Ergodic Properties for Stationary Measures

Łukasz Dębowski

April 2, 2019

1 Distributions on strings

Suppose that we want to imitate generation or prediction of some unknown text of an arbitrary length. Let us assume that the alphabet $\mathbb{X}$, i.e., the set of symbols with which we can write the texts, is countable. Suppose moreover that we are given an infinite sequence $(y_i)_{i\in\mathbb{N}}$ of symbols $y_i \in \mathbb{X}$, from which we can infer a statistical model. We will denote its finite substrings as

$$y_j^k = (y_j, y_{j+1}, ..., y_k). \tag{1.1}$$

Then we may define a function

$$\phi(w||y_1^\infty) := \begin{cases} \lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} \mathbf{1}\Big\{y_{j+1}^{j+|w|} = w\Big\} & \text{if the limit exists,} \\ \perp & \text{else,} \end{cases} \tag{1.2}$$

being the relative frequency of a string $w \in \mathbb{X}^*$ in sequence $(y_i)_{i\in\mathbb{N}}$.

Consider function $p(w) = \phi(w||y_1^\infty)$ given by formula (1.2) for a fixed sequence $(y_i)_{i\in\mathbb{N}}$. Suppose that $p(w) \neq \perp$ for all $w \in \mathbb{X}^*$. Then it can be easily checked that p is a stationary probability distribution in the following sense.

Definition 1.1 (probability distributions on strings) *A probability distribution on strings $\mathbb{X}^*$ is a function $p : \mathbb{X}^* \to [0, 1]$ such that:*

$$p(\lambda) = 1 \text{ for the empty sequence } \lambda, \tag{1.3}$$

$$p(w) \geq 0 \text{ for } w \in \mathbb{X}^*, \tag{1.4}$$

$$\sum_{x\in\mathbb{X}} p(wx) = p(w) \text{ for } w \in \mathbb{X}^*. \tag{1.5}$$

Additionally, a probability distribution p on strings $\mathbb{X}^*$ is called stationary if

$$\sum_{x\in\mathbb{X}} p(xw) = p(w) \text{ for } w \in \mathbb{X}^*. \tag{1.6}$$

For a distribution p and strings $u, w \in \mathbb{X}^*$, let us write the conditional probability

$$p(u|w) := \frac{p(wu)}{p(w)}. \tag{1.7}$$

The Markov distributions are defined as follows.

Definition 1.2 (IID and Markov distributions) *A probability distribution p on strings $\mathbb{X}^*$ is called n-th order Markov if*

$$p(x|wz) = p(x|z) \text{ for } x \in \mathbb{X},\ w \in \mathbb{X}^*,\ z \in \mathbb{X}^n. \tag{1.8}$$

The zeroth order Markov distributions are also called IID distributions and they satisfy

$$p(x_1^n) = p(x_1)...p(x_n) \text{ for } x_i \in \mathbb{X}. \tag{1.9}$$

The first order Markov distributions are simply called Markov distributions and they satisfy

$$p(x_1^n) = p(x_1)p(x_2|x_1)...p(x_n|x_{n-1}) \text{ for } x_i \in \mathbb{X}. \tag{1.10}$$

The first order Markov distributions are stationary if and only if

$$\sum_{x_1 \in \mathbb{X}} p(x_1)p(x_2|x_1) = p(x_2) \text{ for } x_2 \in \mathbb{X}. \tag{1.11}$$

For more background on Markov distributions see Norris (1997).

2 Stationary measures

In this section we would like to show that probability distributions on strings can be extended to probability measures on infinite sequences, which opens a powerful toolbox of probability measure theory at our disposal.

