{-n} C) = \mu(B) m(C) \tag{25.1}$$

for any measurable events B, C.

Proof: Exercise 25.1. □

Theorem 306 *Suppose $\mathcal{X}$ is generated by a field $\mathcal{F}$. Then an AMS measure μ, with stationary mean m, is ergodic if and only if, for all $F \in \mathcal{F}$,*

$$\lim_{t \to \infty} \frac{1}{t} \sum_{n=0}^{t-1} \mu(F \cap T^{-n} F) = \mu(F) m(F) \tag{25.2}$$

i.e., iff Eq. 25.1 holds, taking $B = C = F \in \mathcal{F}$.

Proof: "Only if": Lemma 305. "If": Exercise 25.2. □

25.1.1 Examples of Ergodicity

Example 307 (IID Sequences, Strong Law of Large Numbers) *Every IID sequence is ergodic. This is because the Kolmogorov 0-1 law states that every tail event has either probability 0 or 1, and (Exercise 25.3) every invariant event is a tail event. The strong law of large numbers is thus a two-line corollary of the Birkhoff ergodic theorem.*

Example 308 (Markov Chains) *In the elementary theory of Markov chains, an* ergodic *chain is one which is irreducible, aperiodic and positive recurrent. To see that such a chain corresponds to an ergodic process in the present sense, look at the shift operator on the sequence space. For consistency of notation, let $S_1, S_2, \ldots$ be the values of the Markov chain in Σ, and X be the semi-infinite sequence in sequence space Ξ, with shift operator T, and distribution μ over sequences. μ is the product of an initial distribution $\nu \sim S_1$ and the Markov-family kernel. Now, "irreducible" means that one goes from every state to every other state with positive probability at some lag, i.e., for every $s_1, s_2 \in \Sigma$, there is an n such that $\mathbb{P}(S_n = s_2 | S_1 = s_1) > 0$. But, writing $[s]$ for the cylinder set in Ξ with base s, this means that, for every $[s_1]$, $[s_2]$, $\mu(T^{-n}[s_2] \cap [s_1]) > 0$, provided $\mu([s_1]) > 0$. The Markov property of the S chain, along with positive recurrence, can be used to extend this to all finite-dimensional cylinder sets (Exercise 25.4), and so, by a generating-class argument, to all measurable sets.*

Example 309 (Deterministic Ergodicity: The Logistic Map) *We have seen that the logistic map, $Tx = 4x(1-x)$, has an invariant density (with respect to Lebesgue measure). It has an infinite collection of invariant sets, but the only invariant interval is the whole state space $[0,1]$ — any smaller interval is not invariant. From this, it is easy to show that all the invariant sets either have measure 0 or measure 1 — they differ from $\emptyset$ or from $[0,1]$ by only a countable collection of points. Hence, the invariant measure is ergodic. Notice, too, that the Lebesgue measure on $[0,1]$ is ergodic, but not invariant.*

Example 310 (Invertible Ergodicity: Rotations) *Let $\Xi = [0,1)$, $Tx = x + \phi \bmod 1$, and let μ be the Lebesgue measure on Ξ. (This corresponds to a rotation, where the angle advances by $2\pi\phi$ radians per unit time.) Clearly, T preserve μ. If ϕ is rational, then, for any x, the sequence of iterates will visit only finitely many points, and the process is not ergodic, because one can construct invariant sets whose measure is neither 0 nor 1. (You may construct such a set by taking any one of the periodic orbits, and surrounding its points by internals of sufficiently small, yet positive, width.) If, on the other hand, ϕ is irrational, then $T^n x$ never repeats, and it is easy to show that the process is ergodic, because it is metrically transitive. Nonetheless, T is invertible.*

This example (suitably generalized to multiple coordinates) is very important in physics, because many mechanical systems can be represented in terms of "action-angle" variables, the speed of rotation of the angular variables being set by the actions, which are conserved, energy-like quantities. See Mackey (1992); Arnol'd and Avez (1968) for the ergodicity of rotations and its limitations, and Arnol'd (1978) for action-angle variables. Astonishingly, the result for the one-dimensional case was proved by Nicholas Oresme in the 14th century (von Plato, 1994).

Example 311 *Ergodicity does not ensure a uni-directional evolution of the density. (Some people (Mackey, 1992) believe this has great bearing on the foundations of thermodynamics.) For a particularly extreme example, which also illustrates why elementary Markov chain theory insists on aperiodicity, consider the period-two* deterministic *chain, where state A goes to stae B with probability 1, and vice versa. Every sample path spends just much time in state A as in state B, so every time average will converge on $\mathbf{E}_m\left[f\right]$, where m puts equal probability on both states. It doesn't matter what initial distribution we use, because they are all ergodic (the only invariant sets are the whole space and the empty set, and every distribution gives them probability 1 and 0, respectively). The uniform distribution is the unique* stationary *distribution, but other distributions do not approch it, since $U^{2n}\nu = \nu$ for every integer n. So, $A_t f \rightarrow \mathbf{E}_m\left[f\right]$ a.s., but $\mathcal{L}\left(X_n\right) \not\rightarrow m$. We will see later that aperiodicity of Markov chains connects to "mixing" properties, which do guarantee stronger forms of distributional convergence.*

25.1.2 Consequences of Ergodicity

The most basic consequence of ergodicity is that time-averages converge to deterministic, rather than random, limits.

Theorem 312 *Suppose μ is AMS, with stationary mean m, and T-ergodic. Then, almost surely,*

$$\lim_{t \rightarrow \infty} A_t f(x) = \mathbf{E}_m\left[f\right] \tag{25.3}$$

for μ- and m- almost all x, for any $L_1(m)$ observable f.

PROOF: Because every invariant set has μ-probability 0 or 1, it likewise has m-probability 0 or 1 (Lemma 287). Hence, $\mathbf{E}_m\left[f\right]$ is a version of $\mathbf{E}_m\left[f|\mathcal{I}\right]$. Since $A_t f$ is also a version of $\mathbf{E}_m\left[f|\mathcal{I}\right]$ (Corollary 299), they are equal almost surely. □

An important consequence is the following. Suppose S_t is a strictly stationary random sequence. Let $\Phi_t(S) = f(S_{t+\tau_1}, S_{t+\tau_2}, \ldots S_{t+\tau_n})$ for some fixed collection of shifts τ_n. Then Φ_t is another strictly stationary random sequence. Every strictly stationary random sequence can be represented by a measure-preserving transformation (Theorem 52), where X is the sequence $S_1, S_2, \ldots$, the mapping T is just the shift, and the measure μ is the infinite-dimensional measure of the original stochastic process. Thus $\Phi_t = \phi(X_t)$, for some measurable function ϕ. If the measure is ergodic, and $\mathbf{E}\left[\Phi\right]$ is finite, then the time-average of Φ converges almost surely to its expectation. In particular, let $\Phi_t = S_t S_{t+\tau}$. Then, assuming the mixed moments are finite, $t^{-1}\sum_{t=1}^{\infty} S_t S_{t+\tau} \rightarrow \mathbf{E}\left[S_t S_{t+\tau}\right]$ almost surely, and so the sample covariance converges on the true covariance. More generally, for a stationary ergodic process, if the n-point correlation functions exist, the sample correlation functions converge a.s. on the true correlation functions.

25.2 Preliminaries to Ergodic Decompositions

It is always the case, with a dynamical system, that if x lies within some invariant set A, then all its future iterates stay within A as well. In general, therefore, one might expect to be able to make some predictions about the future trajectory by knowing which invariant sets the initial condition lies within. An ergodic process is one where this is actually not possible. Because all invariants sets have probability 0 or 1, they are all independent of each other, and indeed of every other set. Therefore, knowing which invariant sets x falls into is *completely uninformative* about its future behavior. In the more general non-ergodic case, a limited amount of prediction is however possible on this basis, the limitations being set by the way the state space breaks up into invariant sets of points with the same long-run average behavior — the ergodic components. Put slightly differently, the long-run behavior of an AMS system can be represented as a mixture of stationary, ergodic distributions, and the ergodic components are, in a sense, a minimal parametrically sufficient statistic for this distribution. (They are not in generally *predictively* sufficient.)

The idea of an ergodic decomposition goes back to von Neumann, but was considerably refined subsequently, especially by the Soviet school, who seem to have introduced most of the talk of predictions, and all of the talk of ergodic components as minimal sufficient statistics. Our treatment will follow Gray (1988, ch. 7), and Dynkin (1978). The rest of this lecture will handle some preliminary propositions about combinations of stationary measures.

Proposition 313 *Any convex combination of invariant probability measures is an invariant probability measure.*

Proof: Let μ_1 and μ_2 be two invariant probability measures. It is elementary that for every $0 \leq a \leq 1$, $\nu \equiv a\mu_1 + (1-a)\mu_2$ is a probability measure. Now consider the measure under ν of the pre-image of an arbitrary measurable set $B \in \mathcal{X}$:

$$\begin{aligned} \nu(T-1B) &= a\mu_1(T^{-1}B) + (1-a)\mu_2(T^{-1}B) && (25.4)\\ &= a\mu_1(B) + (1-a)\mu_2(B) && (25.5)\\ &= \nu(B) && (25.6) \end{aligned}$$

so ν is also invariant. □

Proposition 314 *If μ_1 and μ_2 are invariant ergodic measures, then either $\mu_1 = \mu_2$, or they are singular, meaning that there is a set B on which $\mu_1(B) = 0$, $\mu_2(B) = 1$.*

Proof: Suppose $\mu_1 \neq \mu_2$. Then there is at least one set C where $\mu_1(C) \neq \mu_2(C)$. Because both μ_i are stationary and ergodic, $A_t\mathbf{1}_C(x)$ converges to $\mu_i(C)$ for μ_i-almost-all x. So the set

$$\left\{x | \lim_t A_t\mathbf{1}_C(x) = \mu_2(C)\right\}$$

has a μ_2 measure of 1, and a μ_1 measure of 0 (since, by hypothesis, $\mu_1(C) \neq \mu_2(C)$. □

Proposition 315 *Ergodic invariant measures are extremal points of the convex set of invariant measures, i.e., they cannot be written as combinations of other invariant measures.*

Proof: By contradiction. That is, suppose μ is ergodic and invariant, and that there were invariant measures ν and λ, and an $a \in (0,1)$, such that $\mu = a\nu + (1-a)\lambda$. Let C be any invariant set; then $\mu(C) = 0$ or $\mu(C) = 1$. Suppose $\mu(C) = 0$. Then, because a is strictly positive, it must be the case that $\nu(C) = \lambda(C) = 0$. If $\mu(C) = 1$, then C^c is also invariant and has μ-measure 0, so $\nu(C^c) = \lambda(C^c) = 0$, i.e., $\nu(C) = \lambda(C) = 1$. So ν and λ would both have to be ergodic, with the same support as μ. But then (Proposition 314 preceeding) $\lambda = \nu = \mu$. □

Remark: The converse is left as an exercise (25.5).

25.3 Exercises

Exercise 25.1 *Prove Lemma 305.*

Exercise 25.2 *Prove the "if" part of Theorem 306.*

Exercise 25.3 *Prove that every invariant event is a tail event. Does the converse hold?*

Exercise 25.4 *Complete the argument in Example 308, proving that ergodic Markov chains are ergodic processes (in the sense of Definition 300).*

Exercise 25.5 *Prove the converse to Proposition 315: every extermal point of the convex set of invariant measures is an ergodic measure.*

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

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

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, $\mathbb{P}_\theta\left(A|\sigma(\{C_i\})\right) = \lambda(\phi(x), A)$ 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 $\mathcal{L}\left(X_t\right)$ 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 $\mu_0(R_t) \rightarrow 0$ while $\mu_1(R_t) \rightarrow 1$.*

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 $\mu_i(R_t) \rightarrow \mu_i(R)$. 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 $\mu_i(R_t) \rightarrow \mu_i(R) = i$. □

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 $t \rightarrow \infty$, 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 $\mathbb{P}\left(X_1^t \in R_{t,i}\right) \rightarrow 1$ 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. □

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, $T^{-t}\left(B_2 \setminus B_1\right) = T^{-t}B_2 \setminus T^{-t}B_1$. Thus

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

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

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

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.

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

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 $Kf(x) = \mathbf{E}\left[f(X_1)|X_0 = x\right]$. 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, $\mathbf{E}_{\mu}\left[f(X)Kg(X)\right] = \mathbf{E}\left[Uf(X)g(X)\right]\mu$, 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,*

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

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*

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

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 conditions. 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,

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

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

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

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

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

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

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

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 $\mathcal{L}\left(S_n/\sqrt{n}\right)$ 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.

Chapter 28

Shannon Entropy and Kullback-Leibler Divergence

Section 28.1 introduces Shannon entropy and its most basic properties, including the way it measures how close a random variable is to being uniformly distributed.

Section 28.2 describes relative entropy, or Kullback-Leibler divergence, which measures the discrepancy between two probability distributions, and from which Shannon entropy can be constructed. Section 28.2.1 describes some statistical aspects of relative entropy, especially its relationship to expected log-likelihood and to Fisher information.

Section 28.3 introduces the idea of the mutual information shared by two random variables, and shows how to use it as a measure of serial dependence, like a nonlinear version of autocovariance (Section 28.3.1).

Information theory studies stochastic processes as sources of information, or as models of communication channels. It appeared in essentially its modern form with Shannon (1948), and rapidly proved to be an extremely useful mathematical tool, not only for the study of "communication and control in the animal and the machine" (Wiener, 1961), but more technically as a vital part of probability theory, with deep connections to statistical inference (Kullback, 1968), to ergodic theory, and to large deviations theory. In an introduction that's so limited it's almost a crime, we will do little more than build enough theory to see how it can fit in with the theory of inference, and then get what we need to progress to large deviations. If you want to learn more (and you should!), the deservedly-standard modern textbook is Cover and Thomas (1991), and a good treatment, at something more like our level of mathematical rigor, is Gray

(1990).[1]

28.1 Shannon Entropy

The most basic concept of information theory is that of the *entropy* of a random variable, or its distribution, often called Shannon entropy to distinguish it from the many other sorts. This is a measure of the uncertainty or variability associated with the random variable. Let's start with the discrete case, where the variable takes on only a finite or countable number of values, and everything is easier.

Definition 356 (Shannon Entropy (Discrete Case)) *The* Shannon entropy*, or just* entropy*, of a discrete random variable X is*

$$H[X] \equiv -\sum_{x} \mathbb{P}(X = x) \log \mathbb{P}(X = x) = -\mathbf{E}[\log \mathbb{P}(X)] \tag{28.1}$$

when the sum exists. Entropy has units of bits *when the logarithm has base 2, and* nats *when it has base e.*

The joint entropy of two random variables, $H[X,Y]$, is the entropy of their joint distribution.

The conditional entropy of X given Y, $H[X|Y]$ is

$$\begin{aligned} H[X|Y] &\equiv \sum_{y} \mathbb{P}(Y = y) \sum_{x} \mathbb{P}(X = x|Y = y) \log \mathbb{P}(X = x|Y = y) && (28.2)\\ &= -\mathbf{E}[\log \mathbb{P}(X|Y)] && (28.3)\\ &= H[X,Y] - H[Y] && (28.4) \end{aligned}$$

Here are some important properties of the Shannon entropy, presented without proofs (which are not hard).

1. $H[X] \geq 0$
2. $H[X] = 0$ iff $\exists x_0 : \ X = x_0$ a.s.
3. If X can take on $n < \infty$ different values (with positive probability), then $H[X] \leq \log n$. $H[X] = \log n$ iff X is uniformly distributed.
4. $H[X]+H[Y] \geq H[X,Y]$, with equality iff X and Y are independent. (This comes from the logarithm in the definition.)

[1] Remarkably, almost all of the post-1948 development has been either amplifying or refining themes first sounded by Shannon. For example, one of the fundamental results, which we will see in the next chapter, is the "Shannon-McMillan-Breiman theorem", or "asymptotic equipartition property", which says roughly that the log-likelihood per unit time of a random sequence converges to a constant, characteristic of the data-generating process. Shannon's original version was convergence in probability for ergodic Markov chains; the modern form is almost sure convergence for any stationary and ergodic process. Pessimistically, this says something about the decadence of modern mathematical science; optimistically, something about the value of getting it right the first time.

5. $H[X,Y] \geq H[X]$.

6. $H[X|Y] \geq 0$, with equality iff X is a.s. constant given Y, for almost all Y.

7. $H[X|Y] \leq H[X]$, with equality iff X is independent of Y. ("Conditioning reduces entropy".)

8. $H[f(X)] \leq H[X]$, for any measurable function f, with equality iff f is invertible.

