ECE253/CSE208 Introduction to Information Theory

Lecture 6: Entropy Rate

Dr. Yu Zhang

ECE Department
University of California, Santa Cruz

- Chap 4 of *Elements of Information Theory (2nd Edition) by Thomas Cover & Joy Thomas*

Markov Chain

Consider a discrete-time Markov chain $X_1, X_2, \ldots, X_{n+1}$, we have

$$\Pr(X_{n+1} = x_{n+1} | X_n = x_n, \ldots, X_1 = x_1) = \Pr(X_{n+1} = x_{n+1} | X_n = x_n)$$

$$p(x_{n+1}, \ldots, x_1) = p(x_1)p(x_2|x_1)\ldots p(x_{n+1}|x_n)$$

All knowledge of the past states is embedded in the current state.

We can build more memory by using a k^{th}-order MC:

$$p(x_{n+1}|x_n, \ldots, x_1) = p(x_{n+1}|x_n, \ldots, x_{n-k+1})$$

Homogeneous (Time-invariant) Markov chain: If $p(x_{n+1}|x_n)$ does not depend on n:

$$\Pr(X_{n+1} = j | X_n = i) = \Pr(X_2 = j | X_1 = i), \ \forall i, j \in \mathcal{S} := \{1, 2, \ldots, m\}.$$

Then, we can include all transition probabilities in a matrix $\mathbf{P}_{m \times m}$, whose (i, j)-th entry is given as

$$P_{ij} = \Pr(X_{n+1} = j | X_n = i), \ \forall i, j \in \mathcal{S} := \{1, 2, \ldots, m\}. \tag{1}$$

Hence, we see three things in a Markov chain: A sequence of random variables (the chain), a state space (from which the random variables take values), and the rules for transition (the transition probability matrix).

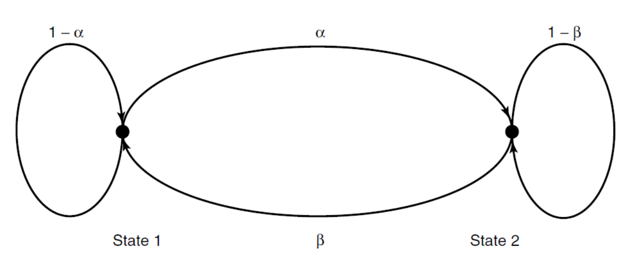
Two-state Markov chain

Example

(Two-state Markov chain).

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

As shown in the figure below:



Q: Given $\Pr(X_n = i)$, find $\Pr(X_{n+1} = j)$, $\forall i, j \in \{1, 2, \ldots, m\}$.

A: $\Pr(X_{n+1} = j) = \sum_i \Pr(X_n = i, X_{n+1} = j) = \sum_i \Pr(X_n = i) \Pr(X_{n+1} = j | X_n = i)$.

Stationary Distribution

Define a row vector to collect all state probabilities at time $n+1$

$$\boldsymbol{\pi}^{(n+1)} = \left[\pi_1^{(n+1)}, \pi_2^{(n+1)}, \ldots, \pi_m^{(n+1)}\right],$$

where $\pi_j^{(n+1)} = \Pr(X_{n+1} = j),\ \forall j \in \{1, 2, \ldots, m\}$. Hence, we have

$$\boldsymbol{\pi}^{(n+1)} = \boldsymbol{\pi}^{(n)}\mathbf{P}.$$

Definition (Stationary distribution)

Stationary distribution of a Markov chain: $\boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi}$ and $\mathbf{0} \leq \boldsymbol{\pi} \leq \mathbf{1}, \boldsymbol{\pi}\mathbf{1} = 1$, where $\mathbf{1}$ is the all-ones column vector with an appropriate dimension.

Hence, stationary distribution $\boldsymbol{\pi}$ is a *fixed point* of the transformation represented by $\mathbf{P}$, which is a *left eigenvector* of $\mathbf{P}$ corresponding to eigenvalue 1.

Example

Find the stationary distribution of the aforementioned example of the two-state MC.

$$\begin{cases} \boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi} \\ \boldsymbol{\pi}\mathbf{1} = 1 \end{cases} \quad \Rightarrow \quad \boldsymbol{\pi} = \left[\frac{\beta}{\alpha+\beta}, \frac{\alpha}{\alpha+\beta}\right].$$

Classification of States

Definition (Accessible and Communicate)

- State j is said to be accessible from state i if $P_{ij}^{(n)} > 0$ for some $n \geq 0$, which is denoted as $i \to j$.
- Two states i and j are said to communicate if they are accessible from each other, which is denoted as $i \leftrightarrow j$. (i.e., there are directed paths between i and j).

