Entropy Rate of a Stochastic Process

Timo Mulder Jorn Peters

Universiteit van Amsterdam

January 2015

Overview

Stochastic Process $\{X_i\}$

Definition (Stochastic Process)

A discrete stochastic process is a sequence of RVs:

$$\ldots, X_{-3}, X_{-2}, X_{-1}, X_0, X_1, X_2, \ldots$$

Stochastic Process $\{X_i\}$

Definition (Stochastic Process)

A discrete stochastic process is a sequence of RVs:

$$\ldots, X_{-3}, X_{-2}, X_{-1}, X_0, X_1, X_2, \ldots$$

- Characterized by its joint probability mass function:

$$P_{X_1,X_2\ldots,X_n}(x_1, x_2 \ldots, x_n)$$

- Arbitrary dependence between RVs

Markov Process $\{X_i\}$

Stochastic process with the Markov property

Definition (Markov Process)

A stochastic process is a Markov process if for $n = 1, 2, \ldots$

$$\begin{aligned} P(X_{n+1} = x_{n+1} \mid X_n = x_n, \ldots, X_1 = x_1) \\ = P(X_{n+1} = x_{n+1} \mid X_n = x_n) \end{aligned}$$

For all $x_1, x_2, \ldots, x_n, x_{n+1} \in \mathcal{X}$.

Markov Process $\{X_i\}$

Stochastic process with the Markov property

Definition (Markov Process)

A stochastic process is a Markov process if for $n = 1, 2, \ldots$

$$P(X_{n+1} = x_{n+1} \mid \boxed{X_n = x_n, \ldots, X_1 = x_1}) = P(X_{n+1} = x_{n+1} \mid \boxed{X_n = x_n})$$

For all $x_1, x_2, \ldots, x_n, x_{n+1} \in \mathcal{X}$.

Random variable only depends on its direct predecessor

Time Invariant Markov Process I

Definition (Time Invariance)

A Markov process is time invariant if for $n = 1, 2, \ldots,$

$$P(X_{n+1} = a \mid X_n = b) = P(X_2 = a \mid X_1 = b)$$

for all $a, b \in \mathcal{X}$.

Defined by:

1. It's initial state
2. A *probability transition matrix* P
   - $P = [P_{ij}], i, j \in \{1, 2, 3, ..., m\}$
   - Where $P_{ij} = Pr\{X_{n+1} = j | X_n = i\}$

Time Invariant Markov Process II

Example

$$\begin{aligned} P(X_{n+1} = b|X_n = a) \\ = P(X_2 = b|X_1 = a) \\ = P(X_9 = b|X_8 = a) \\ \text{etc.} \end{aligned}$$

Stationary Distribution

Given $P_{X_t}(\cdot)$ the probability mass function at time $t+1$ is defined as

$$P_{X_{t+1}}(\alpha) = \sum_{k=1}^{n} P(x_k) P(X_{t+1} = \alpha \mid X_t = x_k)$$
$$= \sum_{k=1}^{n} P(x_k) P_{x_k \alpha}$$

Stationary Distribution

Given $P_{X_t}(\cdot)$ the probability mass function at time $t+1$ is defined as

$$P_{X_{t+1}}(\alpha) = \sum_{k=1}^{n} P(x_k)P(X_{t+1} = \alpha \mid X_t = x_k)$$
$$= \sum_{k=1}^{n} P(x_k)P_{x_k \alpha}$$

If the probability mass at time t and time $t+1$ are the same then the process is a stationary process. In that case μ is the stationary distribution where $\mu_i = P_X(i)$.

Stationary Stochastic Process

More precise:

Definition

A stochastic process is stationary if the joint distribution of any subset of the sequence of RVs is invariant of shifts in the time index.

Stationary Stochastic Process

More precise:

Definition

A stochastic process is stationary if the joint distribution of any subset of the sequence of RVs is invariant of shifts in the time index.
That is,

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

for every n and every shift l and for all $x_1, x_2, ..., x_n \in \mathcal{X}$.

