Random Processes and Ergodicity

Mathias Winther Madsen

November 13, 2015

1 Random Processes

1.1 Motivating Examples

Every model we have seen in the course up to this point has used finite sets of random variables and finite-dimensional sample spaces. However, many real-world processes—texts, diseases, stock prices—develop over time and involve randomness at infinitely many points in time.

Consider for instance the follwing processes:

1. **Random Walk** Start at $X_1 = 0$, and walk left or right with equal probability:

$$0, -1, 0, -1, -2, -3, -4, -3, -4, -3, -2, -1, 0, 1, 2, 1, 2, 1, 2, 1, 2, \ldots$$

2. **Bi-Deterministic Process** Flip a fair coin to decide whether to produce the constant sequence

$$0, 0, 0, 0, 0, 0, \ldots \qquad \text{or} \qquad 1, 1, 1, 1, 1, 1, \ldots$$

3. **Gappy Process** Repeatedly flip a fair coin to choose X_i, but always set $X_i = 0$ deterministically if $X_{i-1} = 1$:

$$0, 0, 1, 0, 1, 0, 0, 1, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, \ldots$$

4. **Letter Repetitions** Repeatedly choose a letter from the English alphabet and print it $k \sim \text{Geometric}(1/2)$ times:

$$\texttt{SSSPMMMMMDHHHKZTDUCAAAIDTTTYHHHHQTTXX}\ldots$$

5. **Beta Urn Model** Put a red ($X = 0$) and a blue ($X = 1$) marble in a bag; repeatedly draw from the bag, adding another marble of the same color after each drawing:

$$0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, \ldots$$

Such processes seem to have a clear mathematical structure, but this structure cannot be described in terms of a finite set of random variables. In order to model them better, we need **random processes**.

1.2 Definition

A random process is an infinite collection of random variables $X_1, X_2, X_3, \ldots$. For the purposes of this course, we will only consider countably infinite collections, and only discrete variables.

As you will remember, a random variable is a function from an underlying sample space Ω to a value space $\mathcal{X}$. At each point $\omega \in \Omega$, the random variable has a value $X(\omega) \in \mathcal{X}$, and for each value $x \in \mathcal{X}$, there is an information cell $X^{-1}(x) \subseteq \Omega$ consistent with the observation $X = x$.

If instead we have a finite family of random variables $X_1, X_2, \ldots, X_n$ with values in $\mathcal{X}$, then every $\omega \in \Omega$ assigns a value vector

$$(X_1(\omega), X_2(\omega), \ldots, X_n(\omega)) \in \mathcal{X}^n$$

to the random vector $(X_1, X_2, \ldots, X_n)$. Reversely, any value vector (or set of value vectors) corresponds to a subset of Ω with a specific probability.

A random process is the infinite generalization of this idea. To define a random process, we need to select a sample space Ω, equip it with a probability distribution P, and then explain how each choice of $\omega \in \Omega$ determines the infinite series

$$X_1(\omega), X_2(\omega), X_3(\omega), \ldots$$

Such a countably infinite series of values is generally called a **sample path**.

Definition 1. A **discrete random process** is a family of random variables $\{X_i\}_{i \in I}$ indexed by a discrete set I, and with values in a discrete space $\mathcal{X}$.

In a nutshell, a random process is thus a probability distribution over a set of sample paths.

1.3 The Extension Theorem

It can be tricky to specify how to spread out the probability budget across the universe of sample paths when we use this definition directly. Typically, we are working with overcountably large bundles of infinitely long sample paths, and it is not always obvious how to define a distribution over such sets.

Fortunately, the following regularity theorem provides some help.

Theorem 1. (The Daniell-Kolmogorov Extention Theorem) Suppose two random processes P_1 and P_2 assign the same marginal probabilities to all finite-dimensional events. Then they assign the same probabilities to all countably-infinite-dimensional events.[1]

[1] P. J. Daniell proved that infinite-dimensional integrals are uniquely determined from finite-dimensional ones when the integral preserves limits (see "Integrals in An Infinite Number of Dimensions," *Annals of Mathematics*, Vol. 20(4), 1919). A. Kolmogorov independently arrived at a similar conclusion about probability measures (see Chapters 2.2 and 3.4 of *Grundbegriffe der Wahrscheinlichkeitsrechnung*, Springer, 1933).