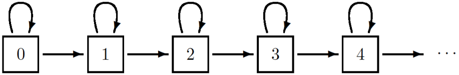


Figure: Each state is accessible from all its previous states, but not vice versa.

Communicate is a *equivalence relation*, meaning that

- Reflexivity: $i \leftrightarrow i$
- Symmetry: If $i \leftrightarrow j$, then $j \leftrightarrow i$
- Transitivity: If $i \leftrightarrow j$ and $j \leftrightarrow k$, then $i \leftrightarrow k$

An equivalence relation divides the state space into disjoint classes of equivalent states that is called **communication classes**.

Irreducible MC

Definition (Irreducible MC: every state can be reached from every other state)

It is possible to go with positive probability from any state to any other state in a finite number of steps. That is, $\exists n < \infty, \Pr(X_n = j | X_0 = i) = P_{ij}^{(n)} > 0, \forall i, j \in \mathcal{S}$.

- A Markov chain is irreducible iif all states belong to one communication class; i.e., all states communicate with each other.
- A Markov chain is reducible iff if there are two or more communication classes.
- A finite Markov chain is irreducible iff its transition graph is strongly connected (there is a path between any pair of two vertices).

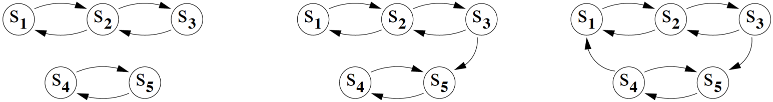


Figure: The first two Markov chains are reducible while the last one is irreducible.

Aperiodic MC

Definition (Aperiodic)

The period of a state i is defined as $k = \gcd\{n > 0, P_{ii}^{(n)} > 0\}$[a]. That is, if $P_{ii}^{(n)} = 0$ when n is not a multiple of k and k is the greatest integer with this property. If $k = 1$, the state is said to be aperiodic. **A Markov chain is aperiodic if every state is aperiodic.**

[a] **gcd** is the greatest common divisor; e.g., $\gcd\{6, 8, 10, \cdots\} = 2$; $\gcd\{3, 5, 7, \cdots\} = 1$.

- All states in the same communication class have the same period.
- An irreducible MC only needs one aperiodic state to imply the chain is aperiodic.

Consider a finite irreducible Markov chain:

- If there is a self transition in the diagram, then the chain is aperiodic.
- Suppose $P_{ii}^{(\ell)} > 0$ and $P_{ii}^{(m)} > 0$. If ℓ and m are co-prime ($\gcd(\ell, m) = 1$), then state i is aperiodic.

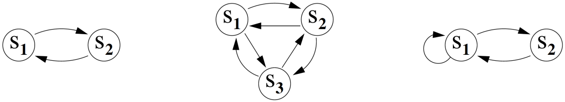


Figure: The first one has period 2 while the last two are aperiodic.

An Exercise[1]

Question: Find all communication classes and their periods.

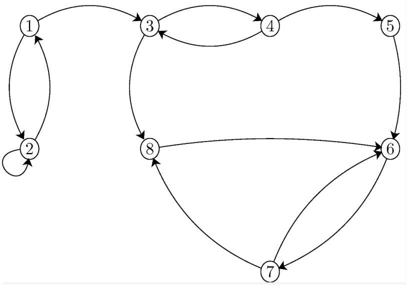


Figure: A state transition diagram.

Answer:

- Class 1={1, 2}, aperiodic
- Class 2={3, 4}, period = 2
- Class 3={5}, period = 0 (transient sate)
- Class 4={6, 7, 8}. aperiodic

[1] https://www.probabilitycourse.com/chapter11/11_2_4_classification_of_states.php

Unique Stationary Distribution and Limiting Distribution

Theorem

For an irreducible, aperiodic, and finite-state Markov chain, there exists a finite integer N *such that* $P_{ij}^{(n)} > 0$*, for all* $i, j \in \mathcal{S}$ *and all* $n \geq N$*.*

Theorem

An irreducible and aperiodic finite-state Markov chain has a unique stationary distribution.