Stationary Stochastic Process

In particular this means that for any stationary stochastic process we have

$$P(X_n = a) = P(X_1 = a), \qquad \forall n, a.$$

Stationary Distribution I

- In our example we can find the stationary distribution by solving

$$\mu^T P = \mu^T$$

- Thus the stationary distribution is related to a left eigenvector of the probability transition matrix P where the eigenvalue equals 1

Irreducible and aperiodic Markov process

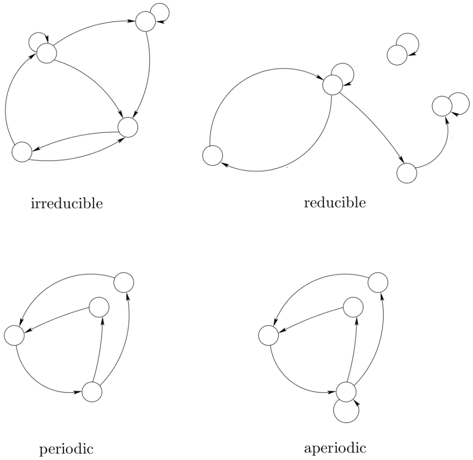

Figure: Taken from Moser, 2013

Irreducible and aperiodic Markov process

Given a time invariant Markov process $\{X_i\}$ that is irreducible and aperiodic.

Remark

$\{X_i\}$ has a unique stationary distribution.

Irreducible and aperiodic Markov process

Given a time invariant Markov process $\{X_i\}$ that is irreducible and aperiodic.

Remark

$\{X_i\}$ has a unique stationary distribution.

Remark

Independent of the starting distribution $P_{X_1}(\cdot)$. $P_{X_k}(\cdot)$ will converge to the stationary distribution μ as $k \to \infty$.

Stationary Distribution II

Example

Let us show that in the example $\mu = [\frac{3}{5}, \frac{2}{5}]$

$P_{X_k}(\cdot)$	$k = 1$
$P_{X_k}(S)$	1
$P_{X_k}(R)$	0
$P_{X_k}(\cdot)$	
$P_{X_k}(S)$	
$P_{X_k}(R)$	

Table: Convergence to stationary distribution when $k \to \infty$. (Taken from Moser, 2013)

Stationary Distribution II

Example

Let us show that in the example $\mu = [\frac{3}{5}, \frac{2}{5}]$

$P_{X_k}(\cdot)$	$k=1$	$k=2$
$P_{X_k}(S)$	1	$\frac{1}{2} = 0.5$
$P_{X_k}(R)$	0	$\frac{1}{2} = 0.5$
$P_{X_k}(\cdot)$		
$P_{X_k}(S)$		
$P_{X_k}(R)$		

Table: Convergence to stationary distribution when $k \to \infty$.
(Taken from Moser, 2013)

Stationary Distribution II

Example

Let us show that in the example $\mu = [\frac{3}{5}, \frac{2}{5}]$

$P_{X_k}(\cdot)$	$k = 1$	$k = 2$	$k = 3$
$P_{X_k}(S)$	1	$\frac{1}{2} = 0.5$	$\frac{5}{8} = 0.625$
$P_{X_k}(R)$	0	$\frac{1}{2} = 0.5$	$\frac{3}{8} = 0.375$
$P_{X_k}(\cdot)$			
$P_{X_k}(S)$			
$P_{X_k}(R)$			

Table: Convergence to stationary distribution when $k \to \infty$.
(Taken from Moser, 2013)

Stationary Distribution II

Example

Let us show that in the example $\mu = [\frac{3}{5}, \frac{2}{5}]$

$P_{X_k}(\cdot)$	$k=1$	$k=2$	$k=3$	$k=4$
$P_{X_k}(S)$	1	$\frac{1}{2}=0.5$	$\frac{5}{8}=0.625$	$\frac{19}{32}=0.59375$
$P_{X_k}(R)$	0	$\frac{1}{2}=0.5$	$\frac{3}{8}=0.375$	$\frac{13}{32}=0.40625$
$P_{X_k}(\cdot)$				
$P_{X_k}(S)$				
$P_{X_k}(R)$				

Table: Convergence to stationary distribution when $k \to \infty$.
(Taken from Moser, 2013)