In other words: once you know the probability of any formula defined in terms of conjunctions, disjunctions, and negations, you also know the probability of any formula defined in terms of conjunctions, disjunctions, negations, and countable quantifications. The proof (which I will not reproduce here) relies crucially on the fact that a probability distribution, by definition, is required to be countably additive. This ensures nice limit behavior.

1.4 Initial-Segment Representation

Note that if you can compute all probabilities of the form

$$P\left(X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n\right),$$

then you can compute the probability of any finite-dimensional event (by setting $A_i = \mathcal{X}$ for the dimensions you are not interested in). In fact, since we are dealing with discrete random processes, it is also enough to just provide the point sequence probabilities of the form

$$P\left(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\right).$$

When the process is discrete, the probability of any finite-dimensional event can be computed by summing up point probabilities of this type. A discrete random process P is therefore uniquely determined once we know the point probability of every initial sequence of values.

An alternative way of describing these initial-segment probabilities is to provide the continuation probabilities

$$P\left(X_n = x_n \,|\, X_1 = x_1, X_2 = x_2, \ldots, X_{n-1} = x_{n-1}\right)$$

along with an unconditional distribution P_{X_1} which provides an initial condition for the first coordinate. Whichever of these two strategies we use, however, we need to specify these probabilities for all n, and for all value vectors $x_1, x_2, \ldots, x_n$.

1.5 Examples

By using the initial-segment formulation of the extension theorem, we can now formally describe the probability distributions P that model the processes mentioned above:

1. **Random Walk** For every $(x_1, x_2, \ldots, x_n) \in \mathbb{Z}^n$:

$$P\left(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\right) = \begin{cases} 2^{1-n} & \text{if } x_1 = 0 \text{and} \\ & |x_i - x_{i+1}| = 1 \text{ for } i < n \\ 0 & \text{otherwise} \end{cases}$$

2. **Bi-Deterministic Process** For every $(x_1, x_2, \ldots, x_n) \in \{0,1\}^n$:

$$P\left(X_1 = x_1, X_2 = x_1, \ldots, X_n = x_1\right) \;=\; \begin{cases} 1/2 & \text{if } x_1 = x_2 = \cdots = x_n = 0 \\ 1/2 & \text{if } x_1 = x_2 = \cdots = x_n = 1 \\ 0 & \text{otherwise} \end{cases}$$

3. **Gappy Process** For every $(x_1, x_2, \ldots, x_n) \in \{0,1\}^n$:

$$P\left(X_n = x_n \,|\, X_1 = x_1, \ldots, X_{n-1} = x_{n-1}\right) \;=\; \begin{cases} 1/2 & \text{if } x_{n-1} = 0 \\ 0 & \text{if } x_n = x_{n-1} = 1 \\ 1 & \text{if } x_{n-1} = 1 \text{and } x_n = 0 \end{cases}$$

with the initial condition $P(X_1 = 0) = P(X_1 = 1) = 1/2$.

4. **Letter Repetitions** For every $(x_1, x_2, \ldots, x_n) \in \{\mathtt{A}, \mathtt{B}, \mathtt{C}, \ldots, \mathtt{Z}\}^n$:

$$P\left(X_n = x_n \,|\, X_1 = x_1, \ldots, X_{n-1} = x_{n-1}\right) \;=\; \begin{cases} 1/2 + 1/2 \cdot 1/27 & \text{if } x_n = x_{n-1} \\ 1/2 \cdot 1/27 & \text{if } x_n \neq x_{n-1} \end{cases}$$

with the initial condition $P(X_1 = x) = 1/27$ for all $x \in \{\mathtt{A}, \mathtt{B}, \mathtt{C}, \ldots, \mathtt{Z}\}$.

5. **Beta Urn Model** For every $(x_1, x_2, \ldots, x_n) \in \{0,1\}^n$:

$$\begin{aligned} P\left(X_n = 1 \,|\, X_1 = x_1, \ldots, X_{n-1} = x_{n-1}\right) &= \frac{1 + x_1 + \cdots + x_n}{2 + n} \\ P\left(X_n = 0 \,|\, X_1 = x_1, \ldots, X_{n-1} = x_{n-1}\right) &= \frac{1 + n - x_1 - \cdots - x_n}{2 + n} \end{aligned}$$

with the initial condition $P(X_1 = x) = 1/2$ for $x = 0, 1$.