The first three properties can be summarized by saying that $H[X]$ is maximized by a uniform distribution, and minimized, to zero, by a degenerate one which is a.s. constant. We can then think of $H[X]$ as the variability of X, something like the log of the effective number of values it can take on. We can also think of it as how uncertain we are about X's value.[2] $H[X,Y]$ is then how much variability or uncertainty is associated with the pair variable X, Y, and $H[Y|X]$ is how much uncertainty remains about Y once X is known, averaging over Y. Similarly interpretations follow for the other properties. The fact that $H[f(X)] = H[X]$ if f is invertible is nice, because then f just relabels the possible values, meshing nicely with this interpretation.

A simple consequence of the above results is particularly important for later use.

Lemma 357 (Chain Rule for Shannon Entropy) *Let $X_1, X_2, \ldots X_n$ be discrete-valued random variables on a common probability space. Then*

$$H[X_1, X_2, \ldots X_n] = H[X_1] + \sum_{i=2}^{n} H[X_n|X_1, \ldots X_{n-1}] \tag{28.5}$$

PROOF: From the definitions, it is easily seen that $H[X_2|X_1] = H[X_2, X_1] - H[X_1]$. This establishes the chain rule for $n = 2$. A simple argument by induction does the rest. □

For non-discrete random variables, it is necessary to introduce a reference measure, and many of the nice properties go away.

Definition 358 (Shannon Entropy (General Case)) *The* Shannon entropy *of a random variable X with distribution μ, with respect to a reference measure ρ, is*

$$H_\rho[X] \equiv -\mathbf{E}_\mu \left[\log \frac{d\mu}{d\rho}\right] \tag{28.6}$$

[2] This line of reasoning is sometimes supplemented by saying that we are more "surprised" to find that $X = x$ the less probable that event is, supposing that surprise should go as the log of one over that probability, and defining entropy as expected surprise. The choice of the logarithm, rather than any other increasing function, is of course retroactive, though one might cobble together some kind of psychophysical justification, since the perceived intensity of a sensation often grows logarithmically with the physical magnitude of the stimulus. More dubious, to my mind, is the idea that there is any surprise *at all* when a fair coin coming up heads.

when $\mu << \rho$. Joint and conditional entropies are defined similarly. We will also write $H_\rho[\mu]$, with the same meaning. This is sometimes called differential entropy *when ρ is Lebesgue measure on Euclidean space, especially $\mathbb{R}$, and then is written $h(X)$ or $h[X]$.*

It remains true, in the general case, that $H_\rho[X|Y] = H_\rho[X,Y] - H_\rho[Y]$, provided all of the entropies are finite. The chain rule remains valid, conditioning still reduces entropy, and the joint entropy is still $\leq$ the sum of the marginal entropies, with equality iff the variables are independent. However, depending on the reference measure, $H_\rho[X]$ can be negative; e.g., if ρ is Lebesgue measure and $\mathcal{L}(X) = \delta(x)$, then $H_\rho[X] = -\infty$.

28.2 Relative Entropy or Kullback-Leibler Divergence

Some of the difficulties associated with Shannon entropy, in the general case, can be evaded by using relative entropy.

Definition 359 (Relative Entropy, Kullback-Leibler Divergence) *Given two probability distributions, $\nu << \mu$, the* relative entropy of ν with respect to μ, *or* the Kullback-Leibler divergence of ν from μ, *is*

$$D(\mu\|\nu) = -\mathbf{E}_\mu\left[\log \frac{d\nu}{d\mu}\right] \tag{28.7}$$

If ν is not absolutely continuous with respect to μ, then $D(\mu\|\nu) = \infty$.

Lemma 360 *$D(\mu\|\nu) \geq 0$, with equality iff $\nu = \mu$ almost everywhere (μ).*

Proof: From Jensen's inequality, $\mathbf{E}_\mu\left[\log \frac{d\nu}{d\mu}\right] \leq \log \mathbf{E}_\mu\left[\frac{d\nu}{d\mu}\right] = \log 1 = 0$. The second part follows from the conditions for equality in Jensen's inequality. □

Lemma 361 (Divergence and Total Variation) *For any two distributions, $D(\mu\|\nu) \geq \frac{1}{2\ln 2}\|\mu - \nu\|_1^2$.*

Proof: Algebra. See, e.g., Cover and Thomas (1991, Lemma 12.6.1, pp. 300–301). □

Definition 362 *The* conditional relative entropy, *$D(\mu(Y|X)\|\nu(Y|X))$ is*

$$D(\mu(Y|X)\|\nu(Y|X)) \equiv -\mathbf{E}_\mu\left[\log \frac{d\nu(Y|X)}{d\mu(Y|X)}\right] \tag{28.8}$$

Lemma 363 (Chain Rule for Relative Entropy) $D(\mu(X,Y)\|\nu(X,Y)) = D(\mu(X)\|\nu(X)) + D(\mu(Y|X)\|\nu(Y|X))$

Proof: Algebra. □

Shannon entropy can be constructed from the relative entropy.

Lemma 364 *The Shannon entropy of a discrete-valued random variable X, with distribution μ, is*

$$H[X] = \log n - D(\mu \| \upsilon) \tag{28.9}$$

where n is the number of values X can take on (with positive probability), and υ is the uniform distribution over those values.

Proof: Algebra. □

A similar result holds for the entropy of a variable which takes values in a finite subset, of volume V, of a Euclidean space, i.e., $H_{\lambda}[X] = \log V - D(\mu \| \upsilon)$, where λ is Lebesgue measure and υ is the uniform probability measure on the range of X.

28.2.1 Statistical Aspects of Relative Entropy

From Lemma 361, "convergence in relative entropy", $D(\mu \| \nu_n) \rightarrow 0$ as $n \rightarrow \infty$, implies convergence in the total variation (L_1) metric. Because of Lemma 360, we can say that KL divergence has some of the properties of a metric on the space of probability distribution: it's non-negative, with equality only when the two distributions are equal (a.e.). Unfortunately, however, it is not symmetric, and it does not obey the triangle inequality. (This is why it's the KL *divergence* rather than the KL *distance.*) Nonetheless, it's *enough* like a metric that it can be used to construct a kind of geometry on the space of probability distributions, and so of statistical models, which can be extremely useful. While we will not be able to go very far into this information geometry[3], it will be important to indicate a few of the connections between information-theoretic notions, and the more usual ones of statistical theory.

Definition 365 (Cross-entropy) *The* cross-entropy of ν and μ*, $Q(\mu \| \nu)$, is*

$$Q_{\rho}(\mu \| \nu) \equiv -\mathbf{E}_{\mu}\left[\log \frac{d\nu}{d\rho}\right] \tag{28.10}$$

where ν is absolutely continuous with respect to the reference measure ρ. If the domain is discrete, we will take the reference measure to be uniform and drop the subscript, unless otherwise noted.

Lemma 366 *Suppose ν and μ are the distributions of two probability models, and $\nu << \mu$. Then the cross-entropy is the expected negative log-likelihood of the model corresponding to ν, when the actual distribution is μ. The actual or empirical negative log-likelihood of the model corresponding to ν is $Q_{\rho}(\nu \| \eta)$, where η is the empirical distribution.*

[3] See Kass and Vos (1997) or Amari and Nagaoka (1993/2000). For applications to statistical inference for stochastic processes, see Taniguchi and Kakizawa (2000). For an easier general introduction, Kulhavý (1996) is hard to beat.

Proof: Obvious from the definitions. □

Lemma 367 *If $\nu << \mu << \rho$, then $Q_\rho(\mu\|\nu) = H_\rho[\mu] + D(\mu\|\nu)$.*

Proof: By the chain rule for densities,

$$\frac{d\nu}{d\rho} = \frac{d\mu}{d\rho}\frac{d\nu}{d\mu} \tag{28.11}$$

$$\log\frac{d\nu}{d\rho} = \log\frac{d\mu}{d\rho} + \log\frac{d\nu}{d\mu} \tag{28.12}$$

$$\mathbf{E}_\mu\left[\log\frac{d\nu}{d\rho}\right] = \mathbf{E}_\mu\left[\log\frac{d\mu}{d\rho}\right] + \mathbf{E}_\mu\left[\log\frac{d\nu}{d\mu}\right] \tag{28.13}$$

The result follows by applying the definitions. □

Corollary 368 (Gibbs's Inequality) *$Q_\rho(\mu\|\nu) \geq H_\rho[\mu]$, with equality iff $\nu = \mu$ a.e.*

Proof: Insert the result of Lemma 360 into the preceding proposition. □

The statistical interpretation of the proposition is this: The log-likelihood of a model, leading to distribution ν, can be broken into two parts. One is the divergence of ν from μ; the other just the entropy of μ, i.e., it is the same for all models. If we are considering the expected log-likelihood, then μ is the actual data-generating distribution. If we are considering the empirical log-likelihood, then μ is the empirical distribution. In either case, to maximize the likelihood is to minimize the relative entropy, or divergence. What we would like to do, as statisticians, is minimize the divergence from the data-generating distribution, since that will let us predict future values. What we *can* do is minimize divergence from the empirical distribution. The consistency of maximum likelihood methods comes down, then, to finding conditions under which a shrinking divergence from the empirical distribution guarantees a shrinking divergence from the true distribution.[4]

Definition 369 *Let $\theta \in \mathbb{R}^k$, $k < \infty$, be the parameter indexing a set $\mathcal{M}$ of statistical models, where for every θ, $\nu_\theta << \rho$, with densities p_θ. Then the* Fisher information matrix *is*

$$I_{ij}(\theta) \equiv \mathbf{E}_{\nu_\theta}\left[\left(\frac{\partial \log p_\theta}{d\theta_i}\right)\left(\frac{\partial \log p_\theta}{d\theta_j}\right)\right] \tag{28.14}$$

Corollary 370 *The Fisher information matrix is equal to the Hessian (second partial derivative) matrix of the relative entropy:*

$$I_{ij}(\theta_0) = \frac{\partial^2}{\partial\theta_i\partial\theta_j} D(\nu_{\theta_0}\|\nu_\theta) \tag{28.15}$$

[4] If we did have a triangle inequality, then we could say $D(\mu\|\nu) \leq D(\mu\|\eta) + D(\eta\|\nu)$, and it would be enough to make sure that both the terms on the RHS went to zero, say by some combination of maximizing the likelihood in-sample, so $D(\eta\|\nu)$ is small, and ergodicity, so that $D(\mu\|\eta)$ is small. While, as noted, there is no triangle inequality, under some conditions this idea is roughly right; there are nice diagrams in Kulhavý (1996).

PROOF: It is a classical result (see, e.g., Lehmann and Casella (1998, sec. 2.6.1)) that $I_{ij}(\theta) = -\mathbf{E}_{\nu_\theta}\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log p_\theta\right]$. The present result follows from this, Lemma 366, Lemma 367, and the fact that $H_\rho[\nu_{\theta_0}]$ is independent of θ. □

28.3 Mutual Information

Definition 371 (Mutual Information) *The mutual information between two random variables, X and Y, is the divergence of the product of their marginal distributions from their actual joint distribution:*

$$I[X;Y] \equiv D(\mathcal{L}(X,Y)\,\|\mathcal{L}(X)\times\mathcal{L}(Y)) \tag{28.16}$$

Similarly, the mutual information among n random variables $X_1, X_2, \ldots X_n$ is

$$I[X_1;X_2;\ldots;X_n] \equiv D(\mathcal{L}(X_1,X_2,\ldots X_n)\,\|\prod_{i=1}^{n}\mathcal{L}(X_i)) \tag{28.17}$$

the divergence of the product distribution from the joint distribution.

Proposition 372 *$I[X;Y] \geq 0$, with equality iff X and Y are independent.*

PROOF: Directly from Lemma 360. □

Proposition 373 *If all the entropies involved are finite,*

$$\begin{aligned} I[X;Y] &= H[X] + H[Y] - H[X,Y] && (28.18)\\ &= H[X] - H[X|Y] && (28.19)\\ &= H[Y] - H[Y|X] && (28.20) \end{aligned}$$

so $I[X;Y] \leq H[X] \wedge H[Y]$.

PROOF: Calculation. □

This leads to the interpretation of the mutual information as the reduction in uncertainty or effective variability of X when Y is known, averaging over their joint distribution. Notice that in the discrete case, we can say $H[X] = I[X;X]$, which is why $H[X]$ is sometimes known as the *self-information.*

28.3.1 Mutual Information Function

Just as with the autocovariance function, we can define a mutual information function for one-parameter processes, to serve as a measure of serial dependence.

Definition 374 (Mutual Information Function) *The* mutual information function *of a one-parameter stochastic process X is*

$$\iota(t_1,t_2) \equiv I[X_{t_1};X_{t_2}] \tag{28.21}$$

which is symmetric in its arguments. If the process is stationary, it is a function of $|t_1 - t_2|$ alone.

Notice that, unlike the autocovariance function, ι includes *nonlinear* dependencies between X_{t_1} and X_{t_2}. Also notice that $\iota(\tau) = 0$ means that the two variables are strictly independent, not just uncorrelated.

Theorem 375 *A stationary process is mixing if* $\iota(\tau) \rightarrow 0$*.*

PROOF: Because then 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, $\mathcal{L}\left(X_{t_1}\right) \mathcal{L}\left(X_{t_2}\right)$, is being forced down towards zero, which implies mixing (Definition 338). □

Chapter 29

Entropy Rates and Asymptotic Equipartition

Section 29.1 introduces the entropy rate — the asymptotic entropy per time-step of a stochastic process — and shows that it is well-defined; and similarly for information, divergence, etc. rates.

Section 29.2 proves the Shannon-MacMillan-Breiman theorem, a.k.a. the asymptotic equipartition property, a.k.a. the entropy ergodic theorem: asymptotically, almost all sample paths of a stationary ergodic process have the same log-probability per time-step, namely the entropy rate. This leads to the idea of "typical" sequences, in Section 29.2.1.

Section 29.3 discusses some aspects of asymptotic likelihood, using the asymptotic equipartition property, and allied results for the divergence rate.

29.1 Information-Theoretic Rates

Definition 376 (Entropy Rate) *The* entropy rate *of a random sequence X is*

$$h(X) \equiv \lim_{n} H_{\rho}[X_1^n] n \tag{29.1}$$

when the limit exists.

Definition 377 (Limiting Conditional Entropy) *The* limiting conditional entropy *of a random sequence X is*

$$h'(X) \equiv \lim_{n} H_{\rho}[X_n | X_1^{n-1}] \tag{29.2}$$

when the limit exists.

Lemma 378 *For a stationary sequence, $H_\rho[X_n|X_1^{n-1}]$ is non-increasing in n. Moreover, its limit exists if X takes values in a discrete space.*

PROOF: Because "conditioning reduces entropy", $H_\rho[X_{n+1}|X_1^n] \leq H[X_{n+1}|X_2^n]$. By stationarity, $H_\rho[X_{n+1}|X_2^n] = H_\rho[X_n|X_1^{n-1}]$. If X takes discrete values, then conditional entropy is non-negative, and a non-increasing sequence of non-negative real numbers always has a limit. □

Remark: Discrete values are a *sufficient* condition for the existence of the limit, not a necessary one.

We now need a natural-looking, but slightly technical, result from real analysis.

Theorem 379 (Cesàro) *For any sequence of real numbers $a_n \rightarrow a$, the sequence $b_n = n^{-1}\sum_{i=1}^{n} a_n$ also converges to a.*

PROOF: For every $\epsilon > 0$, there is an $N(\epsilon)$ such that $|a_n - a| < \epsilon$ whenever $n > N(\epsilon)$. Now take b_n and break it up into two parts, one summing the terms below $N(\epsilon)$, and the other the terms above.

$$\begin{aligned}
\lim_n |b_n - a| &= \lim_n \left|n^{-1}\sum_{i=1}^{n} a_i - a\right| && (29.3)\\
&\leq \lim_n n^{-1}\sum_{i=1}^{n}|a_i - a| && (29.4)\\
&\leq \lim_n n^{-1}\left(\sum_{i=1}^{N(\epsilon)}|a_i - a| + (n - N(\epsilon))\epsilon\right) && (29.5)\\
&\leq \lim_n n^{-1}\left(\sum_{i=1}^{N(\epsilon)}|a_i - a| + n\epsilon\right) && (29.6)\\
&= \epsilon + \lim_n n^{-1}\sum_{i=1}^{N(\epsilon)}|a_i - a| && (29.7)\\
&= \epsilon && (29.8)
\end{aligned}$$

Since ϵ was arbitrary, $\lim b_n = a$. □

Theorem 380 (Entropy Rate) *For a stationary sequence, if the limiting conditional entropy exists, then it is equal to the entropy rate, $h(X) = h'(X)$.*

PROOF: Start with the chain rule to break the joint entropy into a sum of conditional entropies, use Lemma 378 to identify their limit as $h^{]prime}(X)$, and then use Cesàro's theorem:

$$\begin{aligned} h(X) &= \lim_n \frac{1}{n} H_\rho[X_1^n] && (29.9)\\ &= \lim_n \frac{1}{n} \sum_{i=1}^{n} H_\rho[X_i|X_1^{i-1}] && (29.10)\\ &= h'(X) && (29.11)\end{aligned}$$

as required. □

Because $h(X) = h'(X)$ for stationary processes (when both limits exist), it is not uncommon to find what I've called the limiting conditional entropy referred to as the entropy rate.

