

Information Theory and Statistics

Lecture 3: Stationary ergodic processes

Łukasz Dębowski
ldebowsk@ipipan.waw.pl

Ph. D. Programme 2013/2014

Measurable space

Definition (measurable space)

Measurable space (Ω, \mathcal{J}) is a pair where Ω is a certain set (called the *event space*) and $\mathcal{J} \subset 2^\Omega$ is a σ -field. The σ -field \mathcal{J} is an algebra of subsets of Ω which satisfies

- $\Omega \in \mathcal{J}$,
- $\mathbf{A} \in \mathcal{J}$ implies $\mathbf{A}^c \in \mathcal{J}$, where $\mathbf{A}^c := \Omega \setminus \mathbf{A}$,
- $\mathbf{A}, \mathbf{B} \in \mathcal{J}$ implies $\mathbf{A} \cup \mathbf{B} \in \mathcal{J}$,
- $\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3, \dots \in \mathcal{J}$ implies $\bigcup_{n \in \mathbb{N}} \mathbf{A}_n \in \mathcal{J}$.

The elements of \mathcal{J} are called *events*, whereas the elements of Ω are called *elementary events*.

Probability measure

Definition (probability measure)

Probability measure $\mathbf{P} : \mathcal{J} \rightarrow [0, 1]$ is a normalized measure, i.e., a function of events that satisfies

- $\mathbf{P}(\Omega) = 1$,
- $\mathbf{P}(\mathbf{A}) \geq 0$ for $\mathbf{A} \in \mathcal{J}$,
- $\mathbf{P}(\bigcup_{n \in \mathbb{N}} \mathbf{A}_n) = \sum_{n \in \mathbb{N}} \mathbf{P}(\mathbf{A}_n)$ for disjoint events $\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3, \dots \in \mathcal{J}$,

The triple $(\Omega, \mathcal{J}, \mathbf{P})$ is called a *probability space*.

Invariant measures and dynamical systems

Definition (invariant measure)

Consider a probability space $(\Omega, \mathcal{J}, \mathbf{P})$ and an invertible operation $\mathbf{T} : \Omega \rightarrow \Omega$ such that $\mathbf{T}^{-1}\mathbf{A} \in \mathcal{J}$ for $\mathbf{A} \in \mathcal{J}$. Measure \mathbf{P} is called *\mathbf{T} -invariant* and \mathbf{T} is called *\mathbf{P} -preserving* if

$$\mathbf{P}(\mathbf{T}^{-1}\mathbf{A}) = \mathbf{P}(\mathbf{A})$$

for any event $\mathbf{A} \in \mathcal{J}$.

Definition (dynamical system)

A dynamical system $(\Omega, \mathcal{J}, \mathbf{P}, \mathbf{T})$ is a quadruple that consists of a probability space $(\Omega, \mathcal{J}, \mathbf{P})$ and a \mathbf{P} -preserving operation \mathbf{T} .

Stationary processes

Definition (stationary process)

A stochastic process $(\mathbf{X}_i)_{i=-\infty}^{\infty}$, where $\mathbf{X}_i : \Omega \rightarrow \mathbb{X}$ are discrete random variables, is called *stationary* if there exists a distribution of blocks $\mathbf{p} : \mathbb{X}^* \rightarrow [0, 1]$ such that

$$\mathbf{P}(\mathbf{X}_{i+1} = \mathbf{x}_1, \dots, \mathbf{X}_{i+n} = \mathbf{x}_n) = \mathbf{p}(\mathbf{x}_1 \dots \mathbf{x}_n)$$

for each i and n .

Example (IID process)

If variables \mathbf{X}_i are independent and have identical distribution $\mathbf{P}(\mathbf{X}_i = \mathbf{x}) = \mathbf{p}(\mathbf{x})$ then $(\mathbf{X}_i)_{i=-\infty}^{\infty}$ is stationary.