By the extension theorem, each of these families of probability distributions genelize to a unique distribution P on the set of sample paths. Note that the first two processes are defined by means of unconditional initial-segment probabilities, while the last three are defined in terms of conditional continuation probabilities.

2 Markov Chains

2.1 Transition Probabilities

One particularly important kind of random process is the **Markov chain**. A Markov chain

$$X_1 \;\to\; X_2 \;\to\; X_3 \;\to\; X_4 \;\to\; \cdots$$

is a random process whose continuation probabilities at time $n+1$ only depend on the value of X_n. That is, the random variable X_{n+1} is conditionally independent of all the variables $X_1, X_2, \ldots, X_{n-1}$ given X_n. A Markov process is "forgetful" in the sense that only knows what it did one moment ago, but not anything further back in the past.

Definition 2. A random process P is a **Markov chain** if

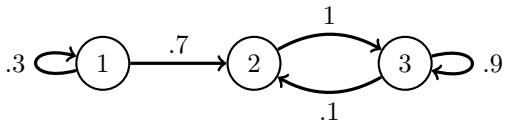
$$P\left(X_{n+1} = x_{n+1} \,|\, X_1 = x_1, \ldots, X_n = x_n\right) \;=\; P\left(X_{n+1} = x_{n+1} \,|\, X_n = x_n\right)$$

for all n and all $x_1, x_2, \ldots, x_n$.[2]

The conditional probabilities on the right-hand side of this equality are usually called the **transition probabilities** of the Markov chain.

Note that several different Markov chains can have the same transition probabilities. If we want these transition probabilities to define a unique random process, we also need to provide an initial condition P_{X_1}. Once we select this initial condition, the choice will propagate out the timeline to all later coordinates, settling their marginal distributions.

The transition probabilities of a Markov chain can be described in several different ways. One option is to draw a **transition diagram**:



Altneratively, we can also provide a **transition matrix**:

$$T \;=\; \begin{pmatrix} .3 & 0 & 0 \\ .7 & 0 & .1 \\ 0 & 1 & .9 \end{pmatrix}.$$

The columns of this matrix represents the conditional distribution functions $P_{X_{n+1}|X_n}$, one for each possible condition.

An initial condition P_{X_1} can be represented be a vector v_1. The marginal distributions of $X_2, X_3, \ldots$ can then be computed by matrix multiplication:

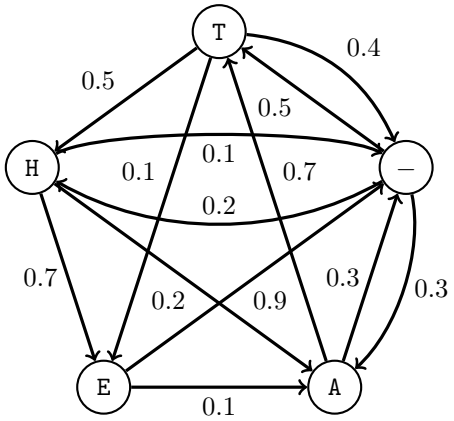
$$\begin{aligned} v_2 &= Tv_1 \\ v_3 &= Tv_2 \\ v_4 &= Tv_3 \\ &\vdots \end{aligned}$$

There is no universally best way of representing a set of transition probabilities. Transition diagrams are compact and easy to inspect visually, while transition matrices easier to use in computational implementations.

[2]This terminology is mainly due to the paper "Extension of the Limit Theorems of Probability Theory to a Sum of Variables Connected in a Chain" (*Notes of the Imperial Academy of Sciences of St. Petersburg*, Vol. 22(9), 1907), in which A. A. Markov proved a convergence result for Markov chains. An English translation is included in R. A. Howard: *Dynamic Probabilistic Systems, Volume I: Markov Models* (Wiley, 1971).

2.2 Markov Models

As Markov himself noticed, Markov chains can be used to model language, and they remain an extremely useful tool in the prediction and description of natural language text. As an illustration of how this might work, consider the following transition diagram:



With the initial condition $P_{X_1}(\texttt{T}) = 1$, these transitions define a unique random process. A representative sample from this process is

```
T_ATE_T_HE_TE_THE_THE_THAT_T_TE_ATHE_AT_ATHE_T_ATHE_TE...
```

This is of course a toy model which was restricted to these five characters in order to make it easier to represent. We can of course also use the whole English alphabet, estimating the transition probabilities from a sample of English text. In that case, we get samples more along the lines of

```
THILY_IMANE_ULDEXTHOUNEDES_E_F_AT_BANIERREDAN'S_SCOPLUPT...
```

This is not English, of course, but for some purposes, they get sufficiently close. Models like these are for instance very good at recognizing the language of a text, mapping sounds to reasonable English, or predicting what you letter you will type next on your phone.

