CS258: Information Theory

Fan Cheng
Shanghai Jiao Tong University

http://www.cs.sjtu.edu.cn/~chengfan/
chengfan@sjtu.edu.cn
Spring, 2020

Outline

- Stochastic Process
- Entropy Rate
- Second Law of Thermodynamics
- Functions of Markov Chains

Stochastic Process

AEP establishes that $nH(X)$ bits suffice on the average to describe n independent and identically distributed random variables. **But what if the random variables $\{X_i\}$ are dependent?**

--entropy rate

- A **stochastic process** $\{X_i\}$ is an indexed sequence of random variables.
- Gambler's Ruin (赌徒的破产)
  - Consider a gambler who starts with an initial fortune of 1 and then on each successive gamble either wins 1 or loses 1 independent of the past with probabilities p and $q = 1 - p$, respectively.
  - The stops playing after getting ruined.
- Let $\{X_i\}$ represent the outcome of the game, then X_{i+1} depends on X_i:

$$X_{i+1} = X_i \pm 1$$

Thus X_i's are not i.i.d.

Stationary Process

A stochastic process is said to be **stationary (穩态)** if the joint distribution of **any subset** of the sequence of random variables is invariant with respect to shifts in the time index; 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 and every shift l and for all $x_1, x_2, \ldots, x_n \in \mathcal{X}$.

- **Shift invariant**
  - $p(X_1) = p(X_2) = \cdots = p(X_n)$
  - $p(X_1, X_3) = p(X_2, X_4)$ ...
- Gaussian process (GP) is stationary
- Stationary Markov Chain
- Strong stationary Vs. Weak stationary: No implication

Time's arrow. Let $\{X_i\}_{i=-\infty}^{\infty}$ be a stationary stochastic process. Prove that

$$H(X_0|X_{-1}, X_{-2}, \ldots, X_{-n}) = H(X_0|X_1, X_2, \ldots, X_n)$$

- $H(X_{-n}, \ldots, X_0) = H(X_0, \ldots, X_n)$
- $H(X_{-n}, \ldots, X_{-1}) = H(X_1, \ldots, X_n)$

Markov Chain

- A discrete stochastic process $X_1, X_2, \ldots$ is said to be a **Markov chain** or a **Markov process** if for $n = 1, 2, \ldots$

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

  for all $x_1, x_2, \ldots, x_n, x_{n+1} \in \mathcal{X}$.
- The **joint distribution** is

$$p(x_1, x_2, \ldots, x_n) = p(x_1)p(x_2|x_1)p(x_3|x_2) \cdots p(x_n|x_{n-1}).$$

- The Markov chain is said to be **time invariant** if the conditional probability $p(x_{n+1}|x_n)$ does not depend on n; that is, for $n = 1, 2, \ldots,$

  $\Pr\{X_{n+1} = b | X\_n = a\} = \Pr\{X_2 = b | X_1 = a\}$ for all $a, b \in \mathcal{X}$.

  We will assume that the Markov chain is time invariant unless otherwise stated
- A time-invariant Markov chain is characterized by its initial state and a probability transition matrix $\boldsymbol{P} = [\boldsymbol{P_{ij}}], \boldsymbol{i}, \boldsymbol{j} \in \{\mathbf{1}, \mathbf{2}, \ldots, \boldsymbol{m}\}$, where