Invariant measures and stationary processes

Example

Let measure \mathbf{P} be \mathbf{T} -invariant and let $\mathbf{X}_0 : \Omega \rightarrow \mathbb{X}$ be a random variable on $(\Omega, \mathcal{J}, \mathbf{P})$. Define random variables $\mathbf{X}_i(\omega) = \mathbf{X}_0(\mathbf{T}^i\omega)$. For

$$\mathbf{A} = (\mathbf{X}_{i+1} = x_1, \dots, \mathbf{X}_{i+n} = x_n)$$

we have

$$\begin{aligned} \mathbf{T}^{-1}\mathbf{A} &= \left\{ \mathbf{T}^{-1}\omega : \mathbf{X}_0(\mathbf{T}^{i+1}\omega) = x_1, \dots, \mathbf{X}_0(\mathbf{T}^{i+n}\omega) = x_n \right\} \\ &= \left\{ \omega : \mathbf{X}_0(\mathbf{T}^{i+2}\omega) = x_1, \dots, \mathbf{X}_0(\mathbf{T}^{i+n+1}\omega) = x_n \right\} \\ &= (\mathbf{X}_{i+2} = x_1, \dots, \mathbf{X}_{i+n+1} = x_n). \end{aligned}$$

Because $\mathbf{P}(\mathbf{T}^{-1}\mathbf{A}) = \mathbf{P}(\mathbf{A})$, process $(\mathbf{X}_i)_{i=-\infty}^{\infty}$ is stationary.

Process theorem

Theorem (process theorem)

Let distribution of blocks $\mathbf{p} : \mathbb{X}^* \rightarrow [0, 1]$ satisfy conditions

$$\sum_{\mathbf{x} \in \mathbb{X}} \mathbf{p}(\mathbf{xw}) = \mathbf{p}(\mathbf{w}) = \sum_{\mathbf{x} \in \mathbb{X}} \mathbf{p}(\mathbf{wx})$$

and $\mathbf{p}(\lambda) = 1$, where λ is the empty word. Let event space

$$\Omega = \{\omega = (\omega_i)_{i=-\infty}^{\infty} : \omega_i \in \mathbb{X}\}$$

consist of infinite sequences and introduce random variables $\mathbf{X}_i(\omega) = \omega_i$. Let \mathcal{J} be the σ -field generated by all cylinder sets

$(\mathbf{X}_i = \mathbf{s}) = \{\omega \in \Omega : \mathbf{X}_i(\omega) = \mathbf{s}\}$. Then there exists a unique probability measure \mathbf{P} on \mathcal{J} that satisfies

$$\mathbf{P}(\mathbf{X}_{i+1} = \mathbf{x}_1, \dots, \mathbf{X}_{i+n} = \mathbf{x}_n) = \mathbf{p}(\mathbf{x}_1 \dots \mathbf{x}_n).$$

Invariant measures and stationary processes

Theorem

Let $(\Omega, \mathcal{J}, \mathbf{P})$ be the probability space constructed in the process theorem. Measure \mathbf{P} is \mathbf{T} -invariant for the operation

$$(\mathbf{T}\omega)_i := \omega_{i+1},$$

called shift. Moreover, we have $\mathbf{X}_i(\omega) = \mathbf{X}_0(\mathbf{T}^i\omega)$.

Proof

By the π - λ theorem it suffices to prove $\mathbf{P}(\mathbf{T}^{-1}\mathbf{A}) = \mathbf{P}(\mathbf{A})$ for $\mathbf{A} = (\mathbf{X}_{i+1} = x_1, \dots, \mathbf{X}_{i+n} = x_n)$. But $\mathbf{X}_i(\omega) = \mathbf{X}_0(\mathbf{T}^i\omega)$. Hence $\mathbf{T}^{-1}\mathbf{A} = (\mathbf{X}_{i+2} = x_1, \dots, \mathbf{X}_{i+n+1} = x_n)$. In consequence, we obtain $\mathbf{P}(\mathbf{T}^{-1}\mathbf{A}) = \mathbf{P}(\mathbf{A})$ by stationarity.