As Shannon noted in his 1948 paper,[3] we can sample from a Markov approximation of English using nothing but pen, paper, and a book of English prose: to pick the next letter (x_{n+1}) given preceding one (x_n), open the book at a random page, look for an occurrence of x_n, and then select the letter following it. This way, the transition probabilities in your sample will be equal to the transition probabilities in the book.

[3]See Part I, §3, of C. E. Shannon: "A Mathematical Theory of Communication," *Bell System Technical Journal*, Vol. 27(3), 1948.

2.3 Stationary Distributions

Consider again the gappy process described above. This process is a Markov chain. Its transition matrix is

$$T = \begin{pmatrix} 1/2 & 1 \\ 1/2 & 0 \end{pmatrix}.$$

Let's further describe the marginal distributions of this process in terms of the vectors

$$v_n = \begin{pmatrix} p_n \\ q_n \end{pmatrix},$$

where $q_n = 1 - p_n$. Since the process is a Markov chain with transition matrix T, we know that $v_{n+1} = Tv_n$ for all n. Plugging in the numbers, we get

$$\begin{pmatrix} p_{n+1} \\ q_{n+1} \end{pmatrix} = \begin{pmatrix} 1/2 & 1 \\ 1/2 & 0 \end{pmatrix} \begin{pmatrix} p_n \\ q_n \end{pmatrix} = \begin{pmatrix} p_n/2 + q_n \\ p_n/2 \end{pmatrix}.$$

We can think about these equations in physical terms, as if they described the flow of water between two communicating vessels. It then makes sense to ask when this system is in equilibrium, that is, for which v^* we have $v^* = Tv^*$.

We can find this equilibrium by solving the equations

$$\begin{pmatrix} p^* \\ q^* \end{pmatrix} = \begin{pmatrix} p^*/2 + q^* \\ p^*/2 \end{pmatrix},$$

which, given the constraint $q^* = 1 - p^*$, has the unique solution

$$\begin{pmatrix} p^* \\ q^* \end{pmatrix} = \begin{pmatrix} 2/3 \\ 1/3 \end{pmatrix}.$$

If this Markov process follows the marginal distribution given by v^* at any point in time, it thus follows it at all times.

This marginal distribution is the unique **stationary distribution** of the Markov chain. Among all the random processes respecting the given transition probabilities, the process described by T and v^* is the only one which has the same marginal distribution at all points in time.

2.4 Attractive Equilibria

The stationary distribution v^* is also an attractor for the transition operation: if we perturb v^* by a small amount so as to produce the distribution

$$\tilde{v}^* = \begin{pmatrix} 2/3 + \varepsilon \\ 1/3 - \varepsilon \end{pmatrix},$$

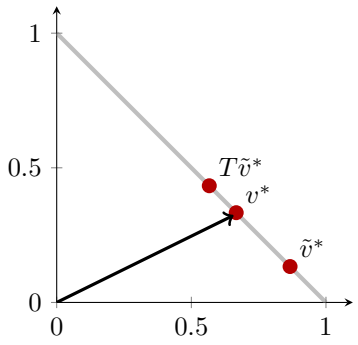
then the transition matrix will map $\tilde{v}^*$ onto

$$T\tilde{v}^* = \begin{pmatrix} (1/3 + \varepsilon/2) + (1/3 - \varepsilon) \\ (1/3 + \varepsilon/2) \end{pmatrix} = \begin{pmatrix} 2/3 - \varepsilon/2 \\ 1/3 + \varepsilon/2 \end{pmatrix}.$$

One step into the future, the distance to the stationary distribution has thus been cut exactly in half, $|T\tilde{v}^* - v^*| = \frac{1}{2}|\tilde{v}^* - v^*|$. The Markov chain thus converges to the stationary distribution in the sense that

$$T^n v \to v^* \qquad (n \to \infty)$$

whatever initial condition v we start with. In fact, since each time step cuts down the remaining distance by a fixed amount, this convergence is exponentially fast.



