Lecture 4 Markov chain and entropy rate

September 9th, 2022

Outline

Stochastic process

Definition

A stochastic process is a collection of random variables that arise from the same probability space. It can be mathematically represented by collection

$$\{X_t,\ t \in I\},$$

where X_t denotes the tth random variable in the process, and the index t runs over an index set I which is arbitrary.

In this course, we focus mostly on discrete-time sources; i.e., sources with the countable index set $I = \{1, 2, \ldots\}$. Each such source is denoted by

$$\mathbf{X} := \{X_n\}_{n=1}^{\infty} = \{X_1, X_2, X_3, \cdots\}$$

as an infinite sequence of random variables, where all the random variables take on values from a common generic alphabet $\mathcal{X} \subset \mathbb{R}$.

The source X completely characterized by the sequence of joint cdf's $\{F_{X^n}\}_{n=1}^{\infty}$. When the alphabet $\mathcal{X}$ is finite, the source can be equivalently described by the sequence of joint probability mass function (pmf's):

$$P_{X^n}(a^n) = Pr[X_1 = a_1, X_2 = a_2, \cdots, X_n = a_n]$$

for all $a^n = (a_1, a_2, \ldots, a_n) \in \mathcal{X}^n$, $n = 1, 2, \ldots$.

Memoryless process

The process **X** is said to be memoryless if its random variables are independent and identically distributed (i.i.d.). Here by independence, we mean that any finite sequence $X_{i_1}, X_{i_2}, \ldots, X_{i_n}$ of random variables satisfies

$$Pr[X_{i_1} = x_1, X_{i_2} = x_2, \ldots, X_{i_n} = x_n] = \Pi_{l=1}^{n} Pr[X_{i_l} = x_l].$$

for all $x_l \in \mathcal{X}$, $l = 1, \ldots, n$; we also say that these random variables are mutually independent. Furthermore, the notion of identical distribution means that

$$Pr[X_i = x] = Pr[X_1 = x]$$

for any $x \in \mathcal{X}$ and $i = 1, 2, \cdots$; i.e., all the process' random variables are governed by the same marginal distribution.

Stationary process

The process X is said to be stationary (or strictly stationary) if the probability of every sequence or event is unchanged by a left (time) shift, or equivalently, if any $j = 1, 2, \ldots$, the joint distribution of $(X_1, X_2, \ldots, X_n)$ satisfies

$$\begin{aligned} &Pr[X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n] \\ &\qquad = Pr[X_{j+1} = x_1, X_{j+2} = x_2, \ldots, X_{j+n} = x_n] \end{aligned}$$

for all $x_l \in \mathcal{X}$, $l = 1, \ldots, n$.

Stationary process

The process X is said to be stationary (or strictly stationary) if the probability of every sequence or event is unchanged by a left (time) shift, or equivalently, if any $j = 1, 2, \ldots$, the joint distribution of $(X_1, X_2, \ldots, X_n)$ satisfies

$$\begin{aligned} &Pr[X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n] \\ &\qquad = Pr[X_{j+1} = x_1, X_{j+2} = x_2, \ldots, X_{j+n} = x_n] \end{aligned}$$

for all $x_l \in \mathcal{X}$, $l = 1, \ldots, n$.

It is direct to verify that a memoryless source is stationary. Also, for a stationary source, its random variables are identically distributed.

For a random process $X = \{X_n\}_{n=1}^{\infty}$ with alphabet $\mathcal{X}$ (i.e., $\mathcal{X} \subset \mathbb{R}$ is the range of each X_i) defined over probability space $(\Omega, \mathcal{F}, P)$, consider $\mathcal{X}^\infty$, the set of all sequences $\mathbf{x} := (x_1, x_2, x_3, \ldots)$ of real numbers in $\mathcal{X}$. An event in $\mathcal{F}_X$, the smallest σ-field generated by all open sets of $\mathcal{X}^\infty$ (i.e., the Borel σ-field of $\mathcal{X}^\infty$), is said to be $\mathbb{T}$-invariant with the left shift (or shift transformation) $\mathbb{T} : \mathcal{X}^\infty \to \mathcal{X}^\infty$ if

$$\mathbb{T}E \subset E,$$

where

$$\mathbb{T}E := \{\mathbb{T}\mathbf{x} : \mathbf{x} \in E\}$$

and

$$\mathbb{T}\mathbf{x} := \mathbb{T}(x_1, x_2, x_3, \ldots) = (x_2, x_3, \ldots).$$

Let

$$\mathbb{T}^{-1}E = E, \tag{1.1}$$

then

$$\mathbb{T}E = \mathbb{T}(\mathbb{T}^{-1}E) = E,$$

and hence E is constituted only by the $\mathbb{T}$-invariant groups because

$$\cdots = \mathbb{T}^{-2}E = \mathbb{T}^{-1}E = E = \mathbb{T}E = \mathbb{T}^2E = \cdots .$$

The sets that satisfy (1.1) are sometimes referred to as *ergodic sets* because as time goes by, the set always stays in the state that it has been before.