The basic construction is as follows. Let $\mathbb{X}^{\mathbb{T}}$, where $\mathbb{T} = \mathbb{N}$ or $\mathbb{T} = \mathbb{Z}$, be the set of one-sided or two-sided infinite sequences, i.e.,

$$\mathbb{X}^{\mathbb{T}} := \{(y_i)_{i \in \mathbb{T}} : y_t \in \mathbb{X}, t \in \mathbb{T}\} \tag{2.1}$$

Subsequently, let us introduce some discrete random variables X_t defined as projections

$$X_t : \mathbb{X}^{\mathbb{T}} \ni (y_i)_{i \in \mathbb{T}} \mapsto y_t \in \mathbb{X}, \quad k \in \mathbb{T}. \tag{2.2}$$

Having these projections, we will subsets of infinite sequences that are fixed on certain positions. These subsets are called cylinder sets $(X_j^k = x_j^k)$ and will be defined

$$(X_j^k = x_j^k) := \left\{(y_i)_{i \in \mathbb{T}} \in \mathbb{X}^{\mathbb{T}} : y_j^k = x_j^k\right\}, \quad j, k \in \mathbb{T}, x_j^k \in \mathbb{X}^*. \tag{2.3}$$

Cylinder sets form a certain class of sets

$$\mathcal{A} := \left\{(X_j^k = x_j^k) : j, k \in \mathbb{T}, x_j^k \in \mathbb{X}^*\right\}. \tag{2.4}$$

Class of sets $\mathcal{A}$ defined in (2.4) induces an important generated σ-field, which will be denoted

$$\mathcal{X}^{\mathbb{T}} := \sigma(\mathcal{A}) = \sigma(\{(X_t \in A) : t \in \mathbb{T}, A \in \mathcal{X}\}). \tag{2.5}$$

Class $\mathcal{X}^{\mathbb{T}}$ will be called the product σ-field.

There is an important theorem concerning the links between probability distributions on finite sequences and probability measures on sets of infinite sequences.

Definition 2.1 (stationary measure) *Let us define the shift operation*

$$T : \mathbb{X}^{\mathbb{Z}} \ni (y_i)_{i\in\mathbb{Z}} \mapsto (y_{i+1})_{i\in\mathbb{Z}} \in \mathbb{X}^{\mathbb{Z}}. \tag{2.6}$$

A probability measure P on $(\mathbb{X}^{\mathbb{Z}}, \mathcal{X}^{\mathbb{Z}})$ is called stationary if for all $G \in \mathcal{X}^{\mathbb{Z}}$ we have

$$P(G) = P(T^{-1}(G)). \tag{2.7}$$

Theorem 2.2 (Kolmogorov process theorem) *We have the following facts:*

1. *If function $p : \mathbb{X}^* \to \mathbb{R}$ is a probability distribution on strings, i.e., it satisfies axioms (1.3)–(1.5) then there exists a unique probability measure P on the measurable space of one-sided infinite sequences $(\mathbb{X}^{\mathbb{N}}, \mathcal{X}^{\mathbb{N}})$ such that for any $x_1^n \in \mathbb{X}^*$,*

$$P(X_1^n = x_1^n) = p(x_1^n). \tag{2.8}$$

2. *If function $p : \mathbb{X}^* \to \mathbb{R}$ is a stationary probability distribution on strings, i.e., it satisfies axioms (1.3)–(1.6) then there exists a unique stationary probability measure P on the measurable space of two-sided infinite sequences $(\mathbb{X}^{\mathbb{Z}}, \mathcal{X}^{\mathbb{Z}})$ such that for any $j \in \mathbb{Z}$ and $x_1^n \in \mathbb{X}^*$,*

$$P(X_{j+1}^{j+n} = x_1^n) = p(x_1^n). \tag{2.9}$$

Proof: See Theorems 36.1–36.2 of Billingsley (1979). □

3 Ergodic measures

There are a few equivalent characterizations of ergodic distributions. The standard definition of ergodic distributions is as follows.

Definition 3.1 (ergodic measure) *Let us define the invariant σ-field*

$$\mathcal{I} := \left\{ G \in \mathcal{X}^{\mathbb{Z}} : G = T^{-1}(G) \right\}. \tag{3.1}$$

A probability measure P on $(\mathbb{X}^{\mathbb{Z}}, \mathcal{X}^{\mathbb{Z}})$ is called ergodic if for all $G \in \mathcal{I}$ we have

$$P(G) \in \{0, 1\}. \tag{3.2}$$

The second, more intuitive, characterization of ergodic measures is provided by the Birkhoff ergodic theorem. Let random variables X_t be the projections (2.2). We notice that the event that the relative frequency of a certain string is equal to its probability belongs to the invariant σ-field, i.e.,

$$\left(\lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} \mathbf{1}\left\{ X_{j+1}^{j+|w|} = w \right\} = p(w) \right) \in \mathcal{I}. \tag{3.3}$$

The Birkhoff ergodic theorem asserts that this event has probability 1 if p is ergodic.