Every random process defined by the transition matrix T thus has the same limiting behavior when we look far enough out into the future. For the transition matrix in this example, the actual initial condition is therefore of relatively limited interest, since all the random processes defined by T have similar behaviors, and all of them have the exact same limiting behavior. For this family of processes, we can thus make predictions about the far future without having any knowledge of the present whatsoever.

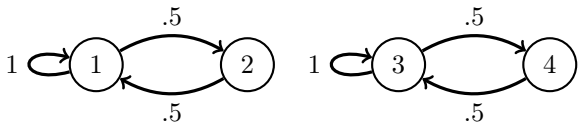
2.5 Uniqueness

In the example above, we saw that a family of random processes had a unique stationary distribution, and that this stationary distribution told us something important about the limit behavior of the process.

This immediately raises two questions:

1. Do all random processes have stationary distributions?
2. When they do, are these stationary distributions unique?

We will address the existence question below. The uniqueness question, however, is settled in the negative by the following transition diagram:



This diagram defines a family of Markov chains whose sample paths move around on two isolated "islands," $\{1, 2\}$ and $\{3, 4\}$. Each of these islands defines a conditional distribution over the sample paths restricted to that island.

The random processes defined by these two conditional distributions happen to be identical, and identical to the gappy example in the previous section. We thus know that

$$\begin{aligned} P(X_n = 1 \,|\, X_1 \in \{1,2\}) &\rightarrow 2/3 \\ P(X_n = 3 \,|\, X_1 \in \{3,4\}) &\rightarrow 2/3 \end{aligned}$$

whatever the marginal distribution of X_1 is (provided that this distribution assigns positive probability to the island we condition on).

We thus have at least two stationary distributions for the transition diagram:

	1	2	3	4
P_1^*	$2/3$	$1/3$	0	0
P_2^*	0	0	$2/3$	$1/3$

These two distributions are attractors for two disjoint classes of initial conditions: any initial condition that places all its probability mass on the first island will eventually converge to P_1^*, while any initial condition restricted to the second island will converge to P_2^*.

Perhaps less obiovusly, when an initial condition endows both islands with a non-trivial amount of probability mass, the two islands will always retain the same relative masses (since the probability mass can never cross from one island to the other). Internally on each island, however, the given endowment will eventually be distributed in the 2:1 proportion suggested by the two stationary distributions.

In addition to the two extremes P_1^* and P_2^*, any convex combination

$$P^* \; = \; \lambda P_1^* + (1-\lambda) P_2^*,$$

with $\lambda \in [0,1]$, is therefore also a stationary distribution for this process. The combination with mixture proportion λ is an attractor for the Markov chains defined by the initial condition

$$\begin{aligned} P_{X_1}(1) + P_{X_1}(2) &= \lambda \\ P_{X_1}(3) + P_{X_1}(4) &= 1-\lambda \end{aligned}$$

This process thus has an overcountably large number of stationary distributions, but they are arranged in a relatively simply geometric pattern.

3 Stationary Random Processes

3.1 Definition

As we saw in the previous section, a Markov chain is uniquely determined by its transition matrix and the initial condition. On their own, the transition probabilities determine a family of Markov chains. The members of this family which are stationary can be identified with a set of initial conditions.

This encoding of the stationary distributions is specific to Markov chains; it does not carry over to more general random processes. In order to define are more general concept, we need to remember that a random process is a probability distribution over sample paths.

What we want to say is then that a distribution P^* is stationary if it cannot to distinguish between the random variables

$$X_1, X_2, X_3, \ldots$$

and its time-shifted twin

$$X_2, X_3, X_4, \ldots$$

This intuition can be expressed in terms of the following definition:

Definition 3. A random process P^* is stationary if

$$P^*\left(X_1 = x_1, \ldots, X_n = x_n\right) \;=\; P^*\left(X_2 = x_1, \ldots, X_{n+1} = x_n\right)$$

for all $n \in \mathbb{N}$ and all value vectors $(x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$.

By induction, this definition implies that a fixed "word" $(x_k, x_{k+1}, \ldots, x_{k+n})$ has the same probability of appearing anywhere in the sample path. The random process P^* therefore has no concept of "early" and "late," only of relative distances in time.

