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

From DPI, we have $X \to Y \to Z \Longleftrightarrow p(x,y,z) = p(x)p(y|x)p(z|y)$.

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

$$P(X_{n+1} = x_{n+1} | X_n = x_n, \cdots, X_1 = x_1) = P(X_{n+1} = x_{n+1} | X_n = x_n) \tag{1}$$

$$p(x_{n+1}, \cdots, x_1) = p(x_1)p(x_2|x_1) \cdots p(x_{n+1}|x_n) \tag{2}$$

Time-invariant Markov chain: If $p(x_{n+1}|x_n)$ does not depend on n; i.e., $P(X_{n+1} = j | X_n = i) = P(X_2 = j | X_1 = i),\ \forall i, j \in \{1, 2, \ldots, m\}$, then we can include all transition probabilities in a matrix $\mathbf{P}_{m \times m}$. Its (i,j)-th entry is given as

$$\mathbf{P}_{ij} = P(X_{n+1} = j | X_n = i),\ \forall i, j \in \{1, 2, \cdots, m\} \tag{3}$$

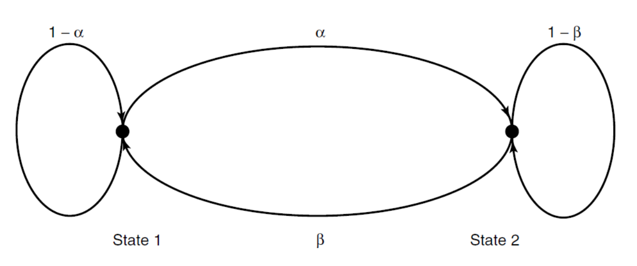
Two-state Markov chain

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

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

Stationary Distribution

Define the following notations:

- $\pi_j^{(n+1)} = P(X_{n+1} = j), \ \forall i, j \in \{1, 2, \ldots, m\}.$
- $\boldsymbol{\pi}^{(n+1)} = \left[\pi_1^{(n+1)}, \pi_2^{(n+1)}, \ldots, \pi_m^{(n+1)}\right].$
- $\boldsymbol{\pi}^{(n+1)} = \boldsymbol{\pi}^{(n)}\mathbf{P}.$

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.

Example

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

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

Irreducible and Aperiodic 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,

$$P(X_n = j | X_0 = i) = P_{ij}^{(n)} > 0, \ \forall i, j \tag{4}$$

Definition (Aperiodic)

The period of a state i is defined as $k = \gcd\{n > 0, P_{ii}^{(n)} > 0\}$. If $k = 1$, the state is said to be aperiodic. **A Markov chain is aperiodic if every state is aperiodic.** An irreducible Markov chain only needs one aperiodic state to imply all states are aperiodic.

Note that *gcd* is the greatest common divisor. For example: $\gcd\{6, 8, 10, 12, \cdots\} = 2$ and $\gcd\{3, 5, 7, \cdots\} = 1$.

Unique Stationary Distribution

Theorem

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

Lemma

For a finite-state Markov chain, if it is irreducible and aperiodic, then any initial distribution converges to the stationary distribution as $n \to \infty$.

Stationary Stochastic Process

Definition (Stationary stochastic process)

A stochastic process $\{X_i\}$ is (strong) *stationary* if the joint distribution of any subset of the sequence is invariant wrt 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 every Markov chain stationary? **A:** No.

Example

Consider a Markov chain with the transition probability matrix $\mathbf{P} = \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix}$ and the initial probability state $\boldsymbol{\pi}^{(0)} = (1,\, 0)$, then $\boldsymbol{\pi}^{(1)} = \boldsymbol{\pi}^{(0)} \cdot \mathbf{P} = (1/2,\, 1/2) \neq \boldsymbol{\pi}^{(0)}$ and $\mathrm{E}(X_1) \neq \mathrm{E}(X_0)$. Hence, it is not stationary.

Entropy Rate for Stochastic Processes

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

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

Definition (Per symbol entropy of n random variables)

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

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* exist; see an example in the textbook.

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

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

Entropy Rate for Stationary Stochastic Processes

Theorem (Entropy rate)

For a stationary stochastic process, $H(\mathcal{X}) = H'(\mathcal{X})$. That is, both limits exist and are equal.

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

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

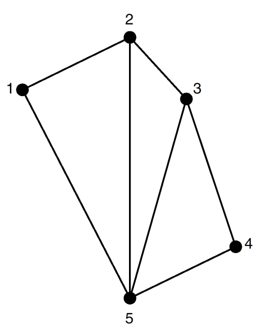
Proof: $H(\mathcal{X}) = H(X_2|X_1) = \sum_i P(X_1 = i) H(X_2|X_1 = i) = \sum_i \mu_i \sum_j P_{ij} \log P_{ij}^{-1}$.