Stationary Distribution II

Example

Let us show that in the example $\mu = [\frac{3}{5}, \frac{2}{5}]$

$P_{X_k}(\cdot)$	$k = 1$	$k = 2$	$k = 3$	$k = 4$
$P_{X_k}(S)$	1	$\frac{1}{2} = 0.5$	$\frac{5}{8} = 0.625$	$\frac{19}{32} = 0.59375$
$P_{X_k}(R)$	0	$\frac{1}{2} = 0.5$	$\frac{3}{8} = 0.375$	$\frac{13}{32} = 0.40625$
$P_{X_k}(\cdot)$	$k = 5$			
$P_{X_k}(S)$	$\frac{77}{128} = 0.6015625$			
$P_{X_k}(R)$	$\frac{51}{128} = 0.3984375$			

Table: Convergence to stationary distribution when $k \to \infty$. (Taken from Moser, 2013)

Stationary Distribution II

Example

Let us show that in the example $\mu = [\frac{3}{5}, \frac{2}{5}]$

$P_{X_k}(\cdot)$	$k=1$	$k=2$	$k=3$	$k=4$
$P_{X_k}(S)$	1	$\frac{1}{2} = 0.5$	$\frac{5}{8} = 0.625$	$\frac{19}{32} = 0.59375$
$P_{X_k}(R)$	0	$\frac{1}{2} = 0.5$	$\frac{3}{8} = 0.375$	$\frac{13}{32} = 0.40625$
$P_{X_k}(\cdot)$	$k=5$		$\cdots$	$k=\infty$
$P_{X_k}(S)$	$\frac{77}{128} = 0.6015625$		$\cdots$	$\frac{3}{5} = 0.6$
$P_{X_k}(R)$	$\frac{51}{128} = 0.3984375$		$\cdots$	$\frac{2}{5} = 0.4$

Table: Convergence to stationary distribution when $k \to \infty$.
(Taken from Moser, 2013)

Entropy Rate

The entropy rate of a state in the example is
$H(X_t) = H(\frac{\alpha}{\alpha+\beta}, \frac{\beta}{\alpha+\beta}) = h(\frac{\alpha}{\alpha+\beta})$

Entropy Rate

The entropy rate of a state in the example is
$H(X_t) = H(\frac{\alpha}{\alpha+\beta}, \frac{\beta}{\alpha+\beta}) = h(\frac{\alpha}{\alpha+\beta})$

This is not the entropy a the stochastic process.

Entropy Rate

The entropy rate of a state in the example is
$H(X_t) = H(\frac{\alpha}{\alpha+\beta}, \frac{\beta}{\alpha+\beta}) = h(\frac{\alpha}{\alpha+\beta})$

This is not the entropy a the stochastic process.

So what is the entropy of a stochastic process?

Entropy Rate: Some Intuition

If $\{X_i\}$ is i.i.d. it makes sense to say that $H(\{X_i\}) = H(X_1)$.

$\rightarrow$ Entropy is average bits per symbol.

Entropy Rate: Some Intuition

If $\{X_i\}$ is i.i.d. it makes sense to say that $H(\{X_i\}) = H(X_1)$.

$\rightarrow$ Entropy is average bits per symbol.

However,

Example

$\{Y_i\}$ is a source with memory such that $P_{Y_1}(0) = P_{Y_1}(1) = \frac{1}{2}$. furthermore assume that

$$P_{Y_2|Y_1}(0 \mid 0) = 0, \quad P_{Y_2|Y_1}(1 \mid 0) = 1$$
$$P_{Y_2|Y_1}(0 \mid 1) = 0, \quad P_{Y_2|Y_1}(1 \mid 1) = 1$$

Then $P_{Y_2}(1) = 1$ which means that $H(Y_2) = 0$, $H(Y_2 \mid Y_1) = 0$, $H(Y_{n+1} \mid Y_n) = 0$ and $H(Y_1, \ldots, Y_n) = 1$.

Entropy Rate: Some Intuition

If $\{X_i\}$ is i.i.d. it makes sense to say that $H(\{X_i\}) = H(X_1)$.
$\rightarrow$ Entropy is average bits per symbol.