For this reason, it also often makes sense to think of P^* as a distribution over doubly infinite sample paths

$$\ldots, x_2, x_{-1}, x_0, x_1, x_2, x_3, \ldots$$

rather than the singly infinity sample paths $x_1, x_2, x_3, \ldots$ that we have seen up until now. This technical modification is a useful way of encoding the fact that infinitely much time has already passed before we start observing P^*, thus avoiding a number of complications imposed by "burn-in" times and other issues specifically related to the beginning of the process.

If we want to recover a distribution over singly infinite sample paths from such a doubly infinite process, we can always impose an initial condition like $X_1 = 0$ so as to make sure that the half-process on the right of the origin looks as we want it to. Note, however, that this conditional process often isn't stationary even though P^* is.

3.2 Examples

The following examples illustrate the concept of stationary random processes:

1. A stationary Markov chain is a stationary random process. This holds because the probability of a value vector $(x_1, x_2, \ldots, x_{1+n})$ only depends on two things, the marginal distribution of X_1 (which by stationary is equal to $X_2 = X_3 = X_4 = \cdots$), and the conditional probabilities of X_{k+1} given X_k (which by the Markov assumption are the same for all $k \in \mathbb{N}$).

2. The transition probabilities on $\mathbb{Z}$ under which we walk left or right with equal probability define a family of Markov chains with countably infinitely many states. None of these processes are stationary, as can be seen from the fact that

$$Var(X_k) \;<\; Var(X_{k+1}) \;<\; Var(X_{k+2}) \;<\; \cdots$$

for any $k \in \mathbb{N}$ and any marginal distribution over X_k. Consequently, there is no probability distribution on the set of sample paths that satisfies these transition probabilities. This reflects that the fact that such a distribution can always become more diffuse over time.

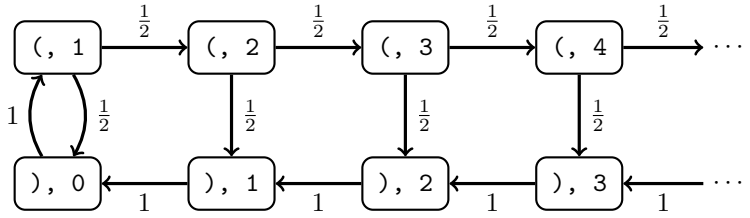
3. Consider the process P that repeatedly selects $k \sim \text{Geometric}(1/2)$ and then prints k left-parentheses and k right-parentheses. A typical sample path drawn from this process might begin

$$()((())))((()))(())(())((()))(()) \ldots$$

This process is not a Markov chain. No finite amount of "memory" is sufficient to simulate this process, since you may need to look indefinitely far back to decide whether there are still open parentheses left over.

The process P does, however, have a closely related cousin P^* which is a stationary: this is a stationary distribution over the doubly infinite sequences of parentheses which is identical to P except that it doesn't have a clear starting point.

The easiest way of constructing this stationary process P^* is to add a hidden memory variable that keeps track of how many opening parentheses that are still left hanging. We can then model P^* as a Markov chain with the transition diagram



This Markov chain has the stationary marginal distribution P^*_X given by

$$\begin{aligned} P^*_X(\text{“(”}, m) &= 2^{-m-1}, \qquad m \geq 1 \\ P^*_X(\text{“)”}, m) &= 2^{-m-2}, \qquad m \geq 0 \end{aligned}$$

By "summing out" the hidden memory variable, this Markov chain defines a distribution over finite strings and therefore over infinite sequences. Perhaps not surprisingly, this distribution has the marginal probabilities

$$P^*(X_k = \text{“(”}) \;=\; P^*(X_k = \text{“)”}) \;=\; 1/2$$

at any single coordinate k. Moreover, if we impose the initial condition $X_0 = (\text{“})\text{”}, 0)$ on P^*, we recover the singly infinite process P.

Note that since we are working with discrete random processes, the strategy employed in the last example—adding hidden state variables as need—works quite generally. Even a random process defined in terms of a completely unconstrained, stochastic Turing machine can be defined as a Markov chain on a countable state of memory states (i.e., tape contents) and a set of transition probabilities.