Lemma 381 *For a stationary sequence $h(X) \leq H[X_1]$, with equality iff the sequence is IID.*

Proof: Conditioning reduces entropy, unless the variables are independent, so $H[X_n|X_1^{n-1}] < H[X_n]$, unless $X_n \perp\!\!\!\perp X_1^{n-1}$. For this to be true of all n, which is what's needed for $h(X) = H[X_1]$, all the values of the sequence must be independent of each other; since the sequence is stationary, this would imply that it's IID. □

Example 382 (Markov Sequences) *If X is a stationary Markov sequence, then $h(X) = H_\rho[X_2|X_1]$, because, by the chain rule, $H_\rho[X_1^n] = H_\rho[X_1] + \sum_{t=2}^{n} H_\rho[X_t|X_1^{t-1}]$. By the Markov property, however, $H_\rho[X_t|X_1^{t-1}] = H_\rho[X_t|X_{t-1}]$, which by stationarity is $H_\rho[X_2|X_1]$. Thus, $H_\rho[X_1^n] = H_\rho[X_1]+(n-1)H_\rho[X_2|X_1]$. Dividing by n and taking the limit, we get $H_\rho[X_1^n] = H_\rho[X_2|X_1]$.*

Example 383 (Higher-Order Markov Sequences) *If X is a k^{th} order Markov sequence, then the same reasoning as before shows that $h(X) = H_\rho[X_{k+1}|X_1^k]$ when X is stationary.*

Definition 384 (Divergence Rate) *The* divergence rate *or* relative entropy rate *of the infinite-dimensional distribution Q from the infinite-dimensional distribution P, $d(P\|Q)$, is*

$$d(P\|Q) = \lim_n \mathbf{E}_P \left[\log \left(\left. \frac{dP}{dQ} \right|_{\sigma(X_{-n}^0)} \right) \right] \tag{29.12}$$

if all the finite-dimensional distributions of Q dominate all the finite-dimensional distributions of P. If P and Q have densities, respectively p and q, with respect to a common reference measure, then

$$d(P\|Q) = \lim_n \mathbf{E}_P \left[\log \frac{p(X_0|X_{-n}^{-1})}{q(X_0|X_{-n}^{-1})} \right] \tag{29.13}$$

29.2 The Shannon-McMillan-Breiman Theorem or Asymptotic Equipartition Property

This is a central result in information theory, acting as a kind of ergodic theorem for the entropy. That is, we want to say that, for almost all ω,

$$-\frac{1}{n}\log \mathbb{P}\left(X_1^n(\omega)\right) \rightarrow \lim_n \frac{1}{n}\mathbf{E}\left[-\log \mathbb{P}\left(X_1^n\right)\right] = h(X)$$

At first, it looks like we should be able to make a nice time-averaging argument. We can always factor the joint probability,

$$\frac{1}{n}\log \mathbb{P}\left(X_1^n\right) = \frac{1}{n}\sum_{t=1}^{n}\log \mathbb{P}\left(X_t | X_1^{t-1}\right)$$

with the understanding that $\mathbb{P}\left(X_1|X_1^0\right) = \mathbb{P}\left(X_1\right)$. This looks rather like the sort of Cesàro average that we became familiar with in ergodic theory. The problem is, there we were averaging $f(T^t\omega)$ for a *fixed* function f. This is not the case here, because we are conditioning on long and longer stretches of the past. There's no problem if the sequence is Markovian, because then the remote past is irrelevant, by the Markov property, and we can just condition on a fixed-length stretch of the past, so we're averaging a fixed function shifted in time. (This is why Shannon's original argument was for Markov chains.) The result nonetheless more broadly, but requires more subtlety than might otherwise be thought. Breiman's original proof of the general case was fairly involved[1], requiring both martingale theory, and a sort of dominated convergence theorem for ergodic time averages. (You can find a simplified version of his argument in Kallenberg, at the end of chapter 11.) We will go over the "sandwiching" argument of Algoet and Cover (1988), which is, to my mind, more transparent.

The idea of the sandwich argument is to show that, for large n, $-n^{-1}\log \mathbb{P}\left(X_1^n\right)$ must lie between an upper bound, h_k, obtained by approximating the sequence by a Markov process of order k, and a lower bound, which will be shown to be h. Once we establish that $h_k \downarrow h$, we will be done.

Definition 385 (Markov Approximation) *For each k, define the order k Markov approximation to X by*

$$\mu_k(X_1^n) = \mathbb{P}\left(X_1^k\right)\prod_{t=k+1}^{n}\mathbb{P}\left(X_t|X_{t-k}^{t-1}\right) \tag{29.14}$$

μ_k is the distribution of a stationary Markov process of order k, where the distribution of X_1^{k+1} matches that of the original process.

[1] Notoriously, the proof in his original paper was actually invalid, forcing him to publish a correction.

Lemma 386 *For each k, the entropy rate of the order k Markov approximation is is equal to $H[X_{k+1}|X_1^k]$.*

Proof: Under the approximation (but not under the original distribution of X), $H[X_1^n] = H[X_1^k]+(n-k)H[X_{k+1}|X_1^k]$, by the Markov property and stationarity (as in Examples 382 and 383). Dividing by n and taking the limit as $n \rightarrow \infty$ gives the result. □

Lemma 387 *If X is a stationary two-sided sequence, then $Y_t = f(X_{-\infty}^t)$ defines a stationary sequence, for any measurable f. If X is also ergodic, then Y is ergodic too.*

Proof: Because X is stationary, it can be represented as a measure-preserving shift on sequence space. Because it is measure-preserving, $\theta X_{-\infty}^t \overset{d}{=} X_{-\infty}^t$, so $Y(t) \overset{d}{=} Y(t+1)$, and similarly for all finite-length blocks of Y. Thus, all of the finite-dimensional distributions of Y are shift-invariant, and these determine the infinite-dimensional distribution, so Y itself must be stationary.

To see that Y must be ergodic if X is ergodic, recall that a random sequence is ergodic iff its corresponding shift dynamical system is ergodic. A dynamical system is ergodic iff all invariant functions are a.e. constant (Theorem 304). Because the Y sequence is obtained by applying a measurable function to the X sequence, a shift-invariant function of the Y sequence is a shift-invariant function of the X sequence. Since the latter are all constant a.e., the former are too, and Y is ergodic. □

Lemma 388 *If X is stationary and ergodic, then, for every k,*

$$\mathbb{P}\left(\lim_n -\frac{1}{n}\log \mu_k(X_1^n(\omega)) = h_k\right) = 1 \tag{29.15}$$

i.e., $-\frac{1}{n}\log \mu_k(X_1^n(\omega))$ converges a.s. to h_k.

Proof: Start by factoring the approximating Markov measure in the way suggested by its definition:

$$-\frac{1}{n}\log \mu_k(X_1^n) = -\frac{1}{n}\log \mathbb{P}\left(X_1^k\right) - \frac{1}{n}\sum_{t=k+1}^{n} \log \mathbb{P}\left(X_t|X_{t-k}^{t-1}\right) \tag{29.16}$$

As n grows, $\frac{1}{n}\log \mathbb{P}\left(X_1^k\right) \rightarrow 0$, for every fixed k. On the other hand, $-\log \mathbb{P}\left(X_t|X_{t-k}^{t-1}\right)$ is a measurable function of the past of the process, and since X is stationary and ergodic, it, too, is stationary and ergodic (Lemma 387). So

$$\begin{aligned}
-\frac{1}{n}\log \mu_k(X_1^n) &\rightarrow -\frac{1}{n}\sum_{t=k+1}^{n} \log \mathbb{P}\left(X_t|X_{t-k}^{t-1}\right) && (29.17)\\
&\overset{a.s.}{\rightarrow} \mathbf{E}\left[-\log \mathbb{P}\left(X_{k+1}|X_1^k\right)\right] && (29.18)\\
&= h_k && (29.19)
\end{aligned}$$

by Theorem 312. □

Definition 389 *The* infinite-order approximation *to the entropy rate of a discrete-valued stationary process X is*

$$h_\infty(X) \equiv \mathbf{E}\left[-\log \mathbb{P}\left(X_0|X_{-\infty}^{-1}\right)\right] \tag{29.20}$$

Lemma 390 *If X is stationary and ergodic, then*

$$\lim_n -\frac{1}{n}\log \mathbb{P}\left(X_1^n|X_{-\infty}^0\right) = h_\infty \tag{29.21}$$

almost surely.

Proof: Via Theorem 312 again, as in Lemma 388. □

Lemma 391 *For a stationary, ergodic, finite-valued random sequence, $h_k(X) \downarrow h_\infty(X)$.*

Proof: By the martingale convergence theorem, for every $x_0 \in \Xi$,

$$\mathbb{P}\left(X_0 = x_0|X_n^{-1}\right) \overset{a.s.}{\rightarrow} \mathbb{P}\left(X_0 = x_0|X_\infty^{-1}\right) \tag{29.22}$$

Since Ξ is finite, the probability of any point in Ξ is between 0 and 1 inclusive, and $p \log p$ is bounded and continuous. So we can apply bounded convergence to get that

$$\begin{aligned} h_k &= \mathbf{E}\left[-\sum_{x_0} \mathbb{P}\left(X_0 = x_0|X_{-k}^{-1}\right)\log \mathbb{P}\left(X_0 = x_0|X_{-k}^{-1}\right)\right] && (29.23)\\ &\rightarrow \mathbf{E}\left[-\sum_{x_0} \mathbb{P}\left(X_0 = x_0|X_{-\infty}^{-1}\right)\log \mathbb{P}\left(X_0 = x_0|X_{-\infty}^{-1}\right)\right] && (29.24)\\ &= h_\infty && (29.25) \end{aligned}$$

Lemma 392 *$h_\infty(X)$ is the entropy rate of X, i.e. $h_\infty(X) = h(X)$.*

Proof: Clear from Theorem 380 and the definition of conditional entropy. □

We are almost ready for the proof, but need one technical lemma first.

Lemma 393 *If $R_n \geq 0$, $\mathbf{E}[R_n] \leq 1$ for all n, then*

$$\limsup_n \frac{1}{n}\log R_n \leq 0 \tag{29.26}$$

almost surely.

Proof: Pick any $\epsilon > 0$.

$$\begin{aligned} \mathbb{P}\left(\frac{1}{n}\log R_n \geq \epsilon\right) &= \mathbb{P}\left(R_n \geq e^{n\epsilon}\right) && (29.27)\\ &\leq \frac{\mathbf{E}[R_n]}{e^{n\epsilon}} && (29.28)\\ &\leq e^{-n\epsilon} && (29.29) \end{aligned}$$

by Markov's inequality. Since $\sum_n e^{-n\epsilon} \leq \infty$, by the Borel-Cantelli lemma, $\limsup_n n^{-1}\log R_n \leq \epsilon$. Since ϵ was arbitrary, this concludes the proof. □

Theorem 394 (Asymptotic Equipartition Property) *For a stationary, ergodic, finite-valued random sequence X,*

$$-\frac{1}{n}\log \mathbb{P}\left(X_1^n\right) \rightarrow h(X) \ a.s. \tag{29.30}$$

Proof: For every k, $\mu_k(X_1^n)/\mathbb{P}\left(X_1^n\right) \geq 0$, and $\mathbf{E}\left[\mu_k(X_1^n)/\mathbb{P}\left(X_1^n\right)\right] \leq 1$. Hence, by Lemma 393,

$$\limsup_n \frac{1}{n}\log\frac{\mu_k(X_1^n)}{\mathbb{P}\left(X_1^n\right)} \leq 0 \tag{29.31}$$

a.s. Manipulating the logarithm,

$$\limsup_n \frac{1}{n}\log \mu_k(X_1^n) \leq -\limsup_n -\frac{1}{n}\log\mathbb{P}\left(X_1^n\right) \tag{29.32}$$

From Lemma 388, $\limsup_n \frac{1}{n}\log\mu_k(X_1^n) = \lim_n \frac{1}{n}\log\mu_k(X_1^n) = -h_k(X)$, a.s. Hence, for each k,

$$h_k(X) \geq \limsup_n -\frac{1}{n}\log\mathbb{P}\left(X_1^n\right) \tag{29.33}$$

almost surely.

A similar manipulation of $\mathbb{P}\left(X_1^n\right)/\mathbb{P}\left(X_1^n|X_{-\infty}^0\right)$ gives

$$h_\infty(X) \leq \liminf_n -\frac{1}{n}\log\mathbb{P}\left(X_1^n\right) \tag{29.34}$$

a.s.

As $h_k \downarrow h_\infty$, it follows that the liminf and the limsup of the normalized log likelihood must be equal almost surely, and so equal to h_∞, which is to say to $h(X)$. □

Why is this called the AEP? Because, to within an $o(n)$ term, all sequences of length n have the same log-likelihood (to within factors of $o(n)$, if they have positive probability at all. In this sense, the likelihood is "equally partitioned" over those sequences.

29.2.1 Typical Sequences

Let's turn the result of the AEP around. For large n, the probability of a given sequence is either approximately 2^{-nh} or approximately zero[2]. To get the total probability to sum up to one, there need to be about 2^{nh} sequences with positive probability. If the size of the alphabet is s, then the fraction of sequences which are actually exhibited is $2^{n(h-\log s)}$, an increasingly small fraction (as $h \leq \log s$). Roughly speaking, these are the *typical* sequences, any one of which, via ergodicity, can act as a representative of the complete process.

[2] Of course that assumes using base-2 logarithms in the definition of entropy.

29.3 Asymptotic Likelihood

29.3.1 Asymptotic Equipartition for Divergence

Using methods analogous to those we employed on the AEP for entropy, it is possible to prove the following.

Theorem 395 *Let P be an asymptotically mean-stationary distribution, with stationary mean $\overline{P}$, with ergodic component function ϕ. Let M be a homogeneous finite-order Markov process, whose finite-dimensional distributions dominate those of P and $\overline{P}$; denote the densities with respect to M by p and $\overline{p}$, respectively. If $\lim_n n^{-1} \log \overline{p}(X_1^n)$ is an invariant function $\overline{P}$-a.e., then*

$$-\frac{1}{n} \log p(X_1^n(\omega)) \overset{a.s.}{\rightarrow} d(\overline{P}_{\phi(\omega)} \| M) \tag{29.35}$$

where $\overline{P}_{\phi(\omega)}$ is the stationary, ergodic distribution of the ergodic component.

Proof: See Algoet and Cover (1988, theorem 4), Gray (1990, corollary 8.4.1).

Remark. The usual AEP is in fact a consequence of this result, with the appropriate reference measure. (Which?)

29.3.2 Likelihood Results

It is left as an exercise for you to obtain the following result, from the AEP for relative entropy, Lemma 367 and the chain rules.

Theorem 396 *Let P be a stationary and ergodic data-generating process, whose entropy rate, with respect to some reference measure ρ, is h. Further let M be a finite-order Markov process which dominates P, whose density, with respect to the reference measure, is m. Then*

$$-\frac{1}{n} \log m(X_1^n) \rightarrow h + d(P \| M) \tag{29.36}$$

P-almost surely.