$$P_{ij} = \Pr\{X_{n+1} = j | X_n = i\}$$

Example

- Gambler's ruin
- Random walk

Stationary Distribution of MC

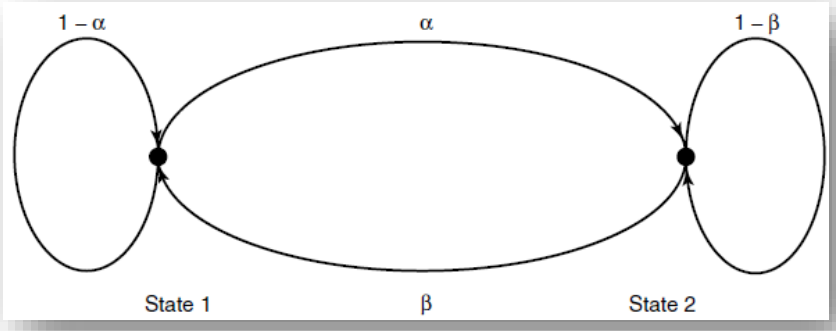
- By the definition of stationary, a Markov chain is **stationary iff** $\boldsymbol{p(X_{n+1}) = p(X_n)}$
- If the probability mass function at time n is $p(x_n)$, then

$$\boldsymbol{p(x_{n+1}) = \sum_{x_n} p(x_n) P_{x_n x_{n+1}} \quad \text{or } x^T P = x^T}$$

- If the initial state of a Markov chain is drawn according to a stationary distribution, the Markov chain is stationary

Consider a two-state Markov chain with a probability transition matrix

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



- $(\mu_1, \mu_2) \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix} = (\mu_1, \mu_2)$
- For stationary distribution, the net probability flow across any cut set in zero

$$\mu_1 \alpha = \mu_2 \beta$$

- $\mu_1 + \mu_2 = 1$

$$\mu_1 = \frac{\beta}{\alpha + \beta} \text{ and } \mu_2 = \frac{\alpha}{\alpha + \beta}$$

Entropy Rate

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

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

when the limits exists

- **Average entropy**
- **How to evaluate**

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

For $H(X_i|X_{i-1}, \dots, X_1)$, we now need to make clear of

- The existence of

$$\lim_{n\to\infty} H(X_n|X_{n-1}, \dots, X_1)$$

- In a series $\{a_n\}$, if $a_n \to a$, the existence of

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} a_i$$

$H'(\mathcal{X})$

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

- $$\begin{aligned} & H(X_{n+1}|X_n,\ldots,X_1) \\ \leq\ & H(X_{n+1}|X_n,\ldots,X_2) \\ =\ & \boldsymbol{H(X_n|X_{n-1},\ldots,X_1)} \end{aligned}$$
- $$H(X_n|X_{n-1},\ldots,X_1) \geq 0$$
- Since $\{H(X_n|X_{n-1},\ldots,X_1)\}$ is nonincreasing and $H(X_n|X_{n-1},\ldots,X_1) \geq 0$, the limit exists.

- Define
$$H'(\mathcal{X}) = \lim_{n\to\infty} H(X_n|X_{n-1},X_{n-2},\ldots,X_1)$$
- (Theorem.) The limits $H'(\mathcal{X})$ exists

Cesaro Mean

If $a_n \to a$ and $b_n = \frac{1}{n}\sum_{i=1}^{n} a_i$, then $b_n \to a$.

Let $\epsilon > 0$. Since $a_n \to a$, there exists a number $N(\epsilon)$ such that $|a_n - a| \leq \epsilon$ for all $n \geq N(\epsilon)$. Hence,

