# 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 15, 2022

- **McMillan inequality**

  Uniquely decodable codes $\Leftrightarrow \sum D^{-\ell_i} \leq 1$.

- **Huffman code**

$$L^* = \min_{\sum D^{-\ell_i} \leq 1} \sum p_i \ell_i$$

$$H_D(X) \leq L^* < H_D(X) + 1.$$

# Outline

- 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.

随机.

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

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

### Definition

A stochastic process is said to be 平稳的.. 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.,

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

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

# Markov Chains

## Definition

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

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

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

In this case, the joint PMF can be written as

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

Hence, a Markov chain is completely characterized by initial distribution $p(x_1)$ and transition probabilities $\underbrace{p(x_n|x_{n-1})}_{=\Pr[X_n = x_n | X_{n-1} = x_{n-1}]}$, $n = 2, 3, 4, \ldots$

$$\Pr = [X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n]$$

$$= p(x_1, x_2 \cdots x_n) = p(x_2, x_3 \cdots x_n | x_1) p(x_1)$$

$$= p(x_3 x_4 \cdots x_n | x_2) p(x_1) p(x_2 | x_1)$$

$$= p(x_4 \cdots x_n | x_3) p(x_1) p(x_2 | x_1) p(x_3 | x_2)$$

$$= p(x_1) \prod_{i=1}^{n-1} p(x_{i+1} | x_i) \rightarrow \text{transition prob.}$$

## Markov Chains

Time Invariant Markov chain isn't stationary

Stationary Markov chain is time invariant

Proof: $Pr[X_1=i, X_2=j] = Pr[X_1=i]\, P_{ij}$

$Pr[X_n=i, X_{n+1}=j] = Pr[X_n=i]\, P_{ij}$

### 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 $n = 1, 2, \ldots,$

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

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 $P = [P_{ij}]$, $i, j \in \{1, 2, \ldots, m\}$, where $P_{ij} = \Pr[X_{n+1} = j|X_n = i]$.

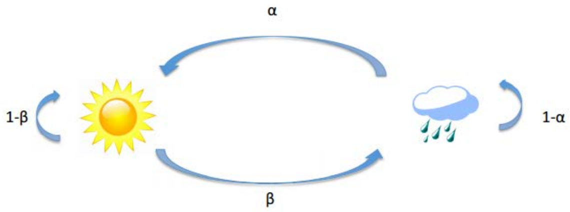
$$P = \begin{bmatrix} P_{11} & P_{12} & \cdots & P_{1n} \\ P_{21} & P_{22} & \cdots & P_{2n} \\ \vdots & \vdots & & \vdots \\ P_{n1} & P_{n2} & \cdots & P_{nn} \end{bmatrix}$$

# Markov Chain Example: Simple Weather Model

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

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

$$P = \begin{array}{c} \begin{array}{cc} p(S|S) & p(R|S) \end{array} \\ \begin{bmatrix} 1-\beta & \beta \\ \alpha & 1-\alpha \end{bmatrix} \\ \begin{array}{cc} p(S|R) & p(R|R) \end{array} \end{array}$$

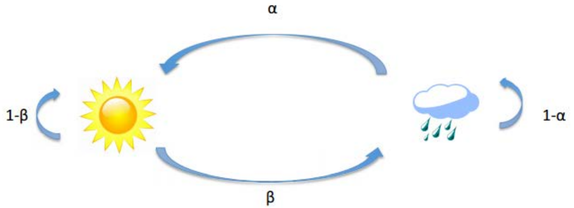


# Markov Chain Example: Simple Weather Model

- 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 $\gamma$, what is the weather distribution of the second day, third day, ...?