29.4 Exercises

Exercise 29.1 *Markov approximations are maximum-entropy approximations. (You may assume that the process X takes values in a finite set.)*

- *a Prove that μ_k, as defined in Definition 385, gets the distribution of sequences of length $k+1$ correct, i.e., for any set $A \in \mathcal{X}^{k+1}$, $\nu(A) = \mathbb{P}\left(X_1^{k+1} \in A\right)$.*
- *b Prove that $\mu_{k'}$, for any any $k' > k$, also gets the distribution of length $k+1$ sequences right.*

c *In a slight abuse of notation, let $H[\nu(X_1^n)]$ stand for the entropy of a sequence of length n when distributed according to ν. Show that $H[\mu_k(X_1^n)] \geq H[\mu_{k'}(X_1^n)]$ if $k' > k$. (Note that the $n \leq k$ case is easy!)*

d *Is it true that that if ν is any other measure which gets the distribution of sequences of length $k+1$ right, then $H[\mu_k(X_1^n)] \geq H[\nu(X_1^n)]$? If yes, prove it; if not, find a counter-example.*

Exercise 29.2 *Prove Theorem 396.*

Chapter 30

General Theory of Large Deviations

> A family of random variables follows the *large deviations principle* if the probability of the variables falling into "bad" sets, representing large deviations from expectations, declines exponentially in some appropriate limit. Section 30.1 makes this precise, using some associated technical machinery, and explores a few consequences. The central one is Varadhan's Lemma, for the asymptotic evaluation of exponential integrals in infinite-dimensional spaces.
>
> Having found one family of random variables which satisfy the large deviations principle, many other, related families do too. Section 30.2 lays out some ways in which this can happen.

As the great forensic statistician C. Chan once remarked, "Improbable events permit themselves the luxury of occurring" (reported in Biggers, 1928). Large deviations theory, as I have said, studies these little luxuries.

30.1 Large Deviation Principles: Main Definitions and Generalities

Some technicalities:

Definition 397 (Level Sets) *For any real-valued function $f : \Xi \mapsto \mathbb{R}$, the* level sets *are the inverse images of intervals from $-\infty$ to c inclusive, i.e., all sets of the form $\{x \in \Xi : f(x) \leq c\}$.*

Definition 398 (Lower Semi-Continuity) *A real-valued function $f : \Xi \mapsto \mathbb{R}$ is* lower semi-continuous *if $x_n \rightarrow x$ implies $\liminf f(x_n) \geq f(x)$.*

Lemma 399 *A function is lower semi-continuous iff either of the following equivalent properties hold.*

i For all $x \in \Xi$, the infimum of f over increasingly small open balls centered at x approaches $f(x)$:

$$\lim_{\delta \rightarrow 0} \inf_{y:\ d(y,x) < \delta} f(y) = f(x) \tag{30.1}$$

ii f has closed level sets.

Proof: A character-building exercise in real analysis, left to the reader. □

Lemma 400 *A lower semi-continuous function attains its minimum on every non-empty compact set, i.e., if C is compact and $\neq \emptyset$, there is an $x \in C$ such that $f(x) = \inf_{y \in C} f(y)$.*

Proof: Another character-building exercise in real analysis. □

Definition 401 (Logarithmic Equivalence) *Two sequences of positive real numbers a_n and b_n are* logarithmically equivalent*, $a_n \simeq b_n$, when*

$$\lim_{n \rightarrow \infty} \frac{1}{n} \left(\log a_n - \log b_n \right) = 0 \tag{30.2}$$

Similarly, for continuous parameterizations by $\epsilon > 0$, $a_\epsilon \simeq b_\epsilon$ when

$$\lim_{\epsilon \rightarrow 0} \epsilon \left(\log a_\epsilon - \log b_\epsilon \right) = 0 \tag{30.3}$$

Lemma 402 ("Fastest rate wins") *For any two sequences of positive numbers, $(a_n + b_n) \simeq a_n \vee b_n$.*

Proof: A character-building exercise in elementary analysis. □

Definition 403 (Large Deviation Principle) *A parameterized family of random variables, X_ϵ, $\epsilon > 0$, taking values in a metric space Ξ with Borel σ-field $\mathcal{X}$,* obeys a large deviation principle with rate $1/\epsilon$, *or just* obeys an LDP, *when, for any set $B \in \mathcal{X}$,*

$$- \inf_{x \in \mathrm{int} B} J(x) \leq \liminf_{\epsilon \rightarrow 0} \epsilon \log \mathbb{P}\left(X_\epsilon \in B \right) \leq \limsup_{\epsilon \rightarrow 0} \epsilon \log \mathbb{P}\left(X_\epsilon \in B \right) \leq - \inf_{x \in \mathrm{cl} B} J(x) \tag{30.4}$$

for some non-negative function $J : \Xi \mapsto [0, \infty]$, its raw rate function*. If J is lower semi-continuous, it is just a* rate function*. If J is lower semi-continuous and has compact level sets, it is a* good rate function.[1] *By a slight abuse of notation, we will write $J(B) = \inf_{x \in B} J(x)$.*

[1] Sometimes what Kallenberg and I are calling a "good rate function" is just "a rate function", and our "rate function" gets demoted to "weak rate function".

Remark: The most common choices of ϵ are $1/n$, in sample-size or discrete sequence problems, or ε^2, in small-noise problems (as in Chapter 22).

Lemma 404 (Uniqueness of Rate Functions) *If X_ϵ obeys the LDP with raw rate function J, then it obeys the LDP with a unique rate function J'.*

PROOF: First, show that a raw rate function can always be replaced by a lower semi-continuous function, i.e. a non-raw (cooked?) rate function. Then, show that non-raw rate functions are unique.

For any raw rate function J, define $J'(x) = \liminf_{y \rightarrow x} J(x)$. This is clearly lower semi-continuous, and $J'(x) \leq J(x)$. However, for any open set B, $\inf_{x \in B} J'(x) = \inf_{x \in B} J(x)$, so J and J' are equivalent for purposes of the LDP.

Now assume that J is a lower semi-continuous rate function, and suppose that $K \neq J$ was too; without loss of generality, assume that $J(x) > K(x)$ at some point x. We can use semi-continuity to find an open neighborhood B of x such that $J(\mathrm{cl}B) > K(x)$. But, substituting into Eq. 30.4, we obtain a contradiction:

$$\begin{aligned} -K(x) &\leq -K(B) && (30.5)\\ &\leq \liminf_{\epsilon \rightarrow 0} \epsilon \log \mathbb{P}(X_\epsilon \in B) && (30.6)\\ &\leq -J(\mathrm{cl}B) && (30.7)\\ &\leq -K(x) && (30.8) \end{aligned}$$

Hence there can be no such rate function K, and J is the unique rate function. □

Lemma 405 *If X_ϵ obeys an LDP with rate function J, then $J(x) = 0$ for some x.*

PROOF: Because $\mathbb{P}(X_\epsilon \in \Xi) = 1$, we must have $J(\Xi) = 0$, and rate functions attain their infima. □

Definition 406 *A Borel set B is J-continuous, for some rate function J, when $J(\mathrm{int}B) = J(\mathrm{cl}B)$.*

Lemma 407 *If X_ϵ satisfies the LDP with rate function J, then for every J-continuous set B,*

$$\lim_{\epsilon \rightarrow 0} \epsilon \log \mathbb{P}(X_\epsilon \in B) = -J(B) \qquad (30.9)$$

PROOF: By J-continuity, the right and left hand extremes of Eq. 30.4 are equal, so the limsup and the liminf sandwiched between them are equal; consequently the limit exists. □

Remark: The obvious implication is that, for small ϵ, $\mathbb{P}(X_\epsilon \in B) \approx c e^{-J(B)/\epsilon}$, which explains why we say that the LDP has rate $1/\epsilon$. (Actually, c need not be constant, but it must be at least $o(\epsilon)$, i.e., it must go to zero faster than ϵ itself does.)

There are several equivalent ways of defining the large deviation principle. The following is especially important, because it is often simplifies proofs.

Lemma 408 *X_ϵ obeys the LDP with rate $1/\epsilon$ and rate function $J(x)$ if and only if*

$$\begin{aligned} \limsup_{\epsilon\rightarrow 0} \epsilon \log \mathbb{P}\left(X_\epsilon \in C\right) & \leq & -J(C) \qquad (30.10)\\ \liminf_{\epsilon\rightarrow 0} \epsilon \log \mathbb{P}\left(X_\epsilon \in O\right) & \geq & -J(O) \qquad (30.11) \end{aligned}$$

for every closed Borel set C and every open Borel set $O \subset \Xi$.

Proof: "If": The closure of any set is closed, and the interior of any set is open, so Eqs. 30.10 and 30.11 imply

$$\begin{aligned} \limsup_{\epsilon\rightarrow 0} \epsilon \log \mathbb{P}\left(X_\epsilon \in \mathrm{cl}B\right) & \leq & -J(\mathrm{cl}B) \qquad (30.12)\\ \liminf_{\epsilon\rightarrow 0} \epsilon \log \mathbb{P}\left(X_\epsilon \in \mathrm{int}B\right) & \geq & -J(\mathrm{int}B) \qquad (30.13) \end{aligned}$$

but $\mathbb{P}\left(X_\epsilon \in B\right) \leq \mathbb{P}\left(X_\epsilon \in \mathrm{cl}B\right)$ and $\mathbb{P}\left(X_\epsilon \in B\right) \geq \mathbb{P}\left(X_\epsilon \in \mathrm{int}B\right)$, so the LDP holds. "Only if": every closed set is equal to its own closure, and every open set is equal to its own interior, so the upper bound in Eq. 30.4 implies Eq. 30.10, and the lower bound Eq. 30.11. □

A deeply important consequence of the LDP is the following, which can be thought of as a version of Laplace's method for infinite-dimensional spaces.

Theorem 409 (Varadhan's Lemma) *If X_ϵ are random variables in a metric space Ξ, obeying an LDP with rate $1/\epsilon$ and rate function J, and $f : \Xi \mapsto \mathbb{R}$ is continuous and bounded from above, then*

$$\Lambda_f \equiv \lim_{\epsilon\rightarrow 0} \epsilon \log \mathbf{E}\left[e^{f(X_\epsilon)/\epsilon}\right] = \sup_{x\in\Xi} f(x) - J(x) \qquad (30.14)$$

Proof: We'll find the limsup and the liminf, and show that they are both $\sup f(x) - J(x)$.

First the limsup. Pick an arbitrary positive integer n. Because f is continuous and bounded above, there exist finitely closed sets, call them $B_1, \ldots B_m$, such that $f \leq -n$ on the complement of $\bigcup_i B_i$, and within each B_i, f varies by at most $1/n$. Now

$$\begin{aligned} & \limsup \epsilon \log \mathbf{E}\left[e^{f(X_\epsilon)/\epsilon}\right] \qquad (30.15)\\ \leq & \quad (-n) \vee \max_{i\leq m} \limsup \epsilon \log \mathbf{E}\left[e^{f(X_\epsilon)/\epsilon}\mathbf{1}_{B_i}(X_\epsilon)\right]\\ \leq & \quad (-n) \vee \max_{i\leq m} \sup_{x\in B_i} f(x) - \inf_{x\in B_i} J(x) \qquad (30.16)\\ \leq & \quad (-n) \vee \max_{i\leq m} \sup_{x\in B_i} f(x) - J(x) + 1/n \qquad (30.17)\\ \leq & \quad (-n) \vee \sup_{x\in\Xi} f(x) - J(x) + 1/n \qquad (30.18) \end{aligned}$$

Letting $n \to \infty$, we get $\limsup \epsilon \log \mathbf{E}\left[e^{f(X_\epsilon)/\epsilon}\right] = \sup f(x) - J(x)$.

To get the liminf, pick any $x \in Xi$ and an arbitrary ball of radius δ around it, $B_{\delta,x}$. We have

$$\begin{aligned}
\liminf \epsilon \log \mathbf{E}\left[e^{f(X_\epsilon)/\epsilon}\right] &\geq \liminf \epsilon \log \mathbf{E}\left[e^{f(X_\epsilon)/\epsilon}\mathbf{1}_{B_{\delta,x}}(X_\epsilon)\right] && (30.19)\\
&\geq \inf_{y \in B_{\delta,x}} f(y) - \inf_{y \in B_{\delta,x}} J(y) && (30.20)\\
&\geq \inf_{y \in B_{\delta,x}} f(y) - J(x) && (30.21)
\end{aligned}$$

Since δ was arbitrary, we can let it go to zero, so (by continuity of f) $\inf_{y \in B_{\delta,x}} f(y) \to f(x)$, or

$$\liminf \epsilon \log \mathbf{E}\left[e^{f(X_\epsilon)/\epsilon}\right] \geq f(x) - J(x) \qquad (30.22)$$

Since this holds for arbitrary x, we can replace the right-hand side by a supremum over all x. Hence $\sup f(x) - J(x)$ is both the liminf and the limsup. □

Remark: The implication of Varadhan's lemma is that, for small ϵ, $\mathbf{E}\left[e^{f(X_\epsilon)/\epsilon}\right] \approx c(\epsilon)e^{\epsilon^{-1}(\sup_{x\in\Xi} f(x)-J(x))}$, where $c(\epsilon) = o(\epsilon)$. So, we can replace the exponential integral with its value at the extremal points, at least to within a multiplicative factor and to first order in the exponent.

An important, if heuristic, consequence of the LDP is that "Highly improbable events tend to happen in the least improbable way". Let us consider two events $B \subset A$, and suppose that $\mathbb{P}(X_\epsilon \in A) > 0$ for all ϵ. Then $\mathbb{P}(X_\epsilon \in B | X_\epsilon \in A) = \mathbb{P}(X_\epsilon \in B)/\mathbb{P}(X_\epsilon \in A)$. Roughly speaking, then, this conditional probability will vanish exponentially, with rate $J(A) - J(B)$. That is, even if we are looking at an exponentially-unlikely large deviation, the vast majority of the probability is concentrated around the *least unlikely* part of the event. More formal statements of this idea are sometimes known as "conditional limit theorems" or "the Gibbs conditioning principle".

30.2 Breeding Large Deviations

Often, the easiest way to prove that one family of random variables obeys a large deviations principle is to prove that another, related family does.

Theorem 410 (Contraction Principle) *If X_ϵ, taking values in a metric space Ξ, obeys an LDP, with rate ϵ and rate function J, and $f : \Xi \mapsto \Upsilon$ is a continuous function from that metric space to another, then $Y_\epsilon = f(X_\epsilon)$ also obeys an LDP, with rate ϵ and raw rate function $K(y) = J(f^{-1}(y))$. If J is a good rate function, then so is K.*

Proof: Since f is continuous, f^{-1} takes open sets to open sets, and closed sets to closed sets. Pick any closed $C \subset \Upsilon$. Then

$$\begin{aligned}
& \limsup_{\epsilon \rightarrow 0} \epsilon \log \mathbb{P}\left(f(X_\epsilon) \in C\right) && (30.23)\\
= & \limsup_{\epsilon \rightarrow 0} \epsilon \log \mathbb{P}\left(X_\epsilon \in f^{-1}(C)\right) &&\\
\leq & -J(f^{-1}(C)) && (30.24)\\
= & -\inf_{x \in f^{-1}(C)} J(x) && (30.25)\\
= & -\inf_{y \in C} \inf_{x \in f^{-1}(y)} J(x) && (30.26)\\
= & -\inf_{y \in C} K(y) && (30.27)
\end{aligned}$$

as required. The argument for open sets in Υ is entirely parallel, establishing that K, as defined, is a raw rate function. By Lemma 404, K can be modified to be lower semi-continuous without affecting the LDP, i.e., we can make a rate function from it. If J is a good rate function, then it has compact level sets. But continuous functions take compact sets to compact sets, so $K = J \circ f^{-1}$ will also have compact level sets, i.e., it will also be a good rate function. □

There are a bunch of really common applications of the contraction principle, relating the large deviations at one level of description to those at coarser levels. To make the most frequent set of implications precise, let's recall a couple of definitions.

Definition 411 (Empirical Mean) *If $X_1, \ldots X_n$ are random variables in a common vector space Ξ, their* empirical mean *is $\overline{X}_n \equiv \frac{1}{n}\sum_{i=1}^{n} X_i$.*

We have already encountered this as the sample average or, in ergodic theory, the finite time average. (Notice that nothing is said about the X_i being IID, or even having a common expectation.)

Definition 412 (Empirical Distribution) *Let $X_1, \ldots X_n$ be random variables in a common measurable space Ξ (not necessarily a vector or metric space). The* empirical distribution *is $\hat{P}_n \equiv \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$, where δ_x is the probability measure that puts all its probability on the point x, i.e., $\delta_x(B) = \mathbf{1}_B(x)$. $\hat{P}_n$ is a random variable taking values in $\mathcal{P}(\Xi)$, the space of all probability measures on Ξ. (Cf. Example 10 in chapter 1 and Example 43 in chapter 4.) $\mathcal{P}(\Xi)$ is a metric space under any of several distances, and a complete separable metric space (i.e., Polish) under, for instance, the total variation metric.*

Definition 413 (Finite-Dimensional Empirical Distributions) *For each k, the* k-dimensional empirical distribution *is*

$$\hat{P}_n^k \equiv \frac{1}{n}\sum_{i=1}^{n} \delta_{(X_i, X_{i+1}, \ldots X_{i+k})} \tag{30.28}$$

where the addition of indices for the delta function is to be done modulo n, i.e., $\hat{P}_3^2 = \frac{1}{3}\left(\delta_{(X_1,X_2)} + \delta_{(X_2,X_3)} + \delta_{(X_3,X_1)}\right)$. $\hat{P}_n^k$ takes values in $\mathcal{P}\left(\Xi^k\right)$.