Generated dynamical system

Definition

The triple $(\Omega, \mathcal{J}, \mathbf{P})$ and the quadruple $(\Omega, \mathcal{J}, \mathbf{P}, \mathbf{T})$ constructed in the previous two theorems will be called the *probability space* and the *dynamical system generated* by a stationary process $(\mathbf{X}_i)_{i=-\infty}^{\infty}$ (with a given block distribution $\mathbf{p} : \mathbb{X}^* \rightarrow [0, 1]$).

Markov processes

Theorem

A Markov chain $(X_i)_{i=-\infty}^{\infty}$ is stationary if and only if it has marginal distribution $P(X_i = k) = \pi_k$ and transition probabilities $P(X_{i+1} = l | X_i = k) = p_{kl}$ which satisfy

$$\pi_l = \sum_k \pi_k p_{kl}.$$

Matrix (p_{kl}) is called the transition matrix.

Proof

If $(X_i)_{i=-\infty}^{\infty}$ is stationary then marginal distribution $P(X_i = k)$ and transition probabilities $P(X_{i+1} = l | X_i = k)$ may not depend on i . We also have

$$P(X_{i+1} = k_1, \dots, X_{i+n} = k_n) = p(k_1 \dots k_n) := \pi_{k_1} p_{k_1 k_2} \dots p_{k_{n-1} k_n}$$

Function $p(k_1 \dots k_n)$ satisfies $\sum_x p(xw) = p(w) = \sum_x p(wx)$. Hence we obtain $\pi_l = \sum_k \pi_k p_{kl}$. On the other hand, if $P(X_i = k) = \pi_k$ and $P(X_{i+1} = l | X_i = k) = p_{kl}$ hold with $\pi_l = \sum_k \pi_k p_{kl}$ then function $p(k_1 \dots k_n)$ satisfies $\sum_x p(xw) = p(w) = \sum_x p(wx)$ and the process is stationary.

For a given transition matrix the stationary distribution may not exist or there may be many stationary distributions.

Example

Let variables X_i assume values in natural numbers and let $P(X_{i+1} = k + 1 | X_i = k) = 1$. Then the process $(X_i)_{i=1}^{\infty}$ is not stationary. Indeed, assume that there is a stationary distribution $P(X_i = k) = \pi_k$. Then we obtain $\pi_{k+1} = \pi_k$ for any k . Such distribution does not exist if there are infinitely many k .

Example

For the transition matrix

$$\begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

we may choose

$$\begin{pmatrix} \pi_1 & \pi_2 \end{pmatrix} = \begin{pmatrix} a & 1 - a \end{pmatrix}, \quad a \in [0, 1].$$

Block entropy

Definition (block)

Blocks of variables are written as $\mathbf{X}_k^l = (\mathbf{X}_i)_{k \leq i \leq l}$.

Definition (block entropy)

The entropy of the block of n variables drawn from a stationary process will be denoted as

$$\mathbf{H}(n) := \mathbf{H}(\mathbf{X}_1^n) = \mathbf{H}(\mathbf{X}_1, \dots, \mathbf{X}_n) = \mathbf{H}(\mathbf{X}_{i+1}, \dots, \mathbf{X}_{i+n}).$$

For convenience, we also put $\mathbf{H}(0) = 0$.

Block entropy (continued)

Theorem

Let Δ be the difference operator, $\Delta F(n) := F(n) - F(n - 1)$. Block entropy satisfies

$$\begin{aligned}\Delta H(n) &= H(X_n | X_1^{n-1}), \\ \Delta^2 H(n) &= -I(X_1; X_n | X_2^{n-1}),\end{aligned}$$

where $H(X_1 | X_1^0) := H(X_1)$ and $I(X_1; X_2 | X_2^1) := I(X_1; X_2)$.

Remark: Hence, for any stationary process, block entropy $H(n)$ is nonnegative ($H(n) \geq 0$), nondecreasing ($\Delta H(n) \geq 0$) and concave ($\Delta^2 H(n) \leq 0$).