Theorem 3.2 (Birkhoff ergodic theorem) *Let P be a stationary probability measure on $(\mathbb{X}^{\mathbb{Z}}, \mathcal{X}^{\mathbb{Z}})$. Measure P is ergodic if and only if for any real random variable Y on $(\mathbb{X}^{\mathbb{Z}}, \mathcal{X}^{\mathbb{Z}})$ such that $\mathbf{E}\,|Y| < \infty$, we have*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} Y \circ T^j = \mathbf{E}\, Y \ \textit{P-almost surely.} \tag{3.4}$$

Proof:

- ($\Longrightarrow$) See Theorem 5.2 in Section 5.
- ($\Longleftarrow$) Assume that equality (3.4) holds for any random variable Y such that $\mathbf{E}\,|Y| < \infty$. If we take $Y = I_G$, where $G \in \mathcal{I}$ is a shift-invariant set then

$$I_G = I_G \lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} I_G \circ T^j = I_G \mathbf{E}\, I_G = I_G P(G) \ P\text{-almost surely}, \tag{3.5}$$

  so applying the expectation to both sides of the above displayed equation yields $P(G) = [P(G)]^2$. Hence $P(G) \in \{0, 1\}$ for all $G \in \mathcal{I}$. Thus equality (3.4) may hold for any integrable random variable Y only if P is a stationary ergodic measure, indeed.

□

Let random variables X_t be the projections (2.2). If we put $Y = \mathbf{1}\Big\{X_1^{|w|} = w\Big\}$ then for an ergodic measure we obtain

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} \mathbf{1}\Big\{X_{j+1}^{j+|w|} = w\Big\} = p(w) \ P\text{-almost surely}, \tag{3.6}$$

since $\mathbf{E}\,\mathbf{1}\Big\{X_1^{|w|} = w\Big\} = P\left(X_1^{|w|} = w\right) = p(w)$. That is, ergodic distributions on strings enjoy a frequency interpretation of probability. Using Theorem 4.4, to be discussed in the next section, it can be shown that a stationary measure P is ergodic if and only if equality (3.6) holds for all strings $w \in \mathbb{X}^*$.

4 How to check ergodicity?

A question arises which particular distributions are ergodic. There is an equivalent characterization of ergodic distributions which constitutes a practical test. The following theorem is the first step to carve it out.

Theorem 4.1 *Let P be a stationary probability measure on $(\mathbb{X}^{\mathbb{Z}}, \mathcal{X}^{\mathbb{Z}})$. Measure P is ergodic if and only if for all $A, B \in \mathcal{X}^{\mathbb{Z}}$ we have*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} P(T^{-j}(A) \cap B) = P(A)P(B). \tag{4.1}$$

Proof: If (4.1) holds for all $A, B \in \mathcal{I}$ then putting $A = B$ we obtain $P(A) = [P(A)]^2 \in \{0, 1\}$ and P is ergodic. Conversely, if P is ergodic then the Brikhoff ergodic theorem (Theorem 3.2) and the dominated convergence yield

$$P(A)P(B) = \int_B P(A)dP = \int_B \left[\lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} I_A \circ T^j\right] dP$$
$$= \lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} P(T^{-j}(A) \cap B) \tag{4.2}$$

for all $A, B \in \mathcal{X}^{\mathbb{Z}}$. □

The above condition inspires a related stronger condition, called mixing. Mixing corresponds in a sense to a stochastic process ultimately forgetting about some details of its past.

Definition 4.2 (mixing measure) *Let P be a stationary probability measure on $(\mathbb{X}^{\mathbb{Z}}, \mathcal{X}^{\mathbb{Z}})$. Measure P is called mixing when for all $A, B \in \mathcal{X}^{\mathbb{Z}}$ we have*

$$\lim_{n\to\infty} P(T^{-n}(A) \cap B) = P(A)P(B). \tag{4.3}$$

Mixing is a stronger property than being ergodic.