Definition 414 (Empirical Process Distribution) *With a finite sequence of random variables $X_1, \ldots X_n$, the* empirical process *is the periodic, infinite random sequence $\tilde{X}_n$ as the repetition of the sample without limit, i.e., $\tilde{X}_n(i) = X_{i \bmod n}$. If T is the shift operator on the sequence space, then the* empirical process distribution *is*

$$\hat{P}_n^\infty \equiv \frac{1}{n} \sum_{i-0}^{n-1} \delta_{T^i \tilde{X}_n} \tag{30.29}$$

$\hat{P}_n^\infty$ takes values in the space of infinite-dimensional distributions for one-sided sequences, $\mathcal{P}\left(\Xi^{\mathbb{N}}\right)$. In fact, it is always a stationary distribution, because by construction it is invariant under the shift T.

Be careful not to confuse this empirical process with the quite distinct empirical process of Examples 10 and 43.

Corollary 415 *The following chain of implications hold:*

i If the empirical process distribution obeys an LDP, so do all the finite-dimensional distributions.

ii If the n-dimensional distribution obeys an LDP, all $m < n$ dimensional distributions do.

iii If any finite-dimensional distribution obeys an LDP, the empirical distribution does.

iv If the empirical distribution obeys an LDP, the empirical mean does.

Proof: In each case, we obtain the lower-level statistic from the higher-level one by applying a continuous function, hence the contraction principle applies. For the distributions, the continuous functions are the projection operators of Chapter 2. □

Corollary 416 ("Tilted" LDP) *In set-up of Theorem 409, let $\mu_\epsilon = \mathcal{L}(X_\epsilon)$. Define the probability measures $\mu_{f,\epsilon}$ via*

$$\mu_{f,\epsilon}(B) \equiv \frac{\mathbf{E}\left[e^{f(X_\epsilon)/\epsilon}\mathbf{1}_B(X_\epsilon)\right]}{\mathbf{E}\left[e^{f(X_\epsilon)/\epsilon}\right]} \tag{30.30}$$

Then $Y_\epsilon \sim \mu_{f,\epsilon}$ obeys an LDP with rate $1/\epsilon$ and rate function

$$J_F(x) = -(f(x) - J(x)) + \sup_{y \in \Xi} f(y) - J(y) \tag{30.31}$$

PROOF: Define a set function $F_\epsilon(B) = \mathbf{E}\left[e^{f(X_\epsilon)/\epsilon}\mathbf{1}_B(X_\epsilon)\right]$; then $\mu_{f,\epsilon}(B) = F_\epsilon(B)/F_\epsilon(\Xi)$. From Varadhan's Lemma, we know that $F_\epsilon(\Xi)$ has asymptotic logarithm $\sup_{y\in\Xi} f(y) - J(y)$, so it is just necessary to show that

$$\begin{aligned} \limsup_{\epsilon} \epsilon \log F_\epsilon(B) &\leq \sup_{x\in \mathrm{cl}B} f(x) - J(x) && (30.32)\\ \liminf_{\epsilon} \epsilon \log F_\epsilon(B) &\geq \sup_{x\in \mathrm{int}B} f(x) - J(x) && (30.33)\end{aligned}$$

which can be done by imitating the proof of Varadhan's Lemma itself. □

Remark: "Tilting" here refers to some geometrical analogy which, in all honesty, has never made any sense to me.

Because the LDP is about exponential decay of probabilities, it is not surprising that several ways of obtaining it require a sort of exponential bound on the dispersion of the probability measure.

Definition 417 (Exponentially Tight) *The parameterized family of random variables X_ϵ, $\epsilon > 0$, is* exponentially tight *if, for every finite real M, there exists a compact set $C \subset \Xi$ such that*

$$\limsup_{\epsilon\rightarrow 0} \epsilon \log \mathbb{P}\left(X_\epsilon \not\in C\right) \leq -M \tag{30.34}$$

The first use of exponential tightness is a converse to the contraction principle: a high-level LDP is implied by the combination of a low-level LDP and high-level exponential tightness.

Theorem 418 (Inverse Contraction Principle) *If X_ϵ are exponentially tight, f is continuous and injective, and $Y_\epsilon = f(X_\epsilon)$ obeys an LDP with rate function K, then X_ϵ obeys an LDP with a good rate function $J(x) = K(f(x))$.*

PROOF: See Kallenberg, Theorem 27.11 (ii). Notice, by the way, that the proof of the upper bound on probabilities (i.e. that $\limsup \epsilon \log \mathbb{P}\left(X_\epsilon \in B\right) \leq -J(B)$ for closed $B \subseteq \Xi$) does not depend on exponential tightness, just the continuity of f. Exponential tightness is only needed for the lower bound. □

Theorem 419 (Bryc's Theorem) *If X_ϵ are exponentially tight, and, for all bounded continuous f, the limit*

$$\Lambda_f \equiv \lim_{\epsilon\rightarrow 0} \epsilon \log \mathbf{E}\left[e^{f(X_\epsilon/\epsilon)}\right] \tag{30.35}$$

exists, then X_ϵ obeys the LDP with good rate function

$$J(x) \equiv \sup_f f(x) - \Lambda_f \tag{30.36}$$

where the supremum extends over all bounded, continuous functions.

PROOF: See Kallenberg, Theorem 27.10, part (ii). □

Remark: This is a converse to Varadhan's Lemma.

Theorem 420 (Projective Limit) *Let $\Xi_1, \Xi_2, \ldots$ be a countable sequence of metric spaces, and let X_ϵ be a random sequence from this space. If, for every n, $X^n_\epsilon = \pi_n X_\epsilon$ obeys the LDP with good rate function J_n, then X_ϵ obeys the LDP with good rate function*

$$J(x) \equiv \sup_n J_n(\pi_n x) \tag{30.37}$$

PROOF: See Kallenberg, Theorem 27.12. □

Definition 421 (Exponentially Equivalent Random Variables) *Two families of random variables, X_ϵ and Y_ϵ, taking values in a common metric space, are* exponentially equivalent *when, for all positive δ,*

$$\lim_{\epsilon \rightarrow 0} \epsilon \log \mathbb{P}\left(d(X_\epsilon, Y_\epsilon) > \delta\right) = -\infty \tag{30.38}$$

Lemma 422 *If X_ϵ and Y_ϵ are exponentially equivalent, one of them obeys the LDP with a good rate function J iff the other does as well.*

PROOF: It is enough to prove that the LDP for X_ϵ implies the LDP for Y_ϵ, with the same rate function. (Draw a truth-table if you don't believe me!) As usual, first we'll get the upper bound, and then the lower.

Pick any closed set C, and let C_δ be its closed δ neighborhood, i.e., $C_\delta = \{x : \ \exists y \in C, \ d(x, y) \leq \delta\}$. Now

$$\mathbb{P}\left(Y_\epsilon \in C_\delta\right) \leq \mathbb{P}\left(X_\epsilon \in C_\delta\right) + \mathbb{P}\left(d(X_\epsilon, Y_\epsilon) > \delta\right) \tag{30.39}$$

Using Eq. 30.38 from Definition 421, the LDP for X_ϵ, and Lemma 402

$$\begin{aligned}
& \limsup \epsilon \log \mathbb{P}\left(Y_\epsilon \in C\right) && (30.40)\\
\leq \quad & \limsup \epsilon \log \mathbb{P}\left(X_\epsilon \in C_\delta\right) + \epsilon \log \mathbb{P}\left(d(X_\epsilon, Y_\epsilon) > \delta\right) && \\
\leq \quad & \limsup \epsilon \log \mathbb{P}\left(X_\epsilon \in C_\delta\right) \vee \limsup \epsilon \log \mathbb{P}\left(d(X_\epsilon, Y_\epsilon) > \delta\right) && (30.41)\\
\leq \quad & -J(C_\delta) \vee -\infty && (30.42)\\
= \quad & -J(C_\delta) && (30.43)
\end{aligned}$$

Since J is a good rate function, we have $J(C_\delta) \uparrow J(C)$ as $\delta \downarrow 0$; since δ was arbitrary to start with,

$$\limsup \epsilon \log \mathbb{P}\left(Y_\epsilon \in C\right) \leq -J(C) \tag{30.44}$$

As usual, to obtain the lower bound on open sets, pick any open set O and any point $x \in O$. Because O is open, there is a $\delta > 0$ such that, for some open neighborhood U of x, not only is $U \subset O$, but $U_\delta \subset O$. In which case, we can say that

$$\mathbb{P}\left(X_\epsilon \in U\right) \leq \mathbb{P}\left(Y_\epsilon \in O\right) + \mathbb{P}\left(d(X_\epsilon, Y_\epsilon) > h\right) \tag{30.45}$$

Proceeding as for the upper bound,

$$\begin{aligned} -J(x) &\leq -J(U) && (30.46)\\ &\leq \liminf \epsilon \log \mathbb{P}\left(X_\epsilon \in U\right) && (30.47)\\ &\leq \liminf \epsilon \log \mathbb{P}\left(Y_\epsilon \in O\right) \vee \limsup \epsilon \log \mathbb{P}\left(d(X_\epsilon, Y_\epsilon) > \delta\right) && (30.48)\\ &= \liminf \epsilon \log \mathbb{P}\left(Y_\epsilon \in O\right) && (30.49) \end{aligned}$$

(Notice that the initial arbitrary choice of δ has dropped out.) Taking the supremum over all x gives $-J(O) \leq \liminf \epsilon \log \mathbb{P}\left(Y_\epsilon \in O\right)$, as required. □

Chapter 31

Large Deviations for IID Sequences: The Return of Relative Entropy

> Section 31.1 introduces the exponential version of the Markov inequality, which will be our major calculating device, and shows how it naturally leads to both the cumulant generating function and the Legendre transform, which we should suspect (correctly) of being the large deviations rate function. We also see the reappearance of relative entropy, as the Legendre transform of the cumulant generating *functional* of distributions.
>
> Section 31.2 proves the large deviations principle for the empirical mean of IID sequences in finite-dimensional Euclidean spaces (Cramér's Theorem).
>
> Section 31.3 proves the large deviations principle for the empirical distribution of IID sequences in Polish spaces (Sanov's Theorem), using Cramér's Theorem for a well-chosen collection of bounded continuous functions on the Polish space, and the tools of Section 30.2. Here the rate function is the relative entropy.
>
> Section 31.4 proves that even the infinite-dimensional empirical process distribution of an IID sequence in a Polish space obeys the LDP, with the rate function given by the relative entropy rate.

The usual approach in large deviations theory is to establish an LDP for some comparatively tractable basic case through explicit calculations, and then use the machinery of Section 30.2 to extend it to LDPs for more complicated cases. This chapter applies this strategy to IID sequences.

31.1 Cumulant Generating Functions and Relative Entropy

Suppose the only inequality we knew in probability theory was Markov's inequality, $\mathbb{P}(X \geq a) \leq \mathbf{E}[X]/a$ when $X \geq 0$. How might we extract an exponential probability bound from it? Well, for any real-valued variable, e^{tX} is positive, so we can say that $\mathbb{P}(X \geq a) = \mathbb{P}\left(e^{tX} \geq e^{ta}\right) \leq \mathbf{E}\left[e^{tX}\right]/e^{ta}$. $\mathbf{E}\left[e^{tX}\right]$ is of course the moment generating function of X. It has the nice property that addition of independent random variables leads to multiplication of their moment generating functions, as $\mathbf{E}\left[e^{t(X_1+X_2)}\right] = \mathbf{E}\left[e^{tX_1}e^{tX_2}\right] = \mathbf{E}\left[e^{tX_1}\right]\mathbf{E}\left[e^{tX_2}\right]$ if $X_1 \perp\!\!\!\perp X_2$. If $X_1, X_2, \ldots$ are IID, then we can get a deviation bound for their sample mean $\overline{X}_n$ through the moment generating function:

$$\begin{aligned}
\mathbb{P}\left(\overline{X}_n \geq a\right) &= \mathbb{P}\left(\sum_{i=1}^{n} X_i \geq na\right) \\
\mathbb{P}\left(\overline{X}_n \geq a\right) &\leq e^{-nta}\left(\mathbf{E}\left[e^{tX_1}\right]\right)^n \\
\frac{1}{n}\log \mathbb{P}\left(\overline{X}_n \geq a\right) &\leq -ta + \log \mathbf{E}\left[e^{tX_1}\right] \\
&\leq \inf_t -ta + \log \mathbf{E}\left[e^{tX_1}\right] \\
&\leq -\sup_t ta - \log \mathbf{E}\left[e^{tX_1}\right]
\end{aligned}$$

This suggests that the functions $\log \mathbf{E}\left[e^{tX}\right]$ and $\sup ta - \log \mathbf{E}\left[e^{tX}\right]$ will be useful to us. Accordingly, we encapsulate them in a pair of definitions.

Definition 423 (Cumulant Generating Function) *The* cumulant generating function *of a random variable X in $\mathbb{R}^d$ is a function $\Lambda : \mathbb{R}^d \mapsto \mathbb{R}$,*

$$\Lambda(t) \equiv \log \mathbf{E}\left[e^{t \cdot X}\right] \tag{31.1}$$

Definition 424 (Legendre Transform) *The* Legendre transform *of a real-valued function f on $\mathbb{R}^d$ is another real-valued function on $\mathbb{R}^d$,*

$$f^*(x) \equiv \sup_{t \in \mathbb{R}^d} t \cdot x - f(t) \tag{31.2}$$

The definition of cumulant generating functions and their Legendre transforms can be extended to arbitrary spaces where some equivalent of the inner product (a real-valued form, bilinear in its two arguments) makes sense; f and f^* then must take arguments from the complementary spaces.

Legendre transforms are particularly important in convex analysis[1], since convexity is preserved by taking Legendre transforms. If f is not convex initially, then f^{**} is (in one dimension) something like the greatest convex lower bound on f; made precise, this statement even remains true in higher dimensions. I make these remarks because of the following fact:

[1] See Rockafellar (1970), or, more concisely, Ellis (1985, ch. VI).

Lemma 425 *The cumulant generating function $\Lambda(t)$ is convex.*

PROOF: Simple calculation, using Hölder's inequality in one step:

$$\begin{aligned}
\Lambda(at+bu) &= \log \mathbf{E}\left[e^{(at+bu)X}\right] && (31.3)\\
&= \log \mathbf{E}\left[e^{atX}e^{buX}\right] && (31.4)\\
&= \log \mathbf{E}\left[\left(e^{tX}\right)^{a}\left(e^{uX}\right)^{b}\right] && (31.5)\\
&\leq \log \left(\mathbf{E}\left[e^{tX}\right]\right)^{a}\left(\mathbf{E}\left[e^{buX}\right]\right)^{b} && (31.6)\\
&= a\Lambda(t)+b\Lambda(u) && (31.7)
\end{aligned}$$

which proves convexity. □

Our previous result, then, is easily stated: if the X_i are IID in $\mathbb{R}$, then

$$\mathbb{P}\left(\overline{X}_n \geq a\right) \leq \Lambda^*(a) \tag{31.8}$$

where $\Lambda^*(a)$ is the Legendre transform of the cumulant generating function of the X_i. This elementary fact is, surprisingly enough, the foundation of the large deviations principle for empirical means.

The notion of cumulant generating functions can be extended to probability measures, and this will be useful when dealing with large deviations of empirical distributions. The definitions follow the pattern one would expect from the complementarity between probability measures and bounded continuous functions.

Definition 426 (Cumulant Generating Functional) *Let X be a random variable on a metric space Ξ, with distribution μ, and let $C_b(\Xi)$ be the class of all bounded, continuous, real-valued functions on Ξ. Then the* cumulant-generating functional $\Lambda : C_b(\Xi) \mapsto \mathbb{R}$ *is*

$$\Lambda(f) \equiv \log \mathbf{E}\left[e^{f(X)}\right] \tag{31.9}$$

Definition 427 *The* Legendre transform *of a real-valued functional F on $C_b(\Xi)$ is*

$$F^*(\nu) \equiv \sup_{f \in C_b(\Xi)} \mathbf{E}_{\nu}\left[f\right] - \Lambda(f) \tag{31.10}$$

where $\nu \in \mathcal{P}\left(\Xi\right)$, the set of all probability measures on Ξ.