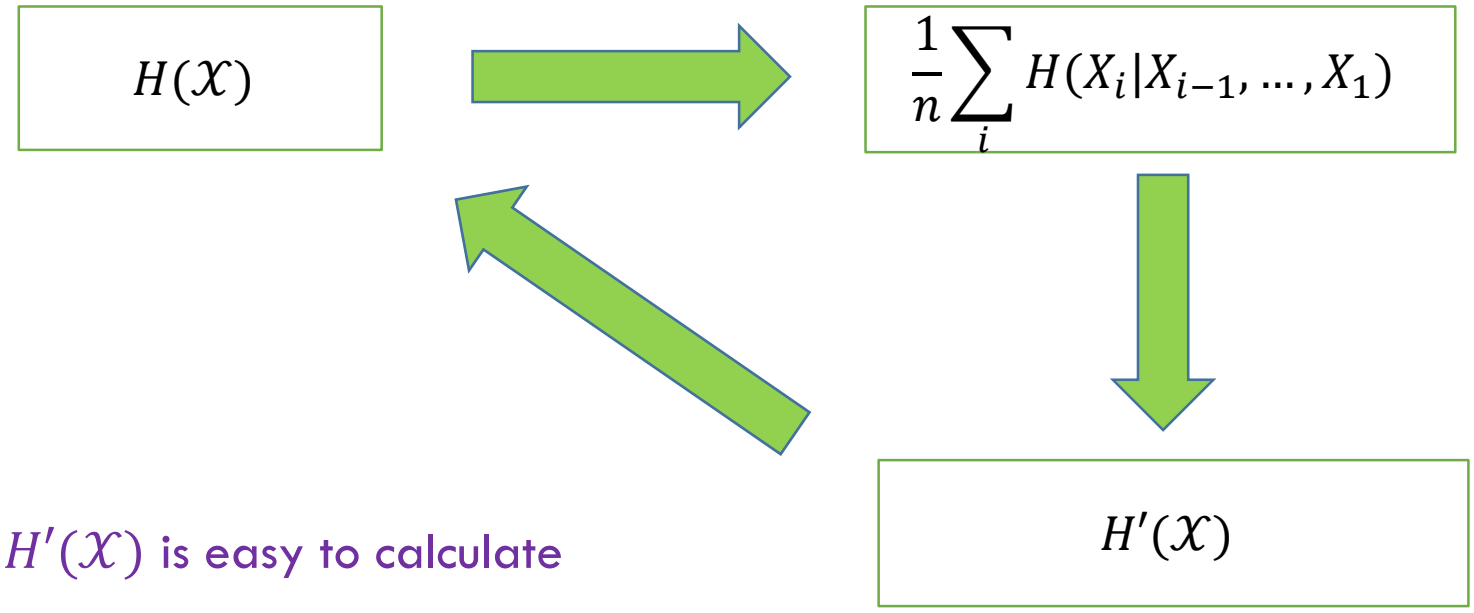
$$
\begin{aligned}
|b_n - a| &= \left| \frac{1}{n} \sum_{i=1}^{n} (a_i - a) \right| \\
&\leq \frac{1}{n} \sum_{i=1}^{n} |(a_i - a)| \\
&\leq \frac{1}{n} \sum_{i=1}^{N(\epsilon)} |a_i - a| + \frac{n - N(\epsilon)}{n} \epsilon \\
&\leq \frac{1}{n} \sum_{i=1}^{N(\epsilon)} |a_i - a| + \epsilon
\end{aligned}
$$

Thus, $|b_n - a| \leq \epsilon'$, for all $n \geq N(\epsilon)$

Entropy Rate (Cont'd)

(Theorem.) For a stationary stochastic process, the limits in $H(\mathcal{X})$ and $H'(\mathcal{X})$ exist and are equal:

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



$H'(\mathcal{X})$ is easy to calculate

Entropy Rate: Markov Chain

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

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

where the conditional entropy is calculated using the given stationary distribution.

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

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

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

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

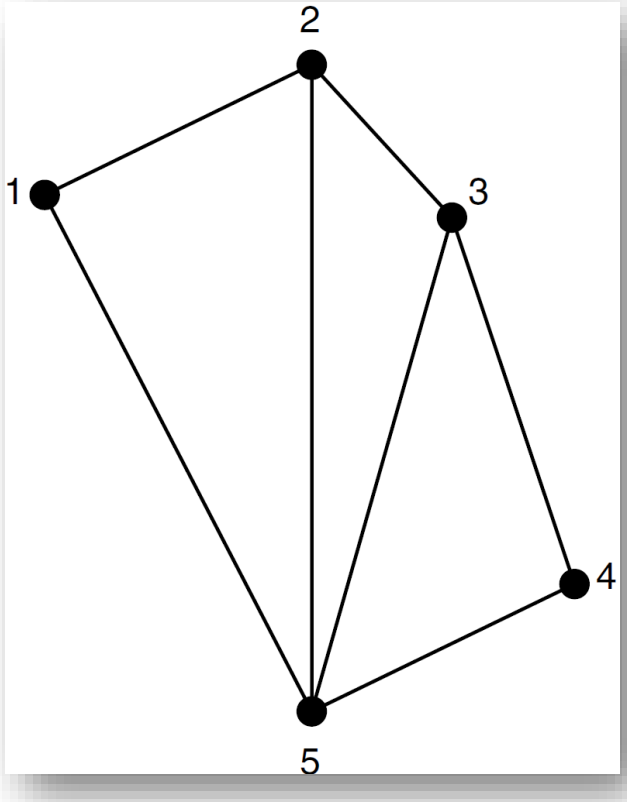
Proof:

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

- The entropy rate of the two-state Markov chain is

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

Entropy Rate: Random Walk



Undirected graph with weight $W_{ij} \geq 0$ and $W_{ij} = W_{ji}$.

$$P_{ij} = W_{ij} / \sum_k W_{ik}$$

$$W_i = \sum_j W_{ij}$$

$$W = \sum_i \frac{W_i}{2}$$

The stationary distribution is

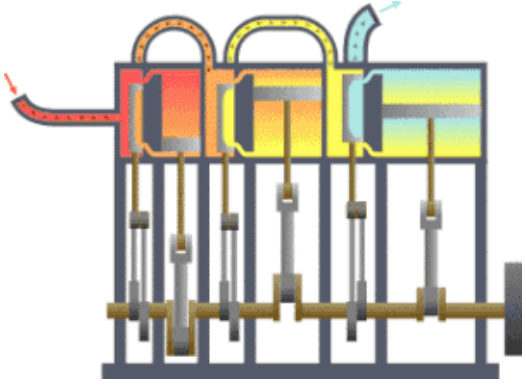
$$\mu_i = \frac{W_i}{2W}$$

Verify it by $\mu P = \mu$.

$$H(\mathcal{X}) = H(X_2|X_1)$$

$$= H\left(\ldots, \frac{W_{ij}}{2W}, \ldots\right) - H(\ldots, \frac{W_i}{2W}, \ldots)$$

Second Law of Thermodynamics



- One of the basic laws of physics, the second law of thermodynamics, states that the entropy of an isolated system is nondecreasing.
- We model the isolated system as a **Markov chain with transitions obeying the physical laws governing the system**.
  - Implicit in this assumption is the notion of an overall state of the system and the fact that knowing the present state, the future of the system is independent of the past.

Some results:

- Relative entropy $\mathrm{D}(\mu_n||\mu'_n)$ decreases with n
- The conditional entropy $H(X_n|X_1)$ increases with n for a stationary Markov process
- Shuffles increase entropy: $H(TX) \geq H(X)$

Reference: Neri Merhav (2010), "Statistical Physics and Information Theory," Foundations and Trends® in Communications and Information Theory

Functions of Markov Chain

$$\begin{array}{cccccc} \boldsymbol{X_1} & & \boldsymbol{X_2} & \cdots & \boldsymbol{X_n} & \cdots \\ \downarrow & & \downarrow & \cdots & \downarrow & \cdots \\ \boldsymbol{Y_1 = \phi(X_1)} & & \boldsymbol{Y_2 = \phi(X_2)} & \cdots & \boldsymbol{Y_n = \phi(X_n)} & \cdots \end{array}$$

Let $X_1, X_2, \ldots, X_n, \ldots$ be a stationary Markov chain, and let $Y_i = \phi(X_i)$ be a process each term of which is a function of the corresponding state in the Markov chain. What is the entropy rate of $H(\mathcal{Y})$?