Theorem 4.3 *If a stationary measure P is mixing then it is ergodic.*

Proof: Condition (4.3) for $A = B \in \mathcal{I}$ yields $P(A) = [P(A)]^2$. Hence $P(A) \in \{0, 1\}$ and p is ergodic. An alternative proof of the same fact observes that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} P(T^{-j}(A) \cap B) = \lim_{n\to\infty} P(T^{-n}(A) \cap B) \tag{4.4}$$

if the second limit exists. □

It turns out that both ergodicity and mixing can be verified by checking conditions (4.1) and (4.3) only for cylinder sets. The rigorous proof is quite technical so we omit it. Interested readers are referred to Gray (2009, Lemma 7.15). Thus we have the following propositions.

Theorem 4.4 (ergodic test) *A stationary probability distribution p is ergodic if and only if*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} \sum_{z\in\mathbb{X}^j} p(uzw) = p(u)p(w) \tag{4.5}$$

for all strings $u, w \in \mathbb{X}^*$.

Theorem 4.5 (mixing test) *A stationary probability distribution p is mixing if and only if*

$$\lim_{n\to\infty} \sum_{z\in\mathbb{X}^n} p(uzw) = p(u)p(w) \tag{4.6}$$

for all strings $u, w \in \mathbb{X}^*$.

Thus, we may check that all IID distributions are mixing, whereas a stationary Markov distribution is mixing if all transition probabilities are greater than some positive constant. Hence the respective processes are also ergodic. First, the IID distributions.

Theorem 4.6 *Any IID distribution is mixing.*

Proof: By Theorem 4.5, it is sufficient to prove condition (4.6). But this is straightforward since $p(uzw) = p(u)p(z)p(w)$. Hence $\sum_{z\in\mathbb{X}^n} p(uzw) = p(u)p(w)$. □

The second proposition concerns Markov distributions.

Theorem 4.7 *A stationary Markov distribution is mixing if there exists an an $\epsilon > 0$ such that $p(x|y) \geq \epsilon$ for all $x, y \in \mathbb{X}$.*

Proof: By Theorem 4.5, it is sufficient to prove condition (4.6). But this holds if

$$\lim_{n\to\infty} p_n(x|y) = p(x), \tag{4.7}$$

where

$$p_n(x|y) := \begin{cases} p(x|y), & n = 1, \\ \sum_{z\in\mathbb{X}} p(x|z)p_{n-1}(z|y), & n > 1. \end{cases} \tag{4.8}$$

Now to prove (4.7), we observe that

$$p_n(x|y) - p(x) = \sum_{z\in\mathbb{X}} p(x|z)\left[p_{n-1}(z|y) - p(z)\right], \tag{4.9}$$

$$\sum_{x\in\mathbb{X}} (p_n(x|y) - p(x)) = 0. \tag{4.10}$$

Hence for $n > 1$, by $p(x|y) \geq \epsilon$, we have

$$\begin{aligned} \sum_{x\in\mathbb{X}} |p_n(x|y) - p(x)| &= \sum_{x\in\mathbb{X}} \left| \sum_{z\in\mathbb{X}} p(x|z)\left[p_{n-1}(z|y) - p(z)\right] \right| \\ &= \sum_{x\in\mathbb{X}} \left| \sum_{z\in\mathbb{X}} \left[p(x|z) - \epsilon\right]\left[p_{n-1}(z|y) - p(z)\right] \right| \\ &\leq \sum_{x\in\mathbb{X}} \left[p(x|z) - \epsilon\right] \sum_{z\in\mathbb{X}} |p_{n-1}(z|y) - p(z)| \\ &\leq (1-\epsilon) \sum_{z\in\mathbb{X}} |p_{n-1}(z|y) - p(z)| . \end{aligned} \tag{4.11}$$

Thus we obtain $|p_n(x|y) - p(x)| \leq 2(1-\epsilon)^{n-1}$, which proves convergence (4.7). Hence the Markov process is mixing. □

5 Proof of the Birkhoff ergodic theorem

In this section we will present a relatively simple proof of the Birkhoff ergodic theorem. First, we will demonstrate an auxiliary fact called the maximal ergodic theorem.