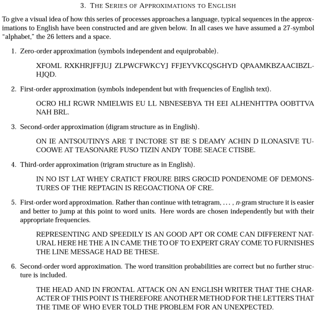
Lemma

For an irreducible, aperiodic, and finite-state Markov chain, any initial distribution converges to the unique stationary distribution as $n \to \infty$*.*

Motivating Question

Shannon used Markov Chain to describe English texts

- Zero-order, first-order, second-order letters
- First-order, second-order words



3. The Series of Approximations to English

To give a visual idea of how this series of processes approaches a language, typical sequences in the approximations to English have been constructed and are given below. In all cases we have assumed a 27-symbol "alphabet," the 26 letters and a space.

1. Zero-order approximation (symbols independent and equiprobable).

   XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGHYD QPAAMKBZAACIBZLHJQD.

2. First-order approximation (symbols independent but with frequencies of English text).

   OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI ALHENHTTPA OOBTTVA NAH BRL.

3. Second-order approximation (digram structure as in English).

   ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY ACHIN D ILONASIVE TUCOOWE AT TEASONARE FUSO TIZIN ANDY TOBE SEACE CTISBE.

4. Third-order approximation (trigram structure as in English).

   IN NO IST LAT WHEY CRATICT FROURE BIRS GROCID PONDENOME OF DEMONSTURES OF THE REPTAGIN IS REGOACTIONA OF CRE.

5. First-order word approximation. Rather than continue with tetragram, . . . , *n*-gram structure it is easier and better to jump at this point to word units. Here words are chosen independently but with their appropriate frequencies.

   REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME CAN DIFFERENT NATURAL HERE HE THE A IN CAME THE TO OF TO EXPERT GRAY COME TO FURNISHES THE LINE MESSAGE HAD BE THESE.

6. Second-order word approximation. The word transition probabilities are correct but no further structure is included.

   THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED.

Figure: Section 3 of the paper "A Mathematical Theory of Communication (1948)"

Shannon asked "Can we define a quantity which will measure, in some sense, how much information is "produced" by such a process, or better, at what rate information is produced?"

Stationary Stochastic Process

Discrete-time Information Sources:

- Communications take place continually rather than a finite period of time; e.g., Internet cellular networks, ratio stations, TV program, etc.
- The info source can be modeled as a discrete-time stochastic process $\{X_k\}_{k=1}^{\infty}$

Definition (Stationary stochastic process)

A stochastic process $\{X_k\}$ is *strongly stationary* if the joint distribution of any subset of the sequence is invariant w.r.t. time shifts. That is,

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

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

Q: Is any Markov chain a stationary stochastic process? **A:** No.

Example

Consider an MC with $\mathbf{P} = \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix}$ and the initial probability state $\boldsymbol{\pi}^{(0)} = (1,\, 0)$, then $\boldsymbol{\pi}^{(1)} = \boldsymbol{\pi}^{(0)} \cdot \mathbf{P} = (1/2,\, 1/2) \neq \boldsymbol{\pi}^{(0)}$. So it is not stationary.

Entropy Rate for Stochastic Processes

Q: How does the entropy of a sequence grow with n?

A: We define the entropy rate as the rate of growth:

Definition (Per symbol entropy of n random variables)

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

when the limit exists.

Special cases:

- $\{X_i\}$ are i.i.d.: $H(\mathcal{X}) = H(X_1)$.
- $\{X_i\}$ are independent but not identical: $H(\mathcal{X}) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} H(X_i)$. But, the limit may not even exist; e.g, $H(X_i) = i$.

Definition (Conditional entropy of the last random variable given the past.)

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

Entropy Rate for Stationary Stochastic Processes

Theorem (Entropy rate)

For a stationary stochastic process, $H(\mathcal{X}) = H'(\mathcal{X})$.

Proof: First, we show that $H'(\mathcal{X})$ is well-defined. Note that due to the stationarity of the process, we have

$$0 \leq H(X_n|X_{n-1},\ldots,X_1) \leq H(X_n|X_{n-1},\ldots,X_2) = H(X_{n-1}|X_{n-2},\ldots,X_1),$$

Therefore, $H(X_n|X_{n-1},\ldots,X_1)$ is a monotonically non-increasing sequence and lower bounded by 0. Hence, the sequence must converge $\lim_{n\to\infty} H(X_n|X_{n-1},\ldots,X_1) = H'(\mathcal{X})$. Recall that $\frac{1}{n}H(X_1,\ldots,X_n) = \frac{1}{n}\sum_{i=1}^{n} H(X_i|X_{i-1},\ldots,X_1)$. By the lemma of Cesáro mean, the RHS converges to $H'(\mathcal{X})$, and so does the LHS.

Lemma (Cesáro mean)

If a sequence $\{a_n\} \to c$, the running average $\{b_n := \frac{1}{n}\sum_{i=1}^{n} a_i\} \to c$.