Entropy Rate of Random Walk over graph

Example (Random walk over a weighted graph)

Consider an undirected and connected graph $G(\mathcal{N}, \mathcal{E}, \mathcal{W})$, where $w_{ij} = w_{ji}$ be the edge weight for the edge between nodes i and j (0 if no edge).

Given $X_n = i$, the probability of moving from node i to node j is given as $P_{ij} = \frac{w_{ij}}{\sum_k w_{ik}} = \frac{w_{ij}}{w_i}$, where $w_i := \sum_k w_{ik}$ is the degree of node i; i.e., the total weight of all edges connecting with node i.



Entropy Rate of Random Walk over graph (Cont'd)

Example (Cont.)

Intuitively, the stationary distribution of node i should be proportional to its degree, which can be finally derived as $\pi_i = \frac{w_i}{2w}$ $\left(w \triangleq \sum_{i,j:j>i} w_{ij}\right)$, $i \in \mathcal{N}$.
Sanity check: $\sum_i \pi_i P_{ij} = \sum_i \frac{w_i}{2w} \frac{w_{ij}}{w_i} = \frac{w_j}{2w} = \pi_j$

Hence, the entropy rate is

$$
\begin{aligned}
H(\mathcal{X}) = H(X_2|X_1) &= -\sum_{ij} \mu_i P_{ij} \log P_{ij} && (5)\\
&= -\sum_{ij} \frac{w_{ij}}{2w} \log \frac{w_{ij}}{w_i} && (6)\\
&= -\sum_{ij} \frac{w_{ij}}{2w} \log \left(\frac{w_{ij}}{2w} \times \frac{2w}{w_i} \right) && (7)\\
&= -\sum_{ij} \frac{w_{ij}}{2w} \log \frac{w_{ij}}{2w} + \sum_i \frac{w_i}{2w} \log \frac{w_i}{2w} && (8)\\
&= H \underbrace{\left(\cdots, \frac{w_{ij}}{2w}, \cdots \right)}_{|\mathcal{N}|^2 \text{ terms}} - H \underbrace{\left(\cdots, \frac{w_i}{2w}, \cdots \right)}_{|\mathcal{N}| \text{ terms}} && (9)
\end{aligned}
$$

Function of Markov Chain

Theorem

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

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

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

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

$$\begin{aligned}
\Pr\left(Y_{n+1}=y_{n+1}|\{Y_k=y_k\}_{k\leq n}\right) &= \Pr\left(X_{n+1}=\phi^{-1}(y_{n+1})|\{X_k=\phi^{-1}(y_k)\}_{k\leq n}\right) && (10)\\
&= \Pr\left(X_{n+1}=\phi^{-1}(y_{n+1})|X_n=\phi^{-1}(y_n)\right) && (11)\\
&= \Pr\left(Y_{n+1}=y_{n+1}|Y_n=y_n\right) && (12)
\end{aligned}$$

Function of Markov Chain (Cont'd)

Proof:

$$
\begin{aligned}
H(Y_n|Y_{n-1},\dots,Y_1,X_1) &= H(Y_n|Y_{n-1},\dots,Y_1,X_1,X_0,\dots,X_{-k}) && (13)\\
&= H(Y_n|Y_{n-1},\dots,Y_1,X_1,X_0,\dots,X_{-k},Y_0,\dots,Y_{-k}) && (14)\\
&\le H(Y_n|Y_{n-1},\dots,Y_1,Y_0,\dots,Y_{-k}) && (15)\\
&= H(Y_{n+k+1}|Y_{n+k},\dots,Y_1) && (16)
\end{aligned}
$$

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

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

Next, we show that $\lim_{n\to\infty} I(X_1;Y_n|Y_{n-1},\dots Y_1) = 0$

$$
\begin{aligned}
H(X_1) &\ge \lim_{n\to\infty} I(X_1;Y_n,Y_{n-1},\dots Y_1) && (17)\\
&= \lim_{n\to\infty} \sum_{i=1}^{n} I(X_1;Y_i|Y_{i-1},\dots Y_1) && (18)\\
&= \sum_{i=1}^{\infty} I(X_1;Y_i|Y_{i-1},\dots Y_1) && (19)
\end{aligned}
$$

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

Hidden Markov Model (HMM)

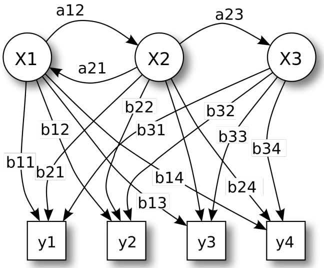


Figure: HMM Diagram (from wiki): X — states; y — possible observations; a — state transition probabilities; b — output (or emission) probabilities; Widely used in many real applications such as speech recognition, handwriting recognition, musical score following, bioinformatics, etc.

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.,

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

Thank You!

Email: <zhangy@ucsc.edu>

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