Theorem 5.1 (maximal ergodic theorem) *Let P be a stationary probability measure on $(\mathbb{X}^{\mathbb{Z}}, \mathcal{X}^{\mathbb{Z}})$. Let Y be a real random variable on $(\mathbb{X}^{\mathbb{Z}}, \mathcal{X}^{\mathbb{Z}})$ such that $\mathbf{E}\,|Y| < \infty$. Define $S_k := \sum_{i=0}^{k-1} Y \circ T^i$ and $M_n := \max(0, S_1, S_2, ..., S_n)$. We have*

$$\int_{M_n>0} Y dP \geq 0. \tag{5.1}$$

Proof: For $1 \leq k \leq n$ we have $M_n \circ T \geq S_k \circ T$. Hence

$$Y + M_n \circ T \geq Y + S_k \circ T = S_{k+1}. \tag{5.2}$$

Let us write it as

$$Y \geq S_{k+1} - M_n \circ T, \quad k = 1, ..., n. \tag{5.3}$$

But we also have

$$Y = S_1 \geq S_1 - M_n \circ T. \tag{5.4}$$

Both inequalities yield $Y \geq \max(S_1, S_2, ..., S_n) - M_n \circ T$. Hence

$$\int_{M_n>0} Y dP \geq \int_{M_n>0} [M_n - M_n \circ T]\, dP \tag{5.5}$$

$$= \int_{M_n>0} M_n dP - \int_{M_n>0} M_n \circ T dP. \tag{5.6}$$

Now we observe that $\int_{M_n>0} M_n dP = \int M_n dP$, whereas $\int_{M_n>0} M_n \circ T dP \leq \int M_n \circ T dP$ since $M_n \circ T \geq 0$. Moreover, by stationarity $\int M_n dP = \int M_n \circ T dP$. Hence

$$\int_{M_n>0} Y dP \geq \int M_n dP - \int M_n \circ T dP \geq 0. \tag{5.7}$$

□

In the next step, we will prove the Birkhoff ergodic theorem.

Theorem 5.2 (Birkhoff ergodic theorem) *Let P be a stationary probability measure on $(\mathbb{X}^{\mathbb{Z}}, \mathcal{X}^{\mathbb{Z}})$. For any real random variable Y on $(\mathbb{X}^{\mathbb{Z}}, \mathcal{X}^{\mathbb{Z}})$ such that $\mathbf{E}\,|Y| < \infty$, we have*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} Y \circ T^j = \mathbf{E}\,[Y|\mathcal{I}] \;\; P\text{-almost surely.} \tag{5.8}$$

Proof: We use the notation from the previous proposition. Without loss of generality, let us assume $\mathbf{E}\left[Y|\mathcal{I}\right] = 0$. Statement (5.8) can be derived applying the proof below to random variable $Y - \mathbf{E}\left[Y|\mathcal{I}\right]$. For a fixed $\epsilon > 0$ denote the shift invariant event

$$G = \left(\limsup_{n\to\infty} S_n/n > \epsilon\right). \tag{5.9}$$

We introduce random variable $Y^* = (Y - \epsilon)I_G$ and, by analogy, we define $S_k^* := \sum_{i=0}^{k-1} Y^* \circ T^i$ and $M_n^* := \max\left(0, S_1^*, S_2^*, ..., S_n^*\right)$ as in the statement of the maximal ergodic theorem. Events

$$(M_n^* > 0) = \left(\max_{1\le k\le n} S_k^* > 0\right) \tag{5.10}$$

converge to

$$\left(\sup_{k\ge 1} S_k^* > 0\right) = \left(\sup_{k\ge 1} S_k^*/k > 0\right) = \left(\sup_{k\ge 1} S_k/k > \epsilon\right) \cap G = G. \tag{5.11}$$