Proof

We have

$$\begin{aligned}H(X_n | X_1^{n-1}) &= H(X_1^n) - H(X_1^{n-1}) \\ &= H(n) - H(n-1) = \Delta H(n), \\ -I(X_1; X_n | X_2^{n-1}) &= H(X_1^n) - H(X_1^{n-1}) - H(X_2^n) + H(X_2^{n-1}) \\ &= H(n) - 2H(n-1) + H(n-2) = \Delta^2 H(n).\end{aligned}$$

Entropy rate

Definition (entropy rate)

The *entropy rate* of a stationary process will be defined as

$$h = \lim_{n \rightarrow \infty} \Delta H(n) = H(1) + \sum_{n=2}^{\infty} \Delta^2 H(n).$$

By the previous theorem, we have $0 \leq h \leq H(1)$.

Example

Let $(X_i)_{i=-\infty}^{\infty}$ be a stationary Markov chain with marginal distribution $P(X_i = k) = \pi_k$ and transition probabilities $P(X_{i+1} = l | X_i = k) = p_{kl}$. We have $\Delta H(n) = H(X_n | X_1^{n-1}) = H(X_n | X_{n-1})$, so

$$h = - \sum_{kl} \pi_k p_{kl} \log p_{kl}.$$

Entropy rate (continued)

Theorem

Entropy rate satisfies equality

$$h = \lim_{n \rightarrow \infty} \frac{H(n)}{n}.$$

Proof

Difference $\Delta H(\cdot)$ is nonincreasing. Hence block entropy $H(n) = H(m) + \sum_{k=m+1}^n \Delta H(k)$ satisfies inequalities

$$H(m) + (n - m) \cdot \Delta H(n) \leq H(n) \leq H(m) + (n - m) \cdot \Delta H(m). \quad (1)$$

Putting $m = 0$ in the left inequality in (1), we obtain

$$\Delta H(n) \leq H(n)/n \quad (2)$$

Putting $m = n - 1$ in the right inequality in (1), we hence obtain $H(n) \leq H(n - 1) + \Delta H(n - 1) \leq H(n - 1) + H(n - 1)/(n - 1)$. Thus $H(n)/n \leq H(n - 1)/(n - 1)$. Because function $H(n)/n$ is nonincreasing, the limit $h' := \lim_{n \rightarrow \infty} H(n)/n$ exists. By (2), we have $h' \geq h$. Now we will prove the converse. Putting $n = 2m$ in the right inequality in (1) and dividing both sides by m we obtain $2h' \leq h' + h$ in the limit. Hence $h' \leq h$.

Concepts for the definition of expectation

Let us write $\mathbf{1}\{\phi\} = \mathbf{1}$ if proposition ϕ is true and $\mathbf{1}\{\phi\} = \mathbf{0}$ if proposition ϕ is false. The characteristic function of a set \mathbf{A} is defined as

$$\mathbf{1}_{\mathbf{A}}(\omega) := \mathbf{1}\{\omega \in \mathbf{A}\}.$$

The supremum $\sup_{\mathbf{a} \in \mathbf{A}} \mathbf{a}$ is defined as the least real number \mathbf{r} such that $\mathbf{r} \geq \mathbf{a}$ for all $\mathbf{a} \in \mathbf{A}$. On the other hand, infimum $\inf_{\mathbf{a} \in \mathbf{A}} \mathbf{a}$ is the largest real number \mathbf{r} such that $\mathbf{r} \leq \mathbf{a}$ for all $\mathbf{a} \in \mathbf{A}$.

Expectation

Definition (expectation)

Let \mathbf{P} be a probability measure. For a discrete random variable $\mathbf{X} \geq \mathbf{0}$, the *expectation* (integral, or average) is defined as

$$\int \mathbf{X} d\mathbf{P} := \sum_{x:\mathbf{P}(\mathbf{X}=x)>0} \mathbf{P}(\mathbf{X} = x) \cdot x.$$

For a real random variable $\mathbf{X} \geq \mathbf{0}$, we define

$$\int \mathbf{X} d\mathbf{P} := \sup_{\mathbf{Y} \leq \mathbf{X}} \int \mathbf{Y} d\mathbf{P},$$

where the supremum is taken over all discrete variables \mathbf{Y} that satisfy $\mathbf{Y} \leq \mathbf{X}$.