However,

Example

$\{Y_i\}$ is a source with memory such that $P_{Y_1}(0) = P_{Y_1}(1) = \frac{1}{2}$. furthermore assume that

$$P_{Y_2|Y_1}(0 \mid 0) = 0, \quad P_{Y_2|Y_1}(1 \mid 0) = 1$$
$$P_{Y_2|Y_1}(0 \mid 1) = 0, \quad P_{Y_2|Y_1}(1 \mid 1) = 1$$

Then $P_{Y_2}(1) = 1$ which means that $H(Y_2) = 0$, $H(Y_2 \mid Y_1) = 0$, $H(Y_{n+1} \mid Y_n) = 0$ and $H(Y_1, \ldots, Y_n) = 1$. This is not the entropy of the process.

Entropy Rate: Definition

The entropy rate of a stochastic process strongly depends on the memory.

Definition (Entropy Rate of $\{X_i\}$)

The entropy rate (the entropy per source symbol) of any stochastic process $\{X_i\}$ is defined as

$$H(\{X_i\}) := \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, \ldots, X_n)$$

if the limit exists.

Entropy Rate: More Intuition

Example

Given a stochastic process $\{X_i\}$. Assume that $\{X_i\}$ is i.i.d. Then the entropy rate of $\{X_i\}$ is

$$H(\{X_i\}) = \lim_{n \to \infty} \frac{1}{n} H(X_1, \ldots, H_n) \quad = \lim_{n \to \infty} \frac{1}{n} n H(X_1) = H(X_1)$$

Entropy Rate: More Intuition

Example

Given a stochastic process $\{X_i\}$. Assume that $\{X_i\}$ is i.i.d. Then the entropy rate of $\{X_i\}$ is

$$H(\{X_i\}) = \lim_{n\to\infty} \frac{1}{n} H(X_1, \ldots, H_n) \quad = \lim_{n\to\infty} \frac{1}{n} n H(X_1) = H(X_1)$$

Example

Given the stochastic process $\{Y_i\}$. Then the entropy rate of $\{Y_i\}$ is

$$H(\{Y_i\} = \lim_{n\to\infty} \frac{1}{n} H(Y_1, \ldots, Y_n) = \lim_{n\to\infty} \frac{1}{n} = 0$$

Entropy Rate: A Related Quantity

We can also define a related quantity for entropy rate:

$$H'(\{X_i\}) = \lim_{n\to\infty} H(X_n \mid X_{n-1}, X_{n-2}, \ldots, X_1)$$

Entropy Rate: A Related Quantity

We can also define a related quantity for entropy rate:

$$H'(\{X_i\}) = \lim_{n\to\infty} H(X_n \mid X_{n-1}, X_{n-2}, \ldots, X_1)$$

$H(\{X_1\})$ is the entropy rate per source symbol of n random variables and $H'(\{X_i\})$ is the entropy rate of the last random variable given the past.

Theorem

For a stationary stochastic process the entropy rate $H(\{X_i\}$ *always exists and is identical to* $H'(\{X_i\})$*:*

$$\begin{aligned} H(\{X_i\}) &= \lim_{n\to\infty} \frac{1}{n} H(X_1, \ldots, X_n) \\ &= \lim_{n\to\infty} H(X_n \mid X_{n-1}, \ldots, X_1) = H'(\{X_i\}) \end{aligned}$$

Furthermore,

1. $H(X_n \mid X_{n-1}, \ldots, X_1)$ *is nonincreasing in* n*;*
2. $\frac{1}{n} H(X_1, \ldots, X_n)$ *is nonincreasing in* n*;*
3. $H(X_n \mid X_{n-1}, \ldots, X_1) \leqslant \frac{1}{n} H(X_1, \ldots, X_n), \quad \forall n \geqslant 1.$

Entropy Rate: Markov Chains

For a stationary Markov chain, the entropy rate is easy to calculate:

$$\begin{aligned} H(\{X_i\}) &= H'(\{X_i\}) \\ &= \lim_{n\to\infty} H(X_n \mid X_{n-1}, \ldots, X_1) \\ &= \lim_{n\to\infty} H(X_n \mid X_{n-1}) \\ &= H(X_2 \mid X_1) \end{aligned}$$

Finaly. . .

- Method to compute the entropy rate of a stochastic process;
- Using this a typical set for 'ergodic sets' can be constructed which has uses in compression/encoding.
- Also stochastic processes are widely used in moddeling in for example AI and the entropy can be used to find optimal models.