Lemma 428 (Donsker and Varadhan) *The Legendre transform of the cumulant generating functional is the relative entropy:*

$$\Lambda^*(\nu) = D\left(\nu \| \mu\right) \tag{31.11}$$

PROOF: First of all, notice that the supremum in Eq. 31.10 can be taken over all bounded *measurable* functions, not just functions in C_b, since C_b is dense. This will let us use indicator functions and simple functions in the subsequent argument.

If $\nu \not\ll \mu$, then $D(\nu\|\mu) = \infty$. But then there is also a set, call it B, with $\mu(B) = 0$, $\nu(B) > 0$. Take $f_n = n\mathbf{1}_B$. Then $\mathbf{E}_\nu[f_n] - \Lambda(f_n) = n\nu(B) - 0$, which can be made arbitrarily large by taking n arbitrarily large, hence the supremum in Eq. 31.10 is ∞.

If $\nu \ll \mu$, then show that $D(\nu\|\mu) \leq \Lambda^*(\nu)$ and $D(\nu\|\mu) \geq \Lambda^*(\nu)$, so they must be equal. To get the first inequality, start with the observation then $\frac{d\nu}{d\mu}$ exists, so set $f = \log\frac{d\nu}{d\mu}$, which is measurable. Then $D(\nu\|\mu)$ is $\mathbf{E}_\nu[f] - \log\mathbf{E}_\mu\left[e^f\right]$. If f is bounded, this shows that $D(\nu\|\mu) \leq \Lambda^*(\nu)$. If f is not bounded, approximate it by a sequence of bounded, measurable functions f_n with $\mathbf{E}_\mu\left[e^{f_n}\right] \to 1$ and $\mathbf{E}_\nu[f_n] \to \mathbf{E}_\nu[f_n]$, again concluding that $D(\nu\|\mu) \leq \Lambda^*(\nu)$.

To go the other way, first consider the special case where $\mathcal{X}$ is finite, and so generated by a partition, with cells $B_1, \ldots B_n$. Then all measurable functions are simple functions, and $\mathbf{E}_\nu[f] - \Lambda(f)$ is

$$g(f) = \sum_{i=1}^{n} f_i\nu(B_i) - \log\sum_{i=1}^{n} e^{f_i}\mu(B_i) \tag{31.12}$$

Now, $g(f)$ is concave on all the f_i, and

$$\frac{\partial g(f)}{\partial f_i} = \nu(B_i) - \frac{1}{\sum_{i=1}^{n} e^{f_i}\mu(B_i)}\mu(B_i)e^{f_i} \tag{31.13}$$

Setting this equal to zero,

$$\frac{\nu(B_i)}{\mu(B_i)} = \frac{1}{\sum_{i=1}^{n}\mu(B_i)e^{f_i}}e^{f_i} \tag{31.14}$$

$$\log\frac{\nu(B_i)}{\mu(B_i)} = f_i \tag{31.15}$$

gives the maximum value of $g(f)$. (Remember that $0\log 0 = 0$.) But then $g(f) = D(\nu\|\mu)$. So $\Lambda^*(\nu) \leq D(\nu\|\mu)$ when the σ-algebra is finite. In the general case, consider the case where f is a simple function. Then $\sigma(f)$ is finite, and $\mathbf{E}_\nu[f] - \log\mathbf{E}_\mu\left[e^f\right] \leq D(\nu\|\mu)$ follows by the finite case and smoothing. Finally, if f is not simple, but is bounded and measurable, there is a simple h such that $\mathbf{E}_\nu[f] - \log\mathbf{E}_\mu\left[e^f\right] \leq \mathbf{E}_\nu[h] - \log\mathbf{E}_\mu\left[e^h\right]$, so

$$\sup_{f\in C_b(\Xi)} \mathbf{E}_\nu[f] - \log\mathbf{E}_\mu\left[e^f\right] \leq D(\nu\|\mu) \tag{31.16}$$

which completes the proof. □

31.2 Large Deviations of the Empirical Mean in $\mathbb{R}^d$

Historically, the oldest and most important result in large deviations is that the empirical mean of an IID sequence of real-valued random variables obeys a large deviations principle with rate n; the oldest version of this proposition goes back to Harald Cramér in the 1930s, and so it is known as Cramér's theorem, even though the modern version, which is both more refined technically and works in arbitrary finite-dimensional Euclidean spaces, is due to Varadhan in the 1960s.

Theorem 429 (Cramér's Theorem) *If X_i are IID random variables in $\mathbb{R}^d$, and $\Lambda(t) < \infty$ for all $t \in \mathbb{R}^d$, then their empirical mean obeys an LDP with rate n and good rate function $\Lambda^*(x)$.*

PROOF: The proof has three parts. First, the upper bound for closed sets; second, the lower bound for open sets, under an additional assumption on $\Lambda(t)$; third and finally, lifting of the assumption on Λ by means of a perturbation argument (related to Lemma 422).

To prove the upper bound for closed sets, we first prove the upper bound for sufficiently small balls around arbitrary points. Then, we take our favorite closed set, and divide it into a compact part close to the origin, which we can cover by a finite number of closed balls, and a remainder which is far from the origin and of low probability.

First the small balls of low probability. Because $\Lambda^*(x) = \sup_u u \cdot x - \Lambda(u)$, for any $\epsilon > 0$, we can find some u such that $u \cdot x - \Lambda(x) > \min 1/\epsilon, \Lambda^*(x) - \epsilon$. (Otherwise, $\Lambda^*(x)$ would not be the *least* upper bound.) Since $u \cdot x$ is continuous in x, it follows that there exists some open ball B of positive radius, centered on x, within which $u \cdot y - \Lambda(x) > \min 1/\epsilon, \Lambda^*(x) - \epsilon$, or $u \cdot y > \Lambda(x) + \min 1/\epsilon, \Lambda^*(x) - \epsilon$. Now use the exponential Markov inequality to get

$$\begin{aligned} \mathbb{P}\left(\overline{X}_n \in B\right) &\leq \mathbf{E}\left[e^{u \cdot n\overline{X}_n - n \inf_{y \in B} u \cdot y}\right] && (31.17) \\ &\leq e^{-n\left(\min \frac{1}{\epsilon}, \Lambda^*(x) - \epsilon\right)} && (31.18) \end{aligned}$$

which is small. To get the the compact set near the origin of high probability, use the exponential decay of the probability at large $\|x\|$. Since $\Lambda(t) < \infty$ for all t, $\Lambda^*(x) \to \infty$ as $\|x\| \to \infty$. So, using (once again) the exponential Markov inequality, for every $\epsilon > 0$, there must exist an $r > 0$ such that

$$\frac{1}{n} \log \mathbb{P}\left(\left\|\overline{X}_n\right\| > r\right) \leq -\frac{1}{\epsilon} \qquad (31.19)$$

for all n.

Now pick your favorite closed measurable set $C \in \mathcal{B}^d$. Then $C \cap \{x : \|x\| \leq r\}$ is compact, and I can cover it by m balls $B_1, \ldots B_m$, with centers $x_1, \ldots x_m$, of the sort built in the previous paragraph. So I can apply a union bound to

$\mathbb{P}\left(\overline{X}_n \in C\right)$, as follows.

$$\begin{aligned}
& \mathbb{P}\left(\overline{X}_n \in C\right) && (31.20)\\
= \ & \mathbb{P}\left(\overline{X}_n \in C \cap \{x : \ \|x\| \leq r\}\right) + \mathbb{P}\left(\overline{X}_n \in C \cap \{x : \ \|x\| > r\}\right) && \\
\leq \ & \mathbb{P}\left(\overline{X}_n \in \bigcup_{i=1}^{m} B_i\right) + \mathbb{P}\left(\left\|\overline{X}_n\right\| > r\right) && (31.21)\\
\leq \ & \left(\sum_{i=1}^{m} \mathbb{P}\left(\overline{X}_n \in B_i\right)\right) + \mathbb{P}\left(\left\|\overline{X}_n\right\| > r\right) && (31.22)\\
\leq \ & \left(\sum_{i=1}^{m} e^{-n\left(\min \frac{1}{\epsilon}, \Lambda^*(x_i) - \epsilon\right)}\right) + e^{-n\frac{1}{\epsilon}} && (31.23)\\
\leq \ & (m+1) e^{-n\left(\min \frac{1}{\epsilon}, \Lambda^*(C) - \epsilon\right)} && (31.24)
\end{aligned}$$

with $\Lambda^*(C) = \inf_{x \in C} \Lambda^*(x)$, as usual. So if I take the log, normalize, and go to the limit, I have

$$\begin{aligned}
\limsup_{n} \frac{1}{n} \log \mathbb{P}\left(\overline{X}_n \in C\right) & \leq -\min \frac{1}{\epsilon}, \Lambda^*(C) - \epsilon && (31.25)\\
& \leq -\Lambda^*(C) && (31.26)
\end{aligned}$$

since ϵ was arbitrary to start with, and I've got the upper bound for closed sets.

To get the lower bound for open sets, pick your favorite open set $O \in \mathcal{B}^d$, and your favorite $x \in O$. Suppose, for the moment, that $\Lambda(t)/\|t\| \to \infty$ as $\|t\| \to \infty$. (This is the growth condition mentioned earlier, which we will left at the end of the proof.) Then, because $\Lambda(t)$ is smooth, there is some u such that $\nabla\Lambda(u) = x$. (You will find it instructive to draw the geometry here.) Now let Y_i be a sequence of IID random variables, whose probability law is given by

$$\mathbb{P}\left(Y_i \in B\right) = \frac{\mathbf{E}\left[e^{uX}\mathbf{1}_B(X)\right]}{\mathbf{E}\left[e^{uX}\right]} = e^{-\Lambda(u)}\mathbf{E}\left[e^{uX}\mathbf{1}_B(X)\right] \tag{31.27}$$

It is not hard to show, by manipulating the cumulant generating functions, that $\Lambda_Y(t) = \Lambda_X(t+u) - \Lambda_X(u)$, and consequently that $\mathbf{E}\left[Y_i\right] = x$. I construct these Y to allow me to pull the following trick, which works if $\epsilon > 0$ is sufficiently small that the first inequality holds (and I can always chose small enough ϵ):

$$\begin{aligned}
\mathbb{P}\left(\overline{X}_n \in O\right) & \geq \mathbb{P}\left(\left\|\overline{X}_n - x\right\| < \epsilon\right) && (31.28)\\
& = e^{n\Lambda(u)}\mathbf{E}\left[e^{-nu\overline{Y}_n}\mathbf{1}\{y : \ \|y - x\| < \epsilon\}(\overline{Y}_n)\right] && (31.29)\\
& \geq e^{n\Lambda(u) - nu\cdot x - n\epsilon\|u\|}\mathbb{P}\left(\left\|\overline{Y}_n - x\right\| < \epsilon\right) && (31.30)
\end{aligned}$$

By the strong law of large numbers, $\mathbb{P}\left(\left\|\overline{Y}_n - x\right\| < \epsilon\right) \rightarrow 1$ for all ϵ, so

$$\begin{aligned}
\liminf \frac{1}{n} \log \mathbb{P}\left(\overline{X}_n \in O\right) & \geq & \Lambda(u) - u \cdot x - \epsilon\|u\| & \quad (31.31)\\
& \geq & -\Lambda^*(x) - \epsilon\|u\| & \quad (31.32)\\
& \geq & -\Lambda^*(x) & \quad (31.33)\\
& \geq & -\inf_{x \in O} \Lambda^*(x) = -\Lambda(O) & \quad (31.34)
\end{aligned}$$

as required. This proves the LDP, as required, if $\Lambda(t)/\|t\| \rightarrow \infty$ as $\|t\| \rightarrow \infty$.

Finally, to lift the last-named restriction (which, remember, only affected the lower bound for open sets), introduce a sequence Z_i of IID standard Gaussian variables, i.e. $Z_i \sim \mathcal{N}(0, I)$, which are completely independent of the X_i. It is easily calculated that the cumulant generating function of the Z_i is $\|t\|^2/2$, so that $\overline{Z}_n$ satisfies the LDP. Another easy calculation shows that $X_i + \sigma Z_i$ has cumulant generating function $\Lambda_X(t) + \frac{\sigma^2}{2}\|t\|^2$, which again satisfies the previous condition. Since $\Lambda_{X+\sigma Z} \geq \Lambda_X$, $\Lambda^*_X \geq \Lambda^*_{X+\sigma Z}$. Now, once again pick any open set O, and any point $x \in O$, and an ϵ sufficiently small that all points within a distance 2ϵ of x are also in O. Since the LDP applies to $X + \sigma Z$,

$$\begin{aligned}
\mathbb{P}\left(\left\|\overline{X}_n + \sigma\overline{Z}_n - x\right\| \leq \epsilon\right) & \geq & -\Lambda^*_{X+\sigma Z}(x) & \quad (31.35)\\
& \geq & -\Lambda^*_X(x) & \quad (31.36)
\end{aligned}$$

On the other hand, basic probability manipulations give

$$\begin{aligned}
\mathbb{P}\left(\left\|\overline{X}_n + \sigma\overline{Z}_n - x\right\| \leq \epsilon\right) & \leq & \mathbb{P}\left(\overline{X}_n \in O\right) + \mathbb{P}\left(\sigma\left\|\overline{Z}_n\right\| \geq \epsilon\right) & \quad (31.37)\\
& \leq & 2 \max \mathbb{P}\left(\overline{X}_n \in O\right), \mathbb{P}\left(\sigma\left\|\overline{Z}_n\right\| \geq \epsilon\right) & (31.38)
\end{aligned}$$

Taking the liminf of the normalized log of both sides,

$$\begin{aligned}
& \liminf \frac{1}{n} \log \mathbb{P}\left(\left\|\overline{X}_n + \sigma\overline{Z}_n - x\right\| \leq \epsilon\right) & \quad (31.39)\\
\leq & \liminf \frac{1}{n} \log\left(\max \mathbb{P}\left(\overline{X}_n \in O\right), \mathbb{P}\left(\sigma\left\|\overline{Z}_n\right\| \geq \epsilon\right)\right) & \\
\leq & \liminf \frac{1}{n} \log \mathbb{P}\left(\overline{X}_n \in O\right) \vee \left(-\frac{\epsilon^2}{2\sigma^2}\right) & \quad (31.40)\\
& & \quad (31.41)
\end{aligned}$$

Since σ was arbitrary, we can let it go to zero, and obtain

$$\begin{aligned}
\liminf \frac{1}{n} \log \mathbb{P}\left(\overline{X}_n \in O\right) & \geq & -\Lambda^*_X(x) & \quad (31.42)\\
& \geq & -\Lambda^*_X(O) & \quad (31.43)
\end{aligned}$$

as required. □

31.3 Large Deviations of the Empirical Measure in Polish Spaces

The Polish space setting is, apparently, more general than $\mathbb{R}^d$, but we will represent distributions on the Polish space in terms of the expectation of a separating set of functions, and then appeal to the Euclidean result.

Proposition 430 *Any Polish space S can be represented as a Borel subset of a compact metric space, namely $[0,1]^{\mathbb{N}} \equiv M$.*

PROOF: See, for instance, Appendix A of Kallenberg. □

Strictly speaking, there should be a function mapping points from S to points in M. However, since this is an embedding, I will silently omit it in what follows.

Proposition 431 *$C_b(M)$ has a countable dense separating set $\mathcal{F} = f_1, f_2, \ldots$.*

PROOF: See Kallenberg again. □

Because $\mathcal{F}$ is separating, to specify a probability distribution on K is equivalent to specifying the expectation value of all the functions in $\mathcal{F}$. Write $f_1^d(X)$ to abbreviate the d-dimensional vector $(f_1(X), f_2(X), \ldots f_d(X))$, and $f_1^\infty(X)$ to abbreviate the corresponding infinite-dimensional vector.

Lemma 432 *Empirical means are expectations with respect to empirical measure. That is, let f be a real-valued measurable function and $Y_i = f(X_i)$. Then $\overline{Y}_n = \mathbf{E}_{\hat{P}_n}[f(X)]$.*

PROOF: Direct calculation.

$$\overline{Y}_n \equiv \frac{1}{n}\sum_{i=1}^{n} f(X_i) \quad (31.44)$$

$$= \frac{1}{n}\sum_{i=1}^{n} \mathbf{E}_{\delta_{X_i}}[f(X)] \quad (31.45)$$

$$\equiv \mathbf{E}_{\hat{P}_n}[f(X)] \quad (31.46)$$

□

Lemma 433 *Let X_i be a sequence of IID random variables in a Polish space Ξ. For each d, the sequence of vectors $(\mathbf{E}_{\hat{P}_n}[f_1], \ldots \mathbf{E}_{\hat{P}_n}[f_d])$ obeys the LDP with rate n and good rate function J_d.*

PROOF: For each d, the sequence of vectors $(f_1(X_i), \ldots f_d(X_i))$ are IID, so, by Cramér's Theorem (429), their empirical mean obeys the LDP with rate n and good rate function

$$J_d(x) = \sup_{t \in \mathbb{R}^d} t \cdot x - \log \mathbf{E}\left[e^{t \cdot f_1^d(X)}\right] \quad (31.47)$$

But, by Lemma 432, the empirical means are expectations over the empirical distributions, so the latter must also obey the LDP, with the same rate and rate function. □

Notice, incidentally, that the fact that the $f_i \in \mathcal{F}$ isn't relevant for the proof of the lemma; it will however be relevant for the proof of the theorem.

