

INFORMATION THEORY & CODING

Entropy Rate

Dr. Rui Wang

Department of Electrical and Electronic Engineering
Southern Univ. of Science and Technology (SUSTech)

Email: wang.r@sustech.edu.cn

November 14, 2023



- On average, $nH(X) + 1$ bits suffices to describe n i.i.d. random variables. But what if the random variables are dependent?
- Markov Chain: a simplest way to model the correlations among random variables in a stochastic process.
- Entropy Rate: average number of bits suffices to describe one random variable in a stochastic process.

How to Model Dependence: Markov Chains

- A **stochastic process** $\{X_i\}$ is an indexed sequence of random variables (X_1, X_2, \dots) characterized by the joint PMF $p(x_1, x_2, \dots, x_n)$, where $(x_1, x_2, \dots, x_n) \in \mathcal{X}^n$ for $n = 0, 1, \dots$

Definition

A **stochastic process** is said to be **stationary** if the joint distribution of any subset of the sequence of random variables is **invariant** with respect to shifts in the time index, i.e.,

$$\begin{aligned}\Pr[X_1 = x_1, X_2 = x_2, \dots, X_n = x_n] \\ = \Pr[X_{1+\ell} = x_1, X_{2+\ell} = x_2, \dots, X_{n+\ell} = x_n]\end{aligned}$$

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

Markov Chains

Definition

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

$$\begin{aligned}\Pr[X_{n+1} = x_{n+1} | X_n = x_n, X_{n-1} = x_{n-1}, \dots, X_1 = x_1] \\ = \Pr[X_{n+1} = x_{n+1} | X_n = x_n]\end{aligned}$$

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

In this case, the joint PMF can be written as

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

Hence, a Markov chain is completely characterized by **initial distribution** $p(x_1)$ and **transition probabilities** $p(x_n|x_{n-1})$, $n = 2, 3, 4, \dots$

Definition

The Markov chain is called **time invariant** if the transition probability $p(x_{n+1}|x_n)$ does **NOT** depend on n , i.e., for $n = 1, 2, \dots$,

$$\Pr[X_{n+1} = b | X_n = a] = \Pr[X_2 = b | X_1 = a], \quad \forall a, b \in \mathcal{X}.$$

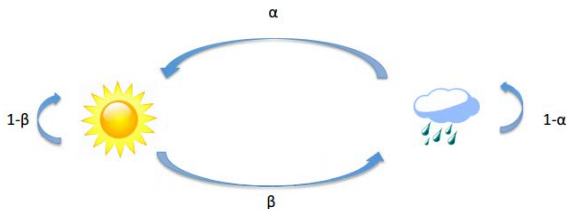
We deal with time invariant Markov chains. If $\{X_i\}$ is a Markov chain, X_n is called the **state** at time n . A time invariant Markov chain is characterized by its initial state and a **probability transition matrix** $P = [P_{ij}]$, $i, j \in \{1, 2, \dots, m\}$, where $P_{ij} = \Pr[X_{n+1} = j | X_n = i]$.

Markov Chain Example: Simple Weather Model

- $\mathcal{X} = \{\text{Sunny: } S, \text{ Rainy: } R\}$

$$p(S|S) = 1 - \beta, p(R|R) = 1 - \alpha, p(R|S) = \beta, p(S|R) = \alpha$$

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

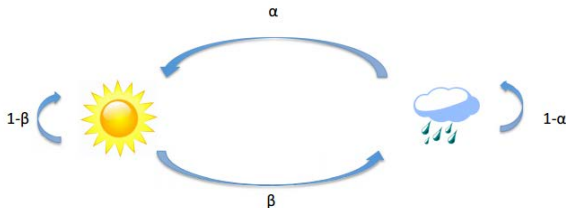


Markov Chain Example: Simple Weather Model

- Probability of seeing a sequence SSRR:

$$p(SSRR) = p(S)p(S|S)p(R|S)p(R|R) = p(S)(1 - \beta)\beta(1 - \alpha)$$

Suppose the first day is "Sunny" with probability γ , what is the weather distribution of the second day, third day, ...?

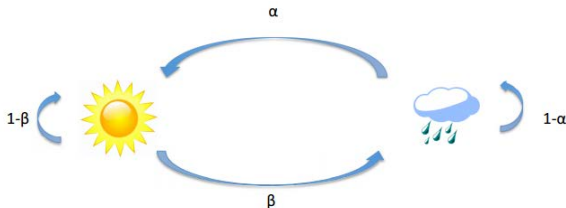


Stationary Distribution

- If $\mu = [\mu_S, \mu_R] = \left[\frac{\alpha}{\alpha+\beta}, \frac{\beta}{\alpha+\beta} \right]$