The process $\mathbf{X}$ is said to be *ergodic* if any ergodic set in $\mathcal{F}_X$ has probability either 1 or 0.

Theorem (Pointwise ergodic theorem)

Consider a discrete-time stationary random process, $\boldsymbol{X} = \{X_n\}_{n=1}^{\infty}$. For real-valued function $f(\cdot)$ on $\mathbb{R}$ with finite mean (i.e., $|E[f(X_n)]| < \infty$), there exists a random variable Y such that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} f(X_k) = Y \text{ with probability } 1.$$

If, in addition to stationarity, the process is also ergodic, then

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} f(X_k) = E[f(X_1)] \text{ with probability } 1.$$

Example

Consider the process $\{X_i\}_{i=1}^{\infty}$ *consisting of a family of i.i.d. binary random variables (obviously, it is stationary and ergodic). Define the function* $f(\cdot)$ *by* $f(0) = 0$ *and* $f(1) = 1$. *Hence,*

$$E[f(X_n)] = P_{X_n}(0)f(0) + P_{X_n}(1)f(1)$$

is finite. By the pointwise ergodic theorem, we have

$$\lim_{n\to\infty} \frac{f(X_1) + f(X_2) + \cdots + f(X_n)}{n} = \lim \frac{X_1 + X_2 + \cdots + X_n}{n} = P_X(1).$$

Outline

Markov chain for three random variables

Example

Three random variables X, Y and Z are said to form a Markov chain if

$$P_{X,Y,Z}(x,y,x) = P_X(x) \cdot P_{Y|X}(y|x) \cdot P_{Z|Y}(z|y);$$

i.e., $P_{Z|X,Y}(z|x,y) = P_{Z|Y}(z|y)$. *This is usually denoted by*

$$X \to Y \to Z.$$

$X \to Y \to Z$ is sometimes read as "X and Z are conditionally independent given Y" because it can be shown that the above definition is equivalent to

$$P_{X,Z|Y}(x, z|y) = P_{X|Y}(x|y) \cdot P_{Z|Y}(z|y).$$

Therefore, $X \to Y \to Z$ is equivalent to $Z \to Y \to X$.
Accordingly, the Markovian notation is sometimes expressed as $X \leftrightarrow Y \leftrightarrow Z$.

kth order Markov chain

The sequence of random variables $\{X_n\}_{n=1}^{\infty} = X_1, X_2, X_3, \ldots$ with common finite-alphabet $\mathcal{X}$ is said to form a kth order Markov chain (or kth order Markov source or process) if for all $n > k$, $x_1 \in \mathcal{X}$, $i = 1, \ldots, n$,

$$\begin{aligned} &Pr[X_n = x_n | X_{n-1} = x_{n-1}, \ldots, X_1 = x_1] \\ &\qquad = Pr[X_n = x_n | X_{n-1} = x_{n-1}, \ldots, X_{n-k} = x_{n-k}]. \end{aligned}$$

Each $x_{n-k}^{n-1} := (x_{n-k}, x_{n-k+1}, \ldots, x_{n-1}) \in \mathcal{X}^k$ is called the state of the Markov chain at time n.

When $k = 1$, then $\{X_n\}_{n=1}^{\infty}$ is called a first-order Markov chain (or just a Markov chain). This means that for any $n > 1$, the random variables $X_1, X_2, \ldots, X_n$ directly satisfy the conditional independence property

$$Pr[X_i = x_i | X^{i-1} = x^{i-1}] = Pr[X_i = x_i | X_{i-1} = x_{i-1}]$$

for all $x_i \in \mathcal{X}$, $i = 1, \ldots, n$; this property is denoted by

$$X_1 \to X_2 \to \cdots \to X_n$$

for $n > 2$.