Expectation II

Definition (expectation)

Integrals over subsets are defined as

$$\int_A \mathbf{X} d\mathbf{P} := \int \mathbf{X} I_A d\mathbf{P}.$$

For random variables that assume negative values, we put

$$\int \mathbf{X} d\mathbf{P} := \int_{\mathbf{X} > 0} \mathbf{X} d\mathbf{P} - \int_{\mathbf{X} < 0} (-\mathbf{X}) d\mathbf{P},$$

unless both terms are infinite. A more frequent notation is

$$\mathbf{E} \mathbf{X} \equiv \mathbf{E}_{\mathbf{P}} \mathbf{X} \equiv \int \mathbf{X} d\mathbf{P},$$

where we suppress the index \mathbf{P} in $\mathbf{E}_{\mathbf{P}} \mathbf{X}$ for probability measure \mathbf{P} .

Invariant algebra

Definition (invariant algebra)

Let $(\Omega, \mathcal{J}, \mathbf{P}, \mathbf{T})$ be a dynamical system. The set of events which are invariant with respect to operation \mathbf{T} ,

$$\mathcal{I} := \left\{ \mathbf{A} \in \mathcal{J} : \mathbf{A} = \mathbf{T}^{-1}\mathbf{A} \right\},$$

will be called the *invariant algebra*.

Examples of invariant events

Example

Let $(\Omega, \mathcal{F}, \mathbf{P}, \mathbf{T})$ be the dynamical system generated by a stationary process $(\mathbf{X}_i)_{i=-\infty}^{\infty}$, where $\mathbf{X}_i : \Omega \rightarrow \{0, 1\}$. The operation \mathbf{T} results in shifting variables \mathbf{X}_i , i.e., $\mathbf{X}_i(\omega) = \mathbf{X}_0(\mathbf{T}^i\omega)$. Thus these events belong to the invariant algebra \mathcal{I} :

$$(\mathbf{X}_i = 1 \text{ for all } i) = \bigcap_{i=-\infty}^{\infty} (\mathbf{X}_i = 1),$$

$$(\mathbf{X}_i = 1 \text{ for infinitely many } i \geq 1) = \bigcap_{i=1}^{\infty} \bigcup_{j=i}^{\infty} (\mathbf{X}_j = 1),$$

$$\left(\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n \mathbf{X}_k = a \right) = \bigcap_{p=1}^{\infty} \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} \left(\left| \frac{1}{n} \sum_{k=1}^n \mathbf{X}_k - a \right| \leq \frac{1}{p} \right).$$

Ergodic processes

If the process $(\mathbf{X}_i)_{i=-\infty}^{\infty}$ is a sequence of independent identically distributed variables then the probability of the mentioned events is **0** or **1**. Following our intuition for independent variables, we may think that a stationary process is well-behaved if the probability of invariant events is **0** or **1**.

Definition (ergodicity)

A dynamical system $(\Omega, \mathcal{J}, \mathbf{P}, \mathbf{T})$ is called *ergodic* if any event from the invariant algebra has probability **0** or **1**, i.e.,

$$\mathbf{A} \in \mathcal{I} \implies \mathbf{P}(\mathbf{A}) \in \{0, 1\}.$$

Analogously, we call a stationary process $(\mathbf{X}_i)_{i=-\infty}^{\infty}$ *ergodic* if the dynamical system generated by this process is ergodic.

Ergodic theorem

We say that Φ holds with probability **1** if $\mathbf{P}(\{\omega : \Phi(\omega) \text{ is true}\}) = \mathbf{1}$.

In 1931, Georg David Birkhoff (1884–1944) showed this fact:

Theorem (ergodic theorem)

Let $(\Omega, \mathcal{J}, \mathbf{P}, \mathbf{T})$ be a dynamical system and define stationary process $\mathbf{X}_i(\omega) := \mathbf{X}_0(\mathbf{T}^i \omega)$ for a real random variable \mathbf{X}_0 on the probability space $(\Omega, \mathcal{J}, \mathbf{P})$. The dynamical system is ergodic if and only if for any real random variable \mathbf{X}_0 where $\mathbf{E} |\mathbf{X}_0| < \infty$ equality

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n \mathbf{X}_k = \mathbf{E} \mathbf{X}_0$$

holds with probability **1**.