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

$$\begin{aligned} p(X_{n+1} = S) &= p(S|S)\mu_S + p(S|R)\mu_R \\ &= (1-\beta)\frac{\alpha}{\alpha+\beta} + \alpha\frac{\beta}{\alpha+\beta} = \frac{\alpha}{\alpha+\beta} = \mu_S. \end{aligned}$$



Stationary Distribution

- If the PMF of the random variable at time n is $\mu_i^n = \Pr[X_n = i]$, the PMF at time $n + 1$, say $\mu_j^{n+1} = \Pr[X_{n+1} = j]$, can be written as

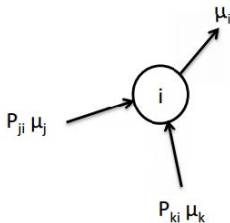
$$\mu_j^{n+1} = \sum_i \mu_i^n \Pr[X_{n+1} = j | X_n = i] = \sum_i \mu_i^n P_{ij}.$$

- $\{\mu_i^n | \forall i\}$ is called a **stationary distribution** if $\mu_i^n = \mu_i^{n+1}, \forall i$.
- For notation convenience, let $\mu_i = \mu_i^n = \mu_i^{n+1}, \forall i$.

Stationary Distribution

- How to calculate stationary distribution?
 - Stationary distribution $\mu_i, i = 1, 2, \dots, |\mathcal{X}|$ satisfies

$$\mu_j = \sum_{i=1}^{|\mathcal{X}|} \mu_i P_{ij} \quad \text{and} \quad \sum_{i=1}^{|\mathcal{X}|} \mu_i = 1.$$



Entropy Rate

- When X_i 's are i.i.d., the entropy

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

- With **dependent** sequences X_i 's, how does $H(X^n)$ grow with n ?
- Entropy rate** characterized the growth rate.

- **Definition 1:** average entropy per symbol

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

- **Definition 2:** conditional entropy of the last r.v. given the past

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

Entropy Rate

Theorem 4.2.2

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

Proof.

$$\begin{aligned} H(X_{n+1}|X_1, X_2, \dots, X_n) &\stackrel{\text{conditional reduces entropy}}{\leq} H(X_{n+1}|X_n, \dots, X_2) \\ &\stackrel{\text{stationary}}{=} H(X_n|X_{n-1}, \dots, X_1), \end{aligned}$$

- $H(X_n|X_{n-1}, \dots, X_1)$ decreases as n increases
- $H(X) \geq 0$
- The limit must exist. □

Theorem 4.2.1

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

Proof.

By the chain rule,

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

- $H(X_n | X_{n-1}, \dots, X_1) \rightarrow H'(\mathcal{X})$
- *Cesaro mean*: If $a_n \rightarrow a$, $b_n = \frac{1}{n} \sum_{i=1}^n a_i$, then $b_n \rightarrow a$.
- So

$$\frac{1}{n}H(X_1, \dots, X_n) \rightarrow H'(\mathcal{X})$$



AEP for Stationary Ergodic Process (chap 16)

$$-\frac{1}{n} \log p(X_1, \dots, X_n) \rightarrow H(\mathcal{X})$$

- $p(X_1, \dots, X_n) \approx 2^{-nH(\mathcal{X})}$
- Typical sequences in typical set of size $2^{-nH(\mathcal{X})}$
- We can use $nH(\mathcal{X})$ bits to represent typical sequences

Entropy Rate for Markov Chain

- For a stationary Markov chain, the entropy rate is

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

- Let $P_{ij} = \Pr[X_2 = j | X_1 = i]$. By definition, entropy rate of stationary Markov chain

$$\begin{aligned} H(\mathcal{X}) &= H(X_2 | X_1) = \sum_i \mu_i \left(\sum_j -P_{ij} \log P_{ij} \right) \\ &= - \sum_{ij} \mu_i P_{ij} \log P_{ij} \end{aligned}$$

To Calculate Entropy Rate

- 1 Find *stationary distribution* μ_i

$$\mu_i = \sum_j \mu_j p_{ji} \text{ and } \sum_{i=1}^{|\mathcal{X}|} \mu_i = 1$$

- 2 User *transition probability* P_{ij}

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

Entropy Rate of Weather Model

- Stationary distribution $\mu(S) = \frac{\alpha}{\alpha+\beta}$, $\mu(R) = \frac{\beta}{\alpha+\beta}$

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

$$\begin{aligned} H(\mathcal{X}) &= \mu(S)H(\beta) + \mu(R)H(\alpha) \\ &= \frac{\alpha}{\alpha+\beta}H(\beta) + \frac{\beta}{\alpha+\beta}H(\alpha) \\ &\stackrel{\text{Jensen's inequality}}{\leq} H\left(2\frac{\alpha\beta}{\alpha+\beta}\right) \end{aligned}$$

Maximum when $\alpha = \beta = 1/2$: degenerate to independent process



Related Sections : Whole Chapter 4