Entropy Rate for Stationary Stochastic Processes (cont'd)

Lemma

For a stationary Markov chain, $H'(\mathcal{X}) = H(X_2|X_1)$.

Proof: $H'(\mathcal{X}) = \lim_{n\to\infty} H(X_n|X_{n-1}, X_{n-2}, \dots X_1) = \lim_{n\to\infty} H(X_n|X_{n-1}) = H(X_2|X_1)$.

Theorem

Let $\{X_i\}$ be a stationary Markov chain with stationary distribution $\boldsymbol{\mu}$ and transition probability matrix $\mathbf{P}$. If $X_1 \sim \boldsymbol{\mu}$, then the entropy rate $H(\mathcal{X}) = -\sum_{ij} \mu_i P_{ij} \log P_{ij}$

Proof: $H(\mathcal{X}) = H(X_2|X_1) = \sum_i \Pr(X_1 = i) H(X_2|X_1 = i) = \sum_i \mu_i \sum_j P_{ij} \log P_{ij}^{-1}$.

Note: For an irreducible, aperiodic, and finite MC, any initial distribution converges to the stationary distribution. So, we do not start from the stationary distribution, the entropy rate $H(\mathcal{X})$ given above is still correct.

The entropy rate of a stationary Markov chain is not dependent on its initial distribution, but only on the transitions between the states and the stationary distribution.

Entropy Rate of Random Walk over Graph

Example (Random walk over a weighted graph)

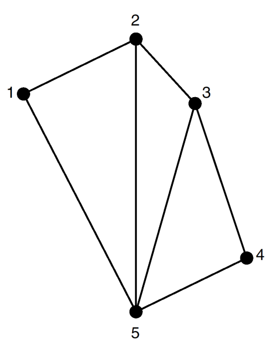


Figure: A weighted graph $G(\mathcal{N}, \mathcal{E}, \mathcal{W})$.

- $w_{ij} = w_{ji}$ denotes the edge weight between nodes i and j (0 if no edges).
- Given $X_n = i$, the probability of moving from node i to j is $P_{ij} = \frac{w_{ij}}{\sum_k w_{ik}} = \frac{w_{ij}}{w_i}$, where $w_i := \sum_k w_{ik}$ is the total weight of all edges connecting with node i.
- Intuitively, the stationary distribution of any node $i \in \mathcal{N}$ should be proportional to its degree w_i, which can be derived as $\pi_i = \frac{w_i}{2w}$, where $w \triangleq \sum_{i,j:j>i} w_{ij}$ is the total weight of all edges.
- Sanity check of the stationary distribution: $\sum_i \pi_i P_{ij} = \sum_i \frac{w_i}{2w} \frac{w_{ij}}{w_i} = \frac{w_j}{2w} = \pi_j, \ \forall j \in \mathcal{N}$.
- Locality property of this stationary distribution: it depends only on the total weight and the weight of edges connected to the node.

Entropy Rate of Random Walk over Graph (cont'd)

Example (cont.)

Hence, the entropy rate is

$$\begin{aligned}
H(\mathcal{X}) = H(X_2|X_1) &= -\sum_{ij} \mu_i P_{ij} \log P_{ij} && (2)\\
&= -\sum_{ij} \frac{w_{ij}}{2w} \log \frac{w_{ij}}{w_i} && (3)\\
&= -\sum_{ij} \frac{w_{ij}}{2w} \log \frac{w_{ij}}{2w} + \sum_{i} \frac{w_i}{2w} \log \frac{w_i}{2w} && (4)\\
&= H \underbrace{\left(\ldots, \frac{w_{ij}}{2w}, \ldots\right)}_{|\mathcal{N}|^2 \text{ terms}} - H \underbrace{\left(\ldots, \frac{w_i}{2w}, \ldots\right)}_{|\mathcal{N}| \text{ terms}} && (5)
\end{aligned}$$

If all the edges have equal weight, the stationary distribution becomes $\pi_i = \frac{D_i}{2D}$, where D_i of node i and D is the total degree of the graph. In this case, the entropy rate is

$$H(\mathcal{X}) = \log(2D) - H\left(\frac{D_1}{2D}, \ldots, \frac{D_{|\mathcal{N}|}}{2D}\right).$$