3.3 Time-Invariant Properties

Having now introduced and rehearsed the concept of general stationary distributions, we will state and prove a theorem about the uniqueness of long-term averages. For this, we first need a new concept:

Definition 4. A bundle of sample paths $B \subseteq \mathcal{X}^{\mathbb{N}}$ is **time-invariant** if

$$(x_1, x_2, x_3, \ldots) \in B \quad \Longrightarrow \quad (x_2, x_3, x_4, \ldots) \in B.$$

If we think of the bundle B as a predicate, then the time-invariant predicates are those that are true or false of the sample path as a whole, regardless of which coordinate we count as “time 0.” Examples of time-invariant properties are:

1. The sample path never visits a certain region $A \subseteq \mathcal{X}$.
2. The sample path never leaves the region $A \subseteq \mathcal{X}$.
3. The sample path visits $A \subseteq \mathcal{X}$ infinitely often.
4. The sample path visits $A \subseteq \mathcal{X}$ with a certain limiting frequency.
5. The sample path is a constant sequence $x, x, x, x, \ldots$.
6. The sample path never transitions directly to $x_2 \in \mathcal{X}$ from $x_1 \in \mathcal{X}$.
7. The sample path converges. (Applicable when $\mathcal{X}$ is a metric space.)

As these examples suggest, it can sometimes be useful to think about time-invariant as “traps” or “black holes” that you cannot get out of however far into the future you go. Note also that time-invariance is preserved under conjunction, disjunction, and negation.

In the context of a random process P, we call a bundle of sample paths **trivial** if it has P-probability 0 or 1. As we shall see below, it is of crucial interest to investigate whether a distribution P has any time-invariant bundles with non-trivial probability. If it does, then different sample paths may suggest radically different pictures of the process P.

3.4 Uniqueness of Limiting Time-Averages

Suppose we define some function f which maps any sample path $x = x_1, x_2, x_3, \ldots$ onto a real number. f might for instance map the sample path x to 1 or 0 depending on whether its first coordinate is an element of some set A; or it might map the whole sample path to its limit, if this makes sense in the context.

Given such a function and a sample path $x = x_1, x_2, x_3, \ldots$, we can then define the nth time-average of f along x as

$$A_n f(x) \;=\; \frac{f(x_1, x_2, \ldots) + f(x_2, x_3, \ldots) + \cdots + f(x_n, x_{n+1}, \ldots)}{n}.$$

$A_n f(x)$ is thus an average n numbers, namely the outputs we get by feeding the first n time-shifted versions of the sample path x into f. Both f and $A_n f$ are thus functions of the sample path x, and therefore random variables.

We then have the following theorem:

Theorem 2. (Uniqueness of Averages) Let P be a random process and $f : \mathcal{X}^{\mathbb{N}} \to \mathbb{R}$ an integrable function whose time-average almost always converge,

$$P\left(x : \lim_{n\to\infty} A_n f(x) \text{ exists}\right) \;=\; 1.$$

Suppose further that P only has trivial trapping sets. Then there is a unique limiting average τ^* such that

$$P\left(x : \lim_{n\to\infty} A_n f(x) = \tau^*\right) \;=\; 1.$$

In other words, the limit of $A_n f$ (which is a random variable) almost always has the same value.

Proof. Statements about limits are time-invariant. By assumption, all time-invariant sets are trivial under P. Hence, any statement about the limit of $A_n f$ either has probability 0 or probability 1. Consequently, the limit statement

$$B_\tau \;=\; \left\{x : \lim_{n\to\infty} A_n f(x) \leq \tau\right\}$$

has probability 1 above some threhold τ^*, and probability 0 for any smaller $\tau < \tau^*$. By the triviality assumption, $P(B_\tau)$ must jump in this way. However, $P(B_\tau)$ is the cumulative distribution of the random variable $\lim_{n\to\infty} A_n f(x)$. Since this cumulative distribution is a step function, $\lim_{n\to\infty} A_n f(x)$ is deterministic. □

As an application of this theorem, let f be the indicator function

$$f(x) \;=\; \mathbb{I}_A(x_1) \;=\; \begin{cases} 1 & x_1 \in A \\ 0 & x_1 \notin A \end{cases}$$

The time-average $A_n f$ is then the relative frequency with which the values $x_1, x_2, \ldots, x_n$ lie in the set A. The theorem above tells us that if the random process has no non-trivial time-invariant properties, then this visiting frequency converges to the same number for $n \to \infty$ with probability 1. Similar statements can be derived for the frequencies of bigrams, trigrams, and other "words."

3.5 A Remark on Existence

While the theorem in the previous section provides the sufficient condition for almost everywhere unique time-averages, the corresponding existence theorem requires a different and much more involved argument.