Inequality $\mathbf{E}\left|Y^*\right| \le \mathbf{E}\left|Y\right| + \epsilon < \infty$ allows to use the Lebesgue dominated convergence theorem, which yields

$$\int_G Y^* dP = \lim_{n\to\infty} \int_{M_n^*>0} Y^* dP \ge 0 \tag{5.12}$$

by the maximal ergodic theorem. But $G \in \mathcal{I}$ so $\int_G Y dP = \int_G \mathbf{E}\left[Y|\mathcal{I}\right] dP = 0$. Hence

$$0 \le \int_G Y^* dP = \int_G Y dP - \epsilon P(G) = -\epsilon P(G), \tag{5.13}$$

and thus $P(G) = 0$. Hence we obtain that $\limsup_{n\to\infty} S_n/n \le \epsilon$ holds P-almost surely for an arbitrary $\epsilon > 0$. Applying the analogous reasoning to random variable $-Y$ yields $\liminf_{n\to\infty} S_n/n \ge -\epsilon$. As a result we derive $\lim_{n\to\infty} S_n/n = 0$, which is the desired claim. □

The original proof of the ergodic theorem given by Birkhoff (1932) was much longer. It was considerably shortened by Garsia (1965), whose proof we have reproduced here.

6 Ergodic decomposition

The ergodic decomposition theorem is the last important characterization of ergodic measures which we will discuss. Let us denote the set of stationary distributions on strings as $\mathbb{S}$ and the set of stationary ergodic distributions on strings as $\mathbb{E}$. There is a simple geometric interpretation of the ergodic distributions. Namely, set $\mathbb{S}$ is *convex*, i.e., for any distributions $r_1, r_2 \in \mathbb{S}$, distribution p satisfying

$$p(w) = q_1 r_1(w) + q_2 r_2(w), \quad w \in \mathbb{X}^*, \tag{6.1}$$

where $0 \leq q_1 = 1 - q_2 \leq 1$, also belongs to $\mathbb{S}$. In contrast, a distribution $p \in \mathbb{S}$ is called *extremal* if we cannot write it as the convex combination (6.1) for $0 < q_1 = 1 - q_2 < 1$ and $r_1 \neq r_2$. It can be argued that extremal distribution should be ergodic. For suppose that $p \in \mathbb{S}$ is not ergodic. Then for some event $G \in \mathcal{I}$ we have $0 < P(G) < 1$ and we may write p as (6.1), where

$$q_1 = P(G), \qquad r_1(x_1^n) = P\left(X_1^n = x_1^n | \, G\right), \tag{6.2}$$

$$q_2 = P(G^c), \qquad r_2(x_1^n) = P\left(X_1^n = x_1^n | \, G^c\right). \tag{6.3}$$

Since $r_1, r_2 \in \mathbb{S}$, hence p is not extremal if $r_1 \neq r_2$.

The following theorem states a stronger fact, namely, that a stationary distribution is extremal if and only if it is ergodic.

Theorem 6.1 (ergodic decomposition) *For any stationary distribution $p \in \mathbb{S}$ there exists a unique probability measure Q on an appropriately defined measurable space $(\mathbb{E}, \mathcal{E})$ of ergodic distributions on strings such that for all events $A \in \mathcal{X}^{\mathbb{Z}}$ we have the integral representation*

$$P(A) = \int_{\mathbb{E}} E(A) dQ(e), \tag{6.4}$$

where P and E are the probability measures on $(\mathbb{X}^{\mathbb{Z}}, \mathcal{X}^{\mathbb{Z}})$ induced by distributions p and e, respectively.

To make some discussion of the above, if $p \in \mathbb{E}$ is an ergodic distribution then, by its uniqueness, measure Q must be concentrated on the singleton set $\{p\}$, i.e., $Q(\{p\}) = 1$. Conversely, if measure Q is concentrated on some singleton set $\{e\}$, then distribution p given by (6.4) is ergodic for the simple reason that $p = e$. Hence, a stationary distribution is extremal if and only if it is ergodic, indeed.