- $\{Y_i\}$: A very special case of hidden Markov model (HMM)
- $\{Y_i\}$ is not a Markov chain in general
- $\{X_i\}$ is stationary ⇒ $\{Y_i\}$ is stationary

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

- Drawback: Hard to ensure the convergence by n
- Solution: We have already known that $\boldsymbol{H(Y_n|Y_{n-1}, \ldots, Y_1)}$ is lower bounded by $H(\mathcal{Y})$
  - Find a lower bound for $H(\mathcal{Y})$ which is close to $H(Y_n|Y_{n-1}, \ldots, Y_1)$
- Let's have a look at X_1
  - X_1 contains much information about Y_n as $Y_1, Y_0, Y_{-1}, \ldots \ldots$

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

(Y_1 could be ignored)

Functions of Markov Chain

(**Theorem**). If $X_1, X_2, \ldots, X_n$ form a stationary Markov chain, and $Y_i = \phi(X_i)$, then

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

and

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

$$\begin{aligned}
&H(Y_n|Y_{n-1},\ldots,Y_2,X_1)\\
&= H(Y_n|Y_{n-1},Y_2,Y_1,X_1)\\
&= H(Y_n|Y_{n-1},\ldots,Y_1,X_1,X_0,X_{-1},\ldots,X_{-k})\\
&= H(Y_n|Y_{n-1},\ldots,Y_1,X_1,X_0,X_{-1},\ldots,X_{-k},Y_0,\ldots,Y_{-k})\\
&\le H(Y_n|Y_{n-1},\ldots,Y_1,Y_0,\ldots,Y_{-k})\\
&= H(Y_{n+k+1}|Y_{n+k},\ldots,Y_1)
\end{aligned}$$

$k \to \infty$,

$$H(Y_n|Y_{n-1},\ldots,Y_2,X_1) \le H(\mathcal{Y})$$

$$\begin{aligned}
&H(Y_n|Y_{n-1},\ldots,Y_1) - H(Y_n|Y_{n-1},\ldots,Y_1,X_1)\\
&\quad = I(X_1;Y_n|Y_{n-1},\ldots,Y_1)\\
&I(X_1;Y_1,Y_2,\ldots,Y_n) \le H(X_1)\\
&H(X_1) \ge \lim_{n\to\infty} I(X_1;Y_1,Y_2,\ldots,Y_n)\\
&\quad = \lim_{n\to\infty} \sum_{i=1}^{n} I(X_1;Y_i|Y_{i-1},\ldots,Y_1)\\
&\quad = \sum_{i=1}^{\infty} I(X_1;Y_i|Y_{i-1},\ldots,Y_1)
\end{aligned}$$

$$I(X_1;Y_n|Y_{n-1},\ldots,Y_2,Y_1) \to 0$$

$$\| $$

$$H(Y_n|Y_{n-1},\ldots,Y_2,Y_1) - H(Y_n|Y_{n-1},\ldots,Y_1,X_1)$$

Problems

Homework
Cover: 4.1, 4.3, 4.7, 4.21

- Monotonicity of entropy per element. For a stationary stochastic process $X_1, X_2, \ldots, X_n$, show that

$$\frac{H(X_1, X_2, \ldots, X_n)}{n} \leq \frac{H(X_1, X_2, \ldots, X_{n-1})}{n-1}$$
$$\frac{H(X_1, X_2, \ldots, X_n)}{n} \geq H(X_n | X_{n-1}, \ldots, X_1)$$

- Initial conditions. Show, for a Markov chain, that

$$H(X_0|X_n) \geq H(X_0|X_{n-1}).$$

Thus, initial conditions X_0 become more difficult to recover as the future X_n unfolds.

- The past has little to say about the future. For a stationary stochastic process $X_1, X_2, \ldots, X_n, \ldots$, show that

$$\lim_{n\to\infty} \frac{1}{2n} I(X_1, X_2, \ldots, X_n; X_{n+1}, X_{n+2}, \ldots, X_{2n}) = 0$$

- Entropy rate. Let $\{X_i\}$ be a discrete stationary stochastic process with entropy rate $H(\mathcal{X})$. Show that

$$\frac{1}{n} H(X_n, \ldots, X_1 | X_0, X_{-1}, \ldots, X_{-k}) \to H(\mathcal{X})$$

for $k = 1, 2, \ldots$

Summary

T. Cover: Ch. 4