We can then state the existence theorem as follows:

Theorem 3. (The Ergodic Theorem) Let P^* be a stationary random process and f an integrable function. Then the time-averages of f converge on almost all sample paths,

$$P^*\left(\lim_{n\to\infty} A_n f \text{ exists}\right) \;=\; 1.$$

Moreover, if all time-invarient sets are trivial under P^*, then

$$P^*\left(\lim_{n\to\infty} A_n f = E^*[f]\right) \;=\; 1,$$

where $E^*[f]$ is the expectation of f under P^*.

The existence statement included in this theorem says that integrable functions interact nicely with the averaging operation.

There are essentially two ways of proving this. One is to show that the space of integrable functions is spanned by a smaller set whose time-dependent components vanish as we increase the number of terms.[4] The other is to proceed via the so-called maximal ergodic theorem, which provides a Markov-like bound on the probability that any of the n first time-averages deviate substantially from the expectation of f.[5] Both proofs are rather complex, and I will not reproduce them here.

3.6 Attractor Distributions

In many situations, we are not drawing our sample path from a stationary distribution P^*, but from some closely related non-stationary distribution P. As we saw in the case of Markov chains, however, the stationary distribution P^* may still sometimes work as an attractor for "similar" processes P, so that the limiting behavior of P is the one modeled by P^*.

In this section, we formulate a condition that guarantees that this attractor status for the stationary distribution. For this, we first need a definition:

Definition 5. We say that a distribution P is **absolutely continuous** with respect to P^* if

$$P^*(B) = 0 \qquad \Longrightarrow \qquad P(B) = 0$$

for all measurable sets B. We write this $P^* \gg P$.

[4] See J. von Neumann: "Proof of the Quasi-ergodic Hypothesis," *Proceedings of the Natural Academy of Sciences of the USA*, Vol. 18(1), 1932.

[5] See G. D. Birkhoff: "Proof of the ergodic theorem," *Proceedings of the Natural Academy of Sciences of the USA*, Vol. 17(12), 1931.

Note that $P^* \gg P$ does not imply $P^*(B) \geq P(B)$ for all B. It only means that setting $P^*(B) = 0$ "squeezes down" $P(B)$ to 0, and setting $P(B) > 0$ "pushes up" $P^*(B)$ above 0. (This is only translates to the inequality $P^*(B) \geq P(B)$ if we consider all positive numbers $p \in (0, 1]$ as equally big.)

Note also that when $P^* \gg P$,

$$P^*(B) = 1 \qquad \Longrightarrow \qquad P(B) = 1,$$

since the complement of a set for which $P^*(B) = 1$ will satisfy $P^*(B^c) = 0$ and hence $P(B^c) = 0$.

We can now state the simple but useful corrollary:

Theorem 4. (Convergence of Averages) Suppose the measures P and P^* define two random process, and that

1. P^* is a stationary distribution;
2. P is absolutely continuous with respect to P^*, $P^* \gg P$.

Then for any integrable f,

$$P\left(\lim_{n\to\infty} A_n f \text{ exists}\right) \;=\; 1.$$

If further all time-invarient sets are trivial under P^*, then

$$P\left(\lim_{n\to\infty} A_n f = E^*[f]\right) \;=\; 1.$$

Proof. The probability of these two events are 1 under P^*. By the absolute continuity assumption, the same holds under P. □

As an application of this theorem, consider a stationary random process P^*. Define the conditional process P as

$$P(B) \;=\; P^*(B \,|\, X_1 = x_1)$$

for some $x_1 \in \mathcal{X}$ with $P^*(X_1 = x_1) > 0$.

The two random processes are then similar except for the fact that we force P to take a detour past x_1 at time $n = 1$. This is relevant, for instance, if we model some kind of diffusion process with a known starting state.

However, since $P^*(X_1 = x_1) > 0$, the new process is absolutely continuous with respect to the old. Any event which is impossible under the stationary distribution is also impossible under its conditionalized version. The two processes thus have the same limit behavior, and the time-averages of both are given by the stationary distribution.

One way of reading the theorem above is that different stationary distributions can act as "sinks" which describe different kind of limit behavior. If a random process has limiting time-averages along almost all sample paths, but time-invariant sets with non-trivial probabilities, then each time-invariant set can describe a different sink with a different limiting behavior.