Theorem 434 (Sanov's Theorem) *Let X_i, $i \in \mathbb{N}$, be IID random variables in a Polish space Ξ, with common probability measure μ. Then the empirical distributions $\hat{P}_n$ obey an LDP with rate n and good rate function $J(\nu) = D(\nu\|\mu)$.*

PROOF: Combining Lemma 433 and Theorem 420, we see that $\mathbf{E}_{\hat{P}_n}\left[f_1^\infty(X)\right]$ obeys the LDP with rate n and good rate function

$$\begin{aligned} J(x) &= \sup_d J_d(\pi_d x) &(31.48)\\ &= \sup_d \sup_{t\in\mathbb{R}^d} t\cdot\pi_d x - \log \mathbf{E}\left[e^{t\cdot f_1^d(X)}\right] &(31.49)\end{aligned}$$

Since $\mathcal{P}(M)$ is compact (so all random sequences in it are exponentially tight), and the mapping from $\nu \in \mathcal{P}(M)$ to $\mathbf{E}_\nu\left[f_1^\infty\right] \in \mathbb{R}^{\mathbb{N}}$ is continuous, apply the inverse contraction principle (Theorem 418) to get that $\hat{P}_n$ satisfies the LDP with good rate function

$$\begin{aligned} J(\nu) &= J(\mathbf{E}_\nu\left[f_1^\infty\right]) &(31.50)\\ &= \sup_d \sup_{t\in\mathbb{R}^d} t\cdot \mathbf{E}_\nu\left[f_1^d\right] - \log \mathbf{E}_\mu\left[e^{t\cdot f_1^d(X)}\right] &(31.51)\\ &= \sup_{f\in\mathrm{span}\mathcal{F}} \mathbf{E}_\nu\left[f\right] - \Lambda(f) &(31.52)\\ &= \sup_{f\in C_b(M)} \mathbf{E}_\nu\left[f\right] - \Lambda(f) &(31.53)\\ &= D(\nu\|\mu) &(31.54)\end{aligned}$$

Notice however that this is an LDP in the space $\mathcal{P}(M)$, not in $\mathcal{P}(\Xi)$. However, the embedding taking $\mathcal{P}(\Xi)$ to $\mathcal{P}(M)$ is continuous, and it is easily verified (see Lemma 27.17 in Kallenberg) that $\hat{P}_n$ is exponentially tight in $\mathcal{P}(\Xi)$, so another application of the inverse contraction principle says that $\hat{P}_n$ must obey the LDP in the restricted space $\mathcal{P}(\Xi)$, and with the same rate. □

31.4 Large Deviations of the Empirical Process in Polish Spaces

A fairly straightforward modification of the proof for Sanov's theorem establishes a large deviations principle for the finite-dimensional empirical distributions of an IID sequence.

Corollary 435 *Let X_i be an IID sequence in a Polish space Ξ, with common measure μ. Then, for every finite positive integer k, the k-dimensional empirical* *distribution $\hat{P}_n^k$, obeys an LDP with rate n and good rate function $J_k(\nu) = D(\nu \| \pi_{k-1}\nu \otimes \mu)$ if $\nu \in \mathcal{P}\left(\Xi^k\right)$ is shift invariant, and $J(\nu) = \infty$ otherwise.*

This leads to the following important generalization.

Theorem 436 *If X_i are IID in a Polish space, with a common measure μ, then the empirical process distribution $\hat{P}_n^\infty$ obeys an LDP with rate n and good rate function $J_\infty(\nu) = d(\nu\|\mu^\infty)$, the relative entropy rate, if ν is a shift-invariant probability measure, and $= \infty$ otherwise.*

Proof: By Corollary 435 and the projective limit theorem 420, $\hat{P}_n^\infty$ obeys an LDP with rate n and good rate function

$$J_\infty(\nu) = \sup_k J_k(\pi_k \nu) = \sup_k D(\pi_k\nu \| \pi_{k-1}\nu \otimes \mu) \tag{31.55}$$

But, applying the chain rule for relative entropy (Lemma 363),

$$\begin{aligned} D(\pi_n\nu\|\mu^n) &= D(\pi_n\nu\|\pi_{n-1}\nu \otimes \mu) + D\left(\pi_{n-1}\nu\|\mu^{n-1}\right) && (31.56)\\ &= \sum_{k=1}^{n} D(\pi_k\nu\|\pi_{k-1}\nu\otimes\mu) && (31.57)\\ \lim \frac{1}{n} D(\pi_n\nu\|\mu^n) &= \lim \frac{1}{n}\sum_{k=1}^{n} D(\pi_k\nu\|\pi_{k-1}\nu\otimes\mu) && (31.58)\\ &= \sup_k D(\pi_k\nu\|\pi_{k-1}\nu\otimes\mu) && (31.59)\end{aligned}$$

But $\lim n^{-1} D(\pi_n\nu\|\mu^n)$ is the relative entropy rate, $d(\nu\|\mu^\infty)$, and we've already identified the right-hand side as the rate function. □

The strength of Theorem 436 lies in the fact that, via the contraction principle (Theorem 410), it implies that the LDP holds for any continuous function of the empirical process distribution. This in particular includes the finite-dimensional distributions, the empirical mean, functions of finite-length trajectories, etc. Moreover, Theorem 410 also provides a means to calculate the rate function for all these quantities.

Chapter 32

Large Deviations for Markov Sequences

This chapter establishes large deviations principles for Markov sequences as natural consequences of the large deviations principles for IID sequences in Chapter 31. (LDPs for continuous-time Markov processes will be treated in the chapter on Freidlin-Wentzell theory.)

Section 32.1 uses the exponential-family representation of Markov sequences to establish an LDP for the two-dimensional empirical distribution ("pair measure"). The rate function is a relative entropy.

Section 32.2 extends the results of Section 32.1 to other observables for Markov sequences, such as the empirical process and time averages of functions of the state.

For the whole of this chapter, let $X_1, X_2, \ldots$ be a homogeneous Markov sequence, taking values in a Polish space Ξ, with transition probability kernel μ, and initial distribution ν and invariant distribution ρ. If Ξ is not discrete, we will assume that ν and ρ have densities n and r with respect to some reference measure, and that $\mu(x, dy)$ has density $m(x, y)$ with respect to that same reference measure, for all x. (LDPs can be proved for Markov sequences without such density assumptions — see, e.g., Ellis (1988) — but the argument is more complicated.)

32.1 Large Deviations for Pair Measure of Markov Sequences

It is perhaps not sufficiently appreciated that Markov sequences form exponential families (Billingsley, 1961; Küchler and Sørensen, 1997). Suppose Ξ is discrete. Then

$$\begin{aligned}
\mathbb{P}\left(X_1^n = x_1^t\right) &= \nu(x_1)\prod_{i=1}^{t-1}\mu(x_i, x_{i+1}) && (32.1)\\
&= \nu(x_1)e^{\sum_{i=1}^{t-1}\log\mu(x_i, x_{i+1})} && (32.2)\\
&= \nu(x_1)e^{\sum_{x,y\in\Xi^2} T_{x,y}(x_1^t)\log\mu(x,y)} && (32.3)
\end{aligned}$$

where $T_{x,y}(x_1^t)$ counts the number of times the state y follows the state x in the sequence x_1^t, i.e., it gives the *transition counts*. What we have just established is that the Markov chains on Ξ with a given initial distribution form an exponential family, whose natural sufficient statistics are the transition counts, and whose natural parameters are the logarithms of the transition probabilities.

(If Ξ is not continuous, but we make the density assumptions mentioned at the beginning of this chapter, we can write

$$\begin{aligned}
p_{X_1^t}(x_1^t) &= n(x_1)\prod_{i=1}^{t-1} m(x_i, x_{i+1}) && (32.4)\\
&= n(x_1)e^{\int_{\Xi^2} dT(x_1^t)\log m(x,y)} && (32.5)
\end{aligned}$$

where now $T(x_1^t)$ puts probability mass $\frac{1}{n-1}$ at x, y for every i such that $x_i = x$, $x_{i+1} = y$.)

We can use this exponential family representation to establish the following basic theorem.

Theorem 437 *Let X_i be a Markov sequence obeying the assumptions set out at the beginning of this chapter, and furthermore that $\mu(x,y)/\rho(y)$ is bounded above (in the discrete-state case) or that $m(x,y)/r(y)$ is bounded above (in the continuous-state case). Then the two-dimensional empirical distribution ("pair measure") $\hat{P}_t^2$ obeys an LDP with rate n and with rate function $J_2(\psi) = D\left(\psi \| \pi_1\psi \times \mu\right)$ if ν is shift-invariant, $J(\nu) = \infty$ otherwise.*

Proof: I will just give the proof for the discrete case, since the modifications for the continuous case are straightforward (given the assumptions made about densities), largely a matter of substituting Roman letters for Greek ones.

First, modify the representation of the probabilities in Eq. 32.3 slightly, so that it refers directly to $\hat{P}_t^2$ (as laid down in Definition 413), rather than to the transition counts.

$$\begin{aligned}
\mathbb{P}\left(X_1^t = x_1^t\right) &= \frac{\nu(x_1)}{\mu(x_t, x_1)}e^{t\sum_{x,y\in\Xi}\hat{P}_t^2(x,y)\log\mu(x,y)} && (32.6)\\
&= \frac{\nu(x_1)}{\mu(x_t, x_1)}e^{n\mathbf{E}_{\hat{P}_t^2}[\log\mu(X,Y)]} && (32.7)
\end{aligned}$$

Now construct a sequence of IID variables Y_i, all distributed according to ρ, the invariant measure of the Markov chain:

$$\mathbb{P}\left(Y_1^t = y_1^t\right) = e^{n\mathbf{E}_{\hat{P}_t^2}[\log\rho(Y)]} \qquad (32.8)$$

The ratio of these probabilities is the Radon-Nikodym derivative:

$$\frac{d\mathbb{P}_X}{d\mathbb{P}_Y}(x_1^t) = \frac{\nu(x_1)}{\mu(x_t, x_1)} e^{t\mathbf{E}_{\hat{P}_n^2}\left[t \log \frac{\mu(X,Y)}{\rho(Y)}\right]} \tag{32.9}$$

(In the continuous-Ξ case, the derivative is the ratio of the densities with respect to the common reference measure, and the principle is the same.) Introducing the functional $F(\nu) = \mathbf{E}_\nu \left[\log \frac{\mu(X,Y)}{\rho(Y)}\right]$, the derivative is equal to $O(1)e^{tF(\hat{P}_t^2)}$, and our initial assumption amounts to saying that F is not just continuous (which it must be) but bounded from above.

Now introduce $Q_{t,X}$, the distribution of the empirical pair measure $\hat{P}_t^2$ under the Markov process, and $Q_{t,Y}$, the distribution of $\hat{P}_t^2$ for the IID samples produced by Y_i. From Eq. 32.9,

$$\begin{aligned} \frac{1}{t}\log \mathbb{P}\left(\hat{P}_t^2 \in B\right) &= \frac{1}{t}\log \int_B dQ_{t,X}(\psi) && (32.10)\\ &= \frac{1}{t}\log \int_B \frac{dQ_{t,X}}{dQ_{t,Y}} dQ_{t,Y}(\psi) && (32.11)\\ &= O\left(\frac{1}{t}\right) + \frac{1}{t}\log \int_B e^{tF(\psi)} dQ_{t,Y}(\psi) && (32.12) \end{aligned}$$

It is thus clear that

$$\begin{aligned} \liminf \frac{1}{t}\log \mathbb{P}\left(\hat{P}_t^2 \in B\right) &= \liminf \frac{1}{t}\log \int_B e^{tF(\psi)} dQ_{t,Y}(\psi) && (32.13)\\ \limsup \frac{1}{t}\log \mathbb{P}\left(\hat{P}_t^2 \in B\right) &= \limsup \frac{1}{t}\log \int_B e^{tF(\psi)} dQ_{t,Y}(\psi) && (32.14) \end{aligned}$$

Introduce a (final) proxy random sequence, also taking values in $\mathcal{P}\left(()\,\Xi^2\right)$, call it Z_t, with $\mathbb{P}\left(Z_t \in B\right) = \int_B e^{tF(\psi)} dQ_{t,Y}(\psi)$. We know (Corollary 435) that, under $Q_{t,Y}$, the empirical pair measure satisfies an LDP with rate t and good rate function $J_Y = D\left(\psi \| \pi_1\psi \otimes \rho\right)$, so by Corollary 416, Z_t satisfies an LDP with rate t and good rate function

$$J_F(\psi) = -(F(\psi) - J_Y(\psi)) + \sup_{\zeta \in \mathcal{P}(\Xi^2)} F(\zeta) - J_Y(\zeta) \tag{32.15}$$

A little manipulation turns this into

$$J_F(\psi) = D\left(\psi \| \pi_1\psi \otimes \mu\right) - \inf_{\zeta \in \mathcal{P}(\Xi^2)} D\left(\zeta \| \pi_1\zeta \otimes \mu\right) \tag{32.16}$$

and the infimum is clearly zero. Since this is the rate function Z_t, in view of Eqs. 32.13 and 32.14 it is also the rate function for $\hat{P}_n^2$, which we have agreed to call J_2. □

Remark 1: The key to making this work is the assumption that F is bounded from above. This can fail if, for instance, the process is not ergodic, although usually in that case one can rescue the general idea by some kind of ergodic decomposition.

Remark 2: The LDP for the pair measure of an IID sequence can now be seen to be a special case of the LDP for the pair measure of a Markov sequence. The same is true, generally speaking, of all the other LDPs for IID and Markov sequences. Calculations are almost always easier for the IID case, however, which permits us to give explicit formulae for the rate functions of empirical means and empirical distributions unavailable (generally speaking) in the Markovian case.

Corollary 438 *The minima of the rate function J_2 are the invariant distributions.*

PROOF: The rate function is $D\left(\psi \| \pi_1\psi \otimes \mu\right)$. Since relative entropy is ≥ 0, and equal to zero iff the two distributions are equal (Lemma 360), we get a minimum of zero in the rate function iff $\psi = \pi_1\psi \otimes \mu$, or $\psi = \rho^2$, for some $\rho \in \mathcal{P}\left(\Xi\right)$ such that $\rho\mu = \rho$. Conversely, if ψ is of this form, then $J_2(\psi) = 0$. □

Corollary 439 *The empirical distribution $\hat{P}_t$ obeys an LDP with rate t and good rate function*

$$J_1(\psi) = \inf_{\zeta \in \mathcal{P}(\Xi^2) : \pi_1\zeta = \psi} D\left(\zeta \| \pi_1\zeta \otimes \mu\right) \tag{32.17}$$

PROOF: This is a direct application of the Contraction Principle (Theorem 410), as in Corollary 415. □

Remark: Observe that if ψ is invariant under the action of the Markov chain, then $J_1(\psi) = 0$ by a combination of the preceding corollaries. This is good, because we know from ergodic theory that the empirical distribution converges on the invariant distribution for an ergodic Markov chain. In fact, in view of Lemma 361, which says that $D\left(\psi \| \rho\right) \geq \frac{1}{2\ln 2}\|\psi - \rho\|_1^2$, the probability that the empirical distribution differs from the invariant distribution ρ by more than δ, in total variation distance, goes down like $O(e^{-t\delta^2/2})$.

Corollary 440 *If Theorem 437 holds, then time averages of observables, $A_t f$, obey a large deviations principle with rate function*

$$J_0(a) = \inf_{\zeta \in \mathcal{P}(\Xi^2) :\ \mathbf{E}_{\pi_1\zeta}[f(X)]} D\left(\zeta \| \pi_1\zeta \otimes \mu\right) \tag{32.18}$$

PROOF: Another application the Contraction Principle, as in Corollary 415. □

Remark: Observe that if $a = \mathbf{E}_\rho\left[f(X)\right]$, with ρ invariant, then the $J_0(a) = 0$. Again, it is reassuring to see that large deviations theory is compatible with ergodic theory, which tells us to expect the almost-sure convergence of $A_t f$ on $\mathbf{E}_\rho\left[f(X)\right]$.