Proof of Theorem 6.1:

- (sketch of existence of Q) The random set function

$$F(A) := P(A|\mathcal{I}) \tag{6.5}$$

satisfies $F(G) \in \{0, 1\}$ almost surely for any $G \in \mathcal{I}$. It can be shown that F is a stationary ergodic probability measure almost surely (since the probability space is countably generated!!). Denoting

$$f(x_1^n) := F(X_{j+1}^{j+n} = x_1^n), \qquad Q(B) := P(f \in B), \tag{6.6}$$

(and taking care of measurability!!), we obtain

$$P(A) = \int P(A|\mathcal{I}) dP = \int F(A) dP = \int \mathbf{1}\{f \in \mathbb{E}\} F(A) dP = \int_{\mathbb{E}} E(A) dQ(e). \tag{6.7}$$

- (sketch of uniqueness of Q) Suppose that we have

$$P(A) = \int_{\mathbb{E}} E(A) dQ_1(e) = \int_{\mathbb{E}} E(A) dQ_2(e), \tag{6.8}$$

for some measures Q_1 and Q_2 and all events $A \in \mathcal{X}^{\mathbb{Z}}$. Define measure $S = Q_1 + Q_2$ and the sets of ergodic distributions

$$B_1 := \left(\frac{dQ_1}{dS} > \frac{dQ_2}{dS} \right) \in \mathbb{E}, \qquad B_2 := \left(\frac{dQ_1}{dS} < \frac{dQ_2}{dS} \right) \in \mathbb{E}. \tag{6.9}$$

For $p \in \mathbb{S}$ define the sets of sequences

$$\Omega_p := \bigcap_{w \in \mathbb{X}^*} \left(\lim_{n \to \infty} \frac{1}{n} \sum_{j=0}^{n-1} \mathbf{1}\Big\{ X_{j+1}^{j+|w|} = w \Big\} = p(w) \right) \in \mathcal{I} \subset \mathcal{X}^{\mathbb{Z}}. \tag{6.10}$$

Now we construct sets of sequences $A_i = \bigcup_{p \in B_i} \Omega_p \in \mathcal{I}$ (we can show that they are measurable!!). Since by the Birkhoff ergodic theorem we have

$$E(\Omega_p) = \mathbf{1}\{p = e\} \text{ for } e \in \mathbb{E}, \tag{6.11}$$

we obtain $E(A_i) = \mathbf{1}\{e \in B_i\}$. In view of our hypothesis, we infer

$$\begin{aligned} 0 &= \int_{\mathbb{E}} E(A_i) dQ_1(e) - \int_{\mathbb{E}} E(A_i) dQ_2(e) \\ &= \int_{\mathbb{E}} E(A_i) \left(\frac{dQ_1}{dS} - \frac{dQ_2}{dS} \right) dS \\ &= \int_{\mathbb{E}} \mathbf{1}\{e \in B_i\} \left(\frac{dQ_1}{dS} - \frac{dQ_2}{dS} \right)(e) dS(e) \quad = \int_{B_i} \left(\frac{dQ_1}{dS} - \frac{dQ_2}{dS} \right) dS, \end{aligned} \tag{6.12}$$

which implies $S(B_i) = 0$. Hence $Q_1 = Q_2$.

□

The idea of the ergodic decomposition comes from Rokhlin (1962). Formula (6.4) is a special case of the Choquet theorem for convex sets on an appropriate vector space. The Choquet theorem states that any element belonging to a convex set can be expressed as a generalized convex combination of extremal elements. However, in general, the representation given by the Choquet need not be unique. For more background in ergodic theorems, an interested reader is referred to Gray (2009).

References

P. Billingsley. *Probability and Measure*. New York: John Wiley, 1979.

G. D. Birkhoff. Proof of the ergodic theorem. *Proceedings of the National Academy of Sciences of the United States of America*, 17:656–660, 1932.

A. M. Garsia. A simple proof of E. Hopf's maximal ergodic theorem. *Journal of Mathematics and Mechanics*, 14:381–382, 1965.

R. M. Gray. *Probability, Random Processes, and Ergodic Properties*. New York: Springer, 2009.

J. R. Norris. *Markov Chains*. Cambridge: Cambridge University Press, 1997.

V. A. Rokhlin. On the fundamental ideas of measure theory. *American Mathematical Society Translations, series 1*, 10:1–54, 1962.