Irreducible Markov chain

A kth order Markov chain is *irreducible* if with some probability, we can go from any state in $\mathcal{X}^k$ to another state in a finite number of steps, i.e., for all $x^k, y^k \in \mathcal{X}^k$ there exists an integer $j \geq 1$ such that

$$Pr\{X_j^{k+j-1} = x^k | X_1^k = y^k\} > 0.$$

A kth order Markov chain is said to be *time-invariant* or *homogeneous*, if for every $n > k$,

$$Pr[X_n = x_n | X_{n-1} = x_{n-1}, \ldots, X_{n-k} = x_{n-k}] \\ = Pr[X_{k+1} = x_{k+1} | X_k = x_k, \ldots, X_1 = x_1].$$

Therefore, a homogeneous first-order Markov chain can be defined through its transition probability:

$$[Pr\{X_2 = x_2 | X_1 = x_1\}]_{|\mathcal{X} \times \mathcal{X}|},$$

and its initial state distribution $P_{X_1}(x)$.

In a first-order Markov chain, the period $d(x)$ of state $x \in \mathcal{X}$ is defined by

$$d(x) := \text{gcd}\{n \in \{1, 2, 3, \ldots\} : Pr\{X_{n+1} = x | X_1 = x\} > 0\},$$

where gcd denotes the greatest common divisor; in other words, if the Markov chain starts in state x, then the chain cannot return to state x at any time that is not a multiple of $d(x)$. If $Pr\{X_{n+1} = x | X_1 = x\} = 0$ for all n, we say that state x has infinite period and write $d(x) = \infty$. We also say that state x is *aperiodic* if $d(x) = 1$ and *periodic* if $d(x) > 1$. Furthermore, the first-order Markov chain is called *aperiodic* if all its states are aperiodic. In other words, the first-order Markov chain ia aperiodic if

$$\text{gcd}\{n \in \{1, 2, 3, \ldots\} : Pr\{X_{n+1} = x | X_1 = x\} > 0\} = 1 \ \forall x \in \mathcal{X}.$$

In an irreducible first-order Markov chain, all states have the same period. Hence, if one state in such a chain is aperiodic, then the entire Markov chain is aperiodic.

A distribution $\pi(\cdot)$ on $\mathcal{X}$ is said to be a *stationary* distribution for a homogeneous first-order Markov chain, if for every $y \in \mathcal{X}$,

$$\pi(y) = \sum_{x \in \mathcal{X}} \pi(x) Pr\{X_2 = y | X_1 = x\}.$$

For a finite-alphabet homogeneous first-order Markov chain, $\pi(\cdot)$ always exists; furthermore, $\pi(\cdot)$ is unique if the Markov chain is irreducible and aperiodic,

$$\lim_{n \to \infty} Pr\{X_{n+1} = y | X_1 = x\} = \pi(y)$$

for all states x and y in $\mathcal{X}$. If the initial state distribution is equal to stationary distribution, then the homogenous first-order Markov chain becomes a stationary distribution, then the homogenous first-order Markov chain becomes a stationary process.

A finite-alphabet stationary Markov chain is an ergodic process if and only if it is irreducible.

Example

Consider a two-state Markov chain with probability transition matrix

$$P = \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix}.$$

The stationary distribution μ can be found by solving the equation $\mu P = \mu$. We have $\mu = (\mu_1, \mu_2)$, where

$$\mu_1 = \frac{\beta}{\alpha+\beta}, \ \mu_2 = \frac{\alpha}{\alpha+\beta}.$$

So the entropy of the state X_n at time n is

$$H(X_n) = H(\frac{\beta}{\alpha+\beta}, \frac{\alpha}{\alpha+\beta}).$$

Outline

Entropy rate

Definition

The entropy rate of a stochastic process $\{X_i\}$ is defined by

$$H(\mathcal{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, \cdots, X_n)$$

when the limit exists.

Example

If $X_1, X_2, \ldots$ are i.i.d. random variables. Then

$$H(\mathcal{X}) = \lim_{n\to\infty} \frac{H(X_1, X_2, \cdots, X_n)}{n} = \lim_{n\to\infty} \frac{nH(X_1)}{n} = H(X_1).$$

Example

If the random variables $X_1, X_2, \cdots, X_n$ are independent but not identically distributed, then

$$H(X_1, X_2, \cdots, X_n) = \sum_{i=1}^{n} H(X_i).$$

We choose a sequence of distributions on $X_1, X_2, \cdots$, such that the limit of $\frac{1}{n}\sum H(X_i)$ does not exist. An example of such a sequence is a random binary sequence where $p_i = P(X_i = 1)$ is not constant but a function of i. For example,

$$p_i = \begin{cases} 0.5 & 2k < \log\log i \leq 2k+1 \\ 0 & 2k+1 < \log\log i \leq 2k+2. \end{cases}$$

The running average of the $H(X_i)$ will oscillate between 0 and 1 and will not have a limit. Thus, $H(\mathcal{X})$ is not defined for this process.