Ergodicity in information theory

In information theory, we often invoke the ergodic theorem in the following way. Namely, for a stationary ergodic process $(\mathbf{X}_i)_{i=-\infty}^{\infty}$, with probability $\mathbf{1}$ we have

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n \left[-\log P(\mathbf{X}_k | \mathbf{X}_{k-m}^{k-1}) \right] &= \mathbf{E} \left[-\log P(\mathbf{X}_0 | \mathbf{X}_{-m}^{-1}) \right] \\ &= H(\mathbf{X}_0 | \mathbf{X}_{-m}^{-1}). \end{aligned}$$

This equality holds since $P(\mathbf{X}_k | \mathbf{X}_{k-m}^{k-1})$ is a random variable on the probability space generated by process $(\mathbf{X}_i)_{i=-\infty}^{\infty}$.

An example of a nonergodic process

Example (nonergodic process)

Let $(U_i)_{i=-\infty}^{\infty}$ and $(W_i)_{i=-\infty}^{\infty}$ be independent stationary ergodic processes having different distributions and let an independent variable Z have distribution $P(Z = 0) = p \in (0, 1)$ and $P(Z = 1) = 1 - p$. We will consider process $(X_i)_{i=-\infty}^{\infty}$, where

$$X_i = 1\{Z = 0\}U_i + 1\{Z = 1\}W_i.$$

Assume that $P(U_1^p = w) \neq P(W_1^p = w)$. Then

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n 1\{X_k^{k+p-1} = w\} = 1\{Z = 0\}P(U_1^p = w) + 1\{Z = 1\}P(W_1^p = w),$$

which is not constant. Hence $(X_i)_{i=-\infty}^{\infty}$ is not ergodic.

Ergodic Markov chains

Theorem

Let $(X_i)_{i=-\infty}^{\infty}$ be a stationary Markov chain, where $P(X_{i+1} = l | X_i = k) = p_{kl}$, $P(X_i = k) = \pi_k$, and the variables take values in a countable set. These conditions are equivalent:

- 1 Process $(X_i)_{i=-\infty}^{\infty}$ is ergodic.
- 2 There are no two disjoint closed sets of states; a set A of states is called closed if $\sum_{l \in A} p_{kl} = 1$ for each $k \in A$.
- 3 For a given transition matrix (p_{kl}) there exists a unique stationary distribution π_k .

Proof (1) \Rightarrow (2): Suppose that there are two disjoint closed sets of states A and B . Then we obtain

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n \mathbf{1}\{X_k \in A\} = \mathbf{1}\{X_1 \in A\} \neq \text{const.}$$

Examples of ergodic processes

Example

A sequence of independent identically distributed random variables is an ergodic process.

Example

Consider a stationary Markov chain $(X_i)_{i=-\infty}^{\infty}$ where $P(X_{n+1} = j | X_n = i) = p_{ij}$ and $P(X_1 = i) = \pi_i$. It is ergodic for

$$\left(\begin{array}{cc} \pi_1 & \pi_2 \end{array} \right) = \left(\begin{array}{cc} 1/2 & 1/2 \end{array} \right), \quad \left(\begin{array}{cc} p_{11} & p_{12} \\ p_{21} & p_{22} \end{array} \right) = \left(\begin{array}{cc} 0 & 1 \\ 1 & 0 \end{array} \right),$$

and nonergodic for

$$\left(\begin{array}{cc} \pi_1 & \pi_2 \end{array} \right) = \left(\begin{array}{cc} a & 1 - a \end{array} \right), \quad \left(\begin{array}{cc} p_{11} & p_{12} \\ p_{21} & p_{22} \end{array} \right) = \left(\begin{array}{cc} 1 & 0 \\ 0 & 1 \end{array} \right).$$