Corollary 441 *If X_i are from a Markov sequence of order $k+1$, then, under conditions analogous to Theorem 437, the $k+1$-dimensional empirical distribution $\hat{P}_t^{k+1}$ obeys an LDP with rate t and good rate function*

$$D\left(\nu \| \pi_{k-1}\nu \otimes \mu\right) \tag{32.19}$$

PROOF: An obvious extension of the argument for Theorem 437, using the appropriate exponential-family representation of the higher-order process. □

Whether all exponential-family stochastic processes (Küchler and Sørensen, 1997) obey LDPs is an interesting question; I'm not sure if anyone knows the answer.

32.2 Higher LDPs for Markov Sequences

In this section, I assume without further comment that the Markov sequence X obeys the LDP of Theorem 437.

Theorem 442 *For all $k \geq 2$, the finite-dimensional empirical distribution $\hat{P}_t^k$ obeys an LDP with rate t and good rate function $J_k(\psi) = D\left(\psi \| \pi_{k-1}\psi \otimes \mu\right)$, if $\psi \in \mathcal{P}\left(\Xi^k\right)$ is shift-invariant, and $= \infty$ otherwise.*

PROOF: The case $k = 2$ is just Theorem 437. However, if $k \geq 3$, the argument preceding that theorem shows that $\mathbb{P}\left(\hat{P}_t^k \in B\right)$ depends only on $\pi_2 \hat{P}_t^k$, the pair measure implied by the k-dimensional distribution, so the proof of that theorem can be adapted to apply to $\hat{P}_t^k$, in conjunction with Corollary 435, establishing the LDP for finite-dimensional distributions of IID sequences. The identification of the rate function follows the same argument, too. □

Theorem 443 *The empirical process distribution obeys an LDP with rate t and good rate function $J_\infty(\psi) = d(\psi \| \rho)$, with ρ here standing for the stationary process distribution of the Markov sequence.*

PROOF: Entirely parallel to the proof of Theorem 436, with Theorem 442 substituting for Corollary 435. □

Consequently, any continuous function of the empirical process distribution has an LDP.

Bibliography

Abramowitz, Milton and Irene A. Stegun (eds.) (1964). *Handbook of Mathematical Functions*. Washington, D.C.: National Bureau of Standards. URL `http://www.math.sfu.ca/~cbm/aands/`.

Algoet, Paul (1992). "Universal Schemes for Prediction, Gambling and Portfolio Selection." *The Annals of Probability*, **20**: 901–941. See also an important Correction, *Annals of Probability*, **23** (1995): 474–478.

Algoet, Paul H. and Thomas M. Cover (1988). "A Sandwich Proof of the Shannon-McMillan-Breiman Theorem." *The Annals of Probability*, **16**: 899–909.

Amari, Shun-ichi and Hiroshi Nagaoka (1993/2000). *Methods of Information Geometry*. Providence, Rhode Island: American Mathematical Society. Translated by Daishi Harada. As *Joho Kika no Hoho*, Tokyo: Iwanami Shoten Publishers.

Arnol'd, V. I. (1973). *Ordinary Differential Equations*. Cambridge, Massachusetts: MIT Press. Trans. Richard A. Silverman from *Obyknovennye differentsial'nye Uravneniya*.

— (1978). *Mathematical Methods of Classical Mechanics*. Berlin: Springer-Verlag. First published as *Matematicheskie metody klassicheskoĭ mekhaniki*, Moscow: Nauka, 1974.

Arnol'd, V. I. and A. Avez (1968). *Ergodic Problems of Classical Mechanics*. Mathematical Physics Monograph Series. New York: W. A. Benjamin.

Badino, Massimiliano (2004). "An Application of Information Theory to the Problem of the Scientific Experiment." *Synthese*, **140**: 355–389. URL `http://philsci-archive.pitt.edu/archive/00001830/`.

Banks, J., J. Brooks, G. Cairns, G. Davis and P. Stacy (1992). "On Devaney's Definition of Chaos." *American Mathematical Monthly*, **99**: 332–334.

Bartlett, M. S. (1955). *An Introduction to Stochastic Processes, with Special Reference to Methods and Applications*. Cambridge, England: Cambridge University Press.

Basharin, Gely P., Amy N. Langville and Valeriy A. Naumov (2004). "The Life and Work of A. A. Markov." *Linear Algebra and its Applications*, **386**: 3–26. URL `http://decision.csl.uiuc.edu/~meyn/pages/Markov-Work-and-life.pdf`.

Biggers, Earl Derr (1928). *Behind That Curtain*. New York: Grosset and Dunlap.

Billingsley, Patrick (1961). *Statistical Inference for Markov Processes*. Chicago: University of Chicago Press.

Blackwell, David and M. A. Girshick (1954). *Theory of Games and Statistical Decisions*. New York: Wiley.

Bosq, Denis (1998). *Nonparametric Statistics for Stochastic Processes: Estimation and Prediction*. Berlin: Springer-Verlag, 2nd edn.

Caires, S. and J. A. Ferreira (2005). "On the Non-parametric Prediction of Conditionally Stationary Sequences." *Statistical Inference for Stochastic Processes*, **8**: 151–184.

Courant, Richard and David Hilbert (1953). *Methods of Mathematical Physics*. New York: Wiley.

Cover, Thomas M. and Joy A. Thomas (1991). *Elements of Information Theory*. New York: Wiley.

Cramér, Harald (1945). *Mathematical Methods of Statistics*. Uppsala: Almqvist and Wiksells. Reprint, Princeton, New Jersey: Princeton University Press, 1999.

Devaney, Robert L. (1992). *A First Course in Chaotic Dynamical Systems: Theory and Experiment*. Reading, Mass.: Addison-Wesley.

Doob, Joseph L. (1953). *Stochastic Processes*. Wiley Publications in Statistics. New York: Wiley.

Doukhan, Paul (1995). *Mixing: Properties and Examples*. New York: Springer-Verlag.

Durrett, Richard (1991). *Probability: Theory and Examples*. Belmont, California: Duxbury.

Dynkin, E. B. (1978). "Sufficient statistics and extreme points." *Annals of Probability*, **6**: 705–730.

Eckmann, Jean-Pierre and David Ruelle (1985). "Ergodic Theory of Chaos and Strange Attractors." *Reviews of Modern Physics*, **57**: 617–656.

Ellis, Richard S. (1985). *Entropy, Large Deviations, and Statistical Mechanics*, vol. 271 of *Grundlehren der mathematischen Wissenschaften*. Berlin: Springer-Verlag.

— (1988). "Large Deviations for the Empirical Measure of a Markov Chain with an Application to the Multivariate Empirical Measure." *The Annals of Probability*, **16**: 1496–1508.

Ethier, Stewart N. and Thomas G. Kurtz (1986). *Markov Processes: Characterization and Convergence*. New York: Wiley.

Eyink, Gregory L. (1996). "Action principle in nonequilbrium statistical dynamics." *Physical Review E*, **54**: 3419–3435.

Forster, Dieter (1975). *Hydrodynamic Fluctuations, Broken Symmetry, and Correlation Functions*. Reading, Massachusetts: Benjamin Cummings.

Freidlin, M. I. and A. D. Wentzell (1998). *Random Perturbations of Dynamical Systems*. Berlin: Springer-Verlag, 2nd edn. First edition first published as *Fluktuatsii v dinamicheskikh sistemakh pod deistviem malykh sluchainykh vozmushchenii*, Moscow: Nauka, 1979.

Frisch, Uriel (1995). *Turbulence: The Legacy of A. N. Kolmogorov*. Cambridge, England: Cambridge University Press.

Fristedt, Bert and Lawrence Gray (1997). *A Modern Approach to Probability Theory*. Probability Theory and Its Applications. Boston: Birkhäuser.

Gikhman, I. I. and A. V. Skorokhod (1965/1969). *Introduction to the Theory of Random Processes*. Philadelphia: W. B. Saunders. Trans. Richard Silverman from *Vvedenie v teoriiu slucainikh protessov*, Moscow: Nauka; reprinted Mineola, New York: Dover, 1996.

Gillespie, John H. (1998). *Population Genetics: A Concise Guide*. Baltimore: Johns Hopkins University Press.

Gnedenko, B. V. and A. N. Kolmogorov (1954). *Limit Distributions for Sums of Independent Random Variables*. Cambridge, Massachusetts: Addison-Wesley. Translated from the Russian and annotated by K. L. Chung, with an Appendix by J. L. Doob.

Gray, Robert M. (1988). *Probability, Random Processes, and Ergodic Properties*. New York: Springer-Verlag. URL `http://ee-www.stanford.edu/~gray/arp.html`.

— (1990). *Entropy and Information Theory*. New York: Springer-Verlag. URL `http://www-ee.stanford.edu/~gray/it.html`.

Howard, Ronald A. (1971). *Markov Models*, vol. 1 of *Dynamic Probabilistic Systems*. New York: Wiley.

Ibragimov, I. A. and R. Z. Has'minskii (1979/1981). *Statistical Estimation: Asymptotic Theory*. New York: Springer-Verlag. Tranlated by Samuel Kotz from *Asimptoticheskaia teoriia ostsenivaniia*, Moscow: Nauka.

Jakubowski, A. and Z. S. Szewczak (1990). "A normal convergence criterion for strongly mixing stationary sequences." In *Limit Theorems in Probability and Statistics* (I. Berkes and Endre Csáki and Pal Révesz, eds.), vol. 57 of *Colloquia mathematica Societatis Janos Bolyai*, pp. 281–292. New York: North-Holland. Proceedings of the Third Hungarian Colloquium on Limit Theorems in Probability and Statistics, held in Pécs, Hungary, July 3-7, 1989 and sponsored by the Bolyai János Mathematical Society.

Kac, Mark (1947). "On the Notion of Recurrence in Discrete Stochastic Processes." *Bulletin of the American Mathematical Society*, **53**: 1002–1010. Reprinted in Kac (1979), pp. 231–239.

— (1979). *Probability, Number Theory, and Statistical Physics: Selected Papers*. Cambridge, Massachusetts: MIT Press. Edited by K. Baclawski and M. D. Donsker.

Kass, Robert E. and Paul W. Vos (1997). *Geometrical Foundations of Asymptotic Inference*. Wiley Series in Probability and Statistics. New York: Wiley.

Katznelson, I. and B. Weiss (1982). "A simple proof opf some ergodic theorems." *Israel Journal of Mathematics*, **42**: 291–296.

Keizer, Joel (1987). *Statistical Thermodynamics of Nonequilibrium Processes*. New York: Springer-Verlag.

Khinchin, Aleksandr Iakovlevich (1949). *Mathematical Foundations of Statistical Mechanics*. New York: Dover Publications. Translated from the Russian by G. Gamow.

Knight, Frank B. (1975). "A Predictive View of Continuous Time Processes." *Annals of Probability*, **3**: 573–596.

— (1992). *Foundations of the Prediction Process*. Oxford: Clarendon Press.

Kontoyiannis, I., P. H. Algoet, Yu. M. Suhov and A. J. Wyner (1998). "Nonparametric entropy estimation for stationary processes and random fields, with applications to English text." *IEEE Transactions on Information Theory*, **44**: 1319–1327. URL `http://www.dam.brown.edu/people/yiannis/PAPERS/suhov2.pdf`.

Küchler, Uwe and Michael Sørensen (1997). *Exponential Families of Stochastic Processes*. Berlin: Springer-Verlag.

Kulhavý, Rudolf (1996). *Recursive Nonlinear Estimation: A Geometric Approach*, vol. 216 of *Lecture Notes in Control and Information Sciences*. Berlin: Springer-Verlag.

Kullback, Solomon (1968). *Information Theory and Statistics*. New York: Dover Books, 2nd edn.

Kurtz, Thomas G. (1970). "Solutions of Ordinary Differential Equations as Limits of Pure Jump Markov Processes." *Journal of Applied Probability*, **7**: 49–58.

— (1971). "Limit Theorems for Sequences of Jump Markov Processes Approximating Ordinary Differential Processes." *Journal of Applied Probability*, **8**: 344–356.

Kushner, Harold J. (1984). *Approximation and Weak Convergence Methods for Random Processes, with Applications to Stochastic Systems Theory*. Cambridge, Massachusetts: MIT Press.

Lasota, Andrzej and Michael C. Mackey (1994). *Chaos, Fractals, and Noise: Stochastic Aspects of Dynamics*. Berlin: Springer-Verlag. First edition, *Probabilistic Properties of Deterministic Systems*, Cambridge University Press, 1985.

Lehmann, E. L. and George Casella (1998). *Theory of Point Estimation*. Springer Texts in Statistics. Berlin: Springer-Verlag, 2nd edn.

Loève, Michel (1955). *Probability Theory*. New York: D. Van Nostrand Company, 1st edn.

Mackey, Michael C. (1992). *Time's Arrow: The Origins of Thermodynamic Behavior*. Berlin: Springer-Verlag.

Øksendal, Bernt (1995). *Stochastic Differential Equations: An Introduction with Applications*. Berlin: Springer-Verlag, 4th edn.

Ornstein, D. S. and B. Weiss (1990). "How Sampling Reveals a Process." *Annals of Probability*, **18**: 905–930.

Pearl, Judea (1988). *Probabilistic Reasoning in Intelligent Systems*. New York: Morgan Kaufmann.

Pollard, David (2002). *A User's Guide to Measure Theoretic Probability*. Cambridge, England: Cambridge University Press.

Rockafellar, R. T. (1970). *Convex Analysis*. Princeton: Princeton University Press.

Rogers, L. C. G. and David Williams (1994). *Foundations*, vol. 1 of *Diffusions, Markov Processes and Martingales*. New York: John Wiley, 2nd edn. Reprinted Cambridge, England: Cambridge University Press, 2000.

— (2000). *Itô Calculus*, vol. 2 of *Diffusions, Markov Processes and Martingales*. Cambridge, England: Cambridge University Press, 2nd edn.

Rosenblatt, Murray (1956). "A Central Limit Theorem and a Strong Mixing Condition." *Proceedings of the National Academy of Sciences (USA)*, **42**: 43–47. URL `http://www.pnas.org/cgi/reprint/42/1/43`.

— (1971). *Markov Processes. Structure and Asymptotic Behavior*. Berlin: Springer-Verlag.

Schroeder, Manfred (1991). *Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise*. San Francisco: W. H. Freeman.

Shannon, Claude E. (1948). "A Mathematical Theory of Communication." *Bell System Technical Journal*, **27**: 379–423. URL `http://cm.bell-labs.com/cm/ms/what/shannonday/paper.html`. Reprinted in Shannon and Weaver (1963). `http://cm.bell-labs.com/cm/ms/what/shannonday/paper.html`.

Shannon, Claude E. and Warren Weaver (1963). *The Mathematical Theory of Communication*. Urbana, Illinois: University of Illinois Press.

Shiryaev, Albert N. (1999). *Essentials of Stochastic Finance: Facts, Models, Theory*. Singapore: World Scientific. Trans. N. Kruzhilin.

Spirtes, Peter, Clark Glymour and Richard Scheines (2001). *Causation, Prediction, and Search*. Cambridge, Massachusetts: MIT Press, 2nd edn.

Taniguchi, Masanobu and Yoshihide Kakizawa (2000). *Asymptotic Theory of Statistical Inference for Time Series*. Berlin: Springer-Verlag.

Tyran-Kamińska, Marta (2005). "An Invariance Principle for Maps with Polynomial Decay of Correlations." *Communications in Mathematical Physics*, **260**: 1–15. URL `http://arxiv.org/abs/math.DS/0408185`.

von Plato, Jan (1994). *Creating Modern Probability: Its Mathematics, Physics and Philosophy in Historical Perspective*. Cambridge, England: Cambridge University Press.

Wiener, Norbert (1949). *Extrapolation, Interpolation, and Smoothing of Stationary Time Series: With Engineering Applications*. Cambridge, Massachusetts: The Technology Press of the Massachusetts Institute of Technology.

— (1958). *Nonlinear Problems in Random Theory*. Cambridge, Massachusetts: The Technology Press of the Massachusetts Institute of Technology.

— (1961). *Cybernetics: Or, Control and Communication in the Animal and the Machine*. Cambridge, Massachusetts: MIT Press, 2nd edn. First edition New York: Wiley, 1948.

Wu, Wei Biao (2005). "Nonlinear system theory: Another look at dependence." *Proceedings of the National Academy of Sciences (USA)*, **102**: 14150–14154.