We can also define a related quantity for entropy rate:

$$H'(\mathcal{X}) = \lim_{n\to\infty} H(X_n|X_{n-1}, X_{n-2}, \cdots, X_1)$$

Theorem

For a stationary stochastic process, the above two limits exist and are equal.

$$H(\mathcal{X}) = H'(\mathcal{X}).$$

Theorem

For a stationary stochastic process, $H(X_n|X_{n-1},\cdots,X_1)$ is nonincreasing in n and has a limit $H'(\mathcal{X})$.

Proof.

We have

$$\begin{aligned} H(X_{n+1}|X_1,X_2,\cdots,X_n) &\leq H(X_{n+1}|X_n,\cdots,X_2) \\ &= H(X_n|X_{n-1},\cdots,X_1). \end{aligned}$$

where the equality follows from the stationary of the process. Since $H(X_n|X_{n-1},\cdots,X_1)$ is a decreasing sequence of nonnegative numbers, it has a limit, $H'(\mathcal{X})$. □

Lemma

If $a_n \to a$ *as* $n \to \infty$ *and*

$$b_n = \frac{1}{n} \sum_{i=1}^{n} a_i,$$

then $b_n \to a$ *as* $n \to \infty$.

Proof of the theorem.

By the chain rule,

$$\frac{H(X_1, X_2, \ldots, X_n)}{n} = \frac{1}{n} \sum_{i=1}^{n} H(X_i | X_{i-1}, \ldots, X_1),$$

that is, the entropy rate is the time average of the conditional entropies. But we know that the conditional entropies tend to a limit H'. Hence, their running average has a limit, which is equal to the limit H' of the terms. Thus,

$$H(\mathcal{X}) = \lim \frac{H(X_1, X_2, \cdots, X_n)}{n} = \lim H(X_n | X_{n-1}, \cdots, X_1) = H'(\mathcal{X}).$$

□

For a stationary Markov chain, the entropy rate is given by

$$H(\mathcal{X}) = H'(\mathcal{X}) = \lim H(X_n|X_{n-1}, \cdots, X_1) = \lim H(X_n|X_{n-1}) = H(X_2|X_1),$$

where the third equality follows from

$$\begin{aligned}
& H(X_n|X_{n-1}, \cdots, X_1) \\
= & -\sum_{x_1, \cdots, x_n} p(x_1, x_2, \cdots, x_{n-1}) p(x_n|x_{n-1}, \cdots, x_1) \log p(x_n|x_{n-1}, \cdots, x_1) \\
= & -\sum_{x_1, \cdots, x_n} p(x_1, x_2, \cdots, x_{n-1}) p(x_n|x_{n-1}) \log p(x_n|x_{n-1}) \\
= & -\sum_{x_{n-1}, x_n} p(x_n|x_{n-1}) \log p(x_n|x_{n-1}) \sum_{x_1, \cdots, x_{n-2}} p(x_1, x_2, \cdots, x_{n-1}) \\
= & -\sum_{x_{n-1}, x_n} p(x_{n-1}) p(x_n|x_{n-1}) \log p(x_n|x_{n-1}) \\
= & -\sum_{x_{n-1}, x_n} p(x_{n-1}, x_n) \log p(x_n|x_{n-1}) \\
= & H(X_n|X_{n-1}).
\end{aligned}$$

Recall that the stationary distribution μ is the solution of the equations

$$\mu_j = \sum_i \mu_i P_{ij} \text{ for all } j.$$

We can express the conditional entropy explicitly in the following theorem.

Theorem

Let $\{X_i\}$ be a stationary Markov chain with stationary distribution μ and transition matrix P. Let $X_1 \sim \mu$. Then the entropy rate is

$$H(\mathcal{X}) = -\sum_{ij} \mu_i P_{ij} \log P_{ij}.$$

Proof.

$$H(\mathcal{X}) = H(X_2|X_1) = \sum_i \mu_i (\sum_j -P_{ij} \log P_{ij}).$$

□

Two-state Markov chain

Example

The entropy rate of the two-state Markov chain is

$$H(\mathcal{X}) = H(X_2|X_1) = \frac{\beta}{\alpha+\beta}H(\alpha) + \frac{\alpha}{\alpha+\beta}H(\beta).$$