Function of Markov Chain

Theorem

Consider a stationary Markov chain $\{X_i\}$ and $Y_i = \phi(X_i)$ for all i. We have:

$$H(Y_n|Y_{n-1},\dots,Y_1,X_1) \le H(\mathcal{Y}) \le H(Y_n|Y_{n-1},\dots,Y_1)$$

$$\lim_{n\to\infty} H(Y_n|Y_{n-1},\dots,Y_1,X_1) = H(\mathcal{Y}) = \lim_{n\to\infty} H(Y_n|Y_{n-1},\dots,Y_1).$$

First, note that $\{X_i\}$ is a stationary MC $\implies$ $\{Y_i\}$ is stationary, but not necessarily a MC (unless ϕ is injective).

$$\begin{aligned}
\Pr\left(Y_{n+1}=y_{n+1}|\{Y_k=y_k\}_{k\le n}\right) &= \Pr\left(X_{n+1}=\phi^{-1}(y_{n+1})|\{X_k=\phi^{-1}(y_k)\}_{k\le n}\right) && (6)\\
&= \Pr\left(X_{n+1}=\phi^{-1}(y_{n+1})|X_n=\phi^{-1}(y_n)\right) && (7)\\
&= \Pr\left(Y_{n+1}=y_{n+1}|Y_n=y_n\right) && (8)
\end{aligned}$$

Function of Markov Chain (cont'd)

Proof: We have proved the upper bound. For the lower bounded,

$$\begin{aligned}
H(Y_n|Y_{n-1},\dots,Y_1,X_1) &= H(Y_n|Y_{n-1},\dots,Y_1,X_1,X_0,\dots,X_{-k}) && (9)\\
&= H(Y_n|Y_{n-1},\dots,Y_1,X_1,X_0,\dots,X_{-k},Y_0,\dots,Y_{-k}) && (10)\\
&\le H(Y_n|Y_{n-1},\dots,Y_1,Y_0,\dots,Y_{-k}) && (11)\\
&= H(Y_{n+k+1}|Y_{n+k},\dots,Y_1) && (12)
\end{aligned}$$

The inequality is true for all k, it is also true in the limit.

$$H(Y_n|Y_{n-1},\dots,Y_1,X_1) \le \lim_{k\to\infty} H(Y_{n+k+1}|Y_{n+k},\dots,Y_1) = H(\mathcal{Y}).$$

Function of Markov Chain (cont'd)

Next, we show that the upper and lower bounds converge to the same value:

$$\lim_{n\to\infty} H(Y_n|Y_{n-1},\ldots,Y_1,X_1) = \lim_{n\to\infty} H(Y_n|Y_{n-1},\ldots,Y_1),$$

which is equivalent to $\lim_{n\to\infty} I(X_1;Y_n|Y_{n-1},\ldots Y_1) = 0.$

Note that $\underbrace{I(X_1;Y_n,\ldots,Y_1)}_{\text{increases in } n} = H(X_1) - H(X_1|Y_n,\ldots,Y_1) \leq H(X_1)$. Hence, we have

$$H(X_1) \geq \lim_{n\to\infty} I(X_1;Y_n,Y_{n-1},\ldots Y_1) \tag{13}$$

$$= \lim_{n\to\infty} \sum_{i=1}^{n} I(X_1;Y_i|Y_{i-1},\ldots Y_1) \tag{14}$$

$$= \sum_{i=1}^{\infty} I(X_1;Y_i|Y_{i-1},\ldots Y_1) \tag{15}$$

The infinite sum of nonnegative terms is finite $\implies$ the terms must tend to 0.

Hidden Markov Model (HMM)[2]

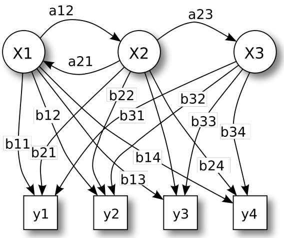


Figure: HMM diagram:
$X \rightarrow$ states;
$y \rightarrow$ possible observations;
$a \rightarrow$ state transition probabilities;
$b \rightarrow$ output (or emission) probabilities.

Given a Markov process $\{X_n\}$, each Y_i is drawn according to $p(y_i|x_i)$, conditionally independent of all the other X_j, $j \neq i$; i.e.,

$$\boxed{p(x^n, y^n) = p(x^n)p(y^n|x^n) = p(x_1)\prod_{i=1}^{n-1} p(x_{i+1}|x_i)\prod_{i=1}^{n} p(y_i|x_i)}$$

- Lower bound the entropy rate by conditioning it on the underlying Markov state.

[2] Wiki: HMM is widely used in many real applications such as speech recognition, handwriting recognition, musical score following, bioinformatics, etc.

Thank You!

Email: `<zhangy@ucsc.edu>`

Homepage: `https://people.ucsc.edu/~yzhan419/`