$\Pr[X_1=S]=a$ . $\Pr[X_1=R]=b$

$\Pr[X_2=S]=\Pr[X_2=S, X_1=S]+\Pr[X_2=S, X_1=R]$

$=\Pr[X_1=R]P_{RS}+\Pr[X_1=S]P_{SS}$

$= bP_{RS}+aP_{SS} = (a\ b)\begin{pmatrix}P_{SS}\\ P_{RS}\end{pmatrix}$

$\Pr[X_2=R] = bP_{RR}+aP_{SR} = (a\ b)\begin{pmatrix}P_{SR}\\ P_{RR}\end{pmatrix}$

$(\Pr[X_2=S]\ \ \Pr[X_2=R]) = (a\ b)\begin{pmatrix}P_{SS} & P_{SR}\\ P_{RS} & P_{RR}\end{pmatrix}$

$=(a\ b)P$

$(\Pr[X_n=S]\ \ \Pr[X_n=R]) = (a\ b)P^{n-1}$

when $(a\ b)=(a\ b)P$,

Time invariant / Markov Chain is stationary.

$$\left\{\begin{array}{l} a = a(1-\beta)+b\alpha \Rightarrow a\beta=b\alpha \\ b = a\beta + b(1-\alpha) \Rightarrow a\beta = b\alpha \\ a+b=1 \end{array}\right\} \rightarrow \left\{\begin{array}{l} a=\frac{\alpha}{\alpha+\beta} \\ b=\frac{\beta}{\alpha+\beta} \end{array}\right.$$

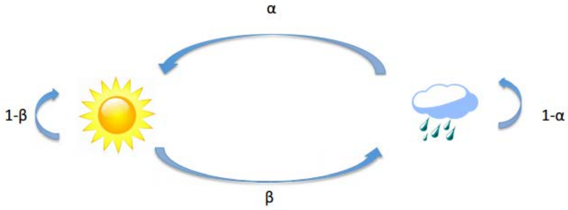
Stationary distribution.

# Stationary Distribution

- If $\mu = [\mu_S, \mu_R] = \begin{bmatrix} \frac{\alpha}{\alpha+\beta} & \frac{\beta}{\alpha+\beta} \end{bmatrix}$

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

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



## 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 $\mu_i = \mu_i^n = \mu_i^{n+1}$, $\forall i$.

General Case:

$X_1, X_2, \cdots X_n. \rightarrow$ n random variable

$\mathcal{X} = \{1, 2, \cdots, m\} \rightarrow$ m case.

$\{\mu_1, \mu_2, \cdots \mu_m\} \rightarrow$ probility of cases.

$$(\mu_1, \mu_2, \cdots, \mu_m) \begin{bmatrix} P_{11} & P_{12} & \cdots & P_{1m} \\ P_{21} & P_{22} & \cdots & P_{2m} \\ P_{31} & P_{32} & \cdots & P_{3m} \\ \vdots & & & \\ P_{m1} & P_{m2} & \cdots & P_{mm} \end{bmatrix} = (\mu_1, \mu_2, \cdots \mu_m).$$

$1 \times m$.

$\vec{\mu} P = \vec{\mu} \Rightarrow \vec{\mu}(P - I) = 0 \rightarrow$ (m个未知数, m-1个独立方程).

$\mathrm{Rank}\{P - I\} = m-1$ $\quad \vec{\mu}$ 与 $P-I$ 正交.

$\rightarrow \vec{\mu}\vec{e} = 1$ $\left[\vec{e} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}\right]$

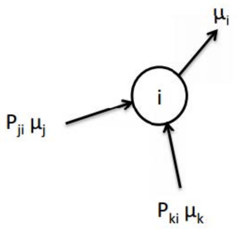
$\rightarrow (\vec{\mu}(P-I) \quad \vec{\mu}\vec{e}) = (\vec{0} \quad 1).$ $\rightarrow 1 \times (m+1)$

$\rightarrow \vec{\mu} \underbrace{[P - I \quad \vec{e}]}_{\tilde{P}} = (\underbrace{0 \cdots 0}_{m} \quad 1)$

$\rightarrow \vec{\mu}\tilde{P} = (\underbrace{0 \cdots 0}_{m} \quad 1)$

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

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




等式两边右乘 $\tilde{P}^T(\underbrace{\tilde{P}\tilde{P}^T}_{m\times m})^{-1}$

$\vec{\mu} = (0 \cdots 1)\hat{P}^T(\hat{P}\hat{P}^T)^{-1}$

# Entropy Rate

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

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

- With dependent sequences $X_i$'s, how does $H(X^n)$ grow with $n$?

- Entropy rate characterized the growth rate.

## Entropy Rate

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

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

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

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

Proof: △If $a_n \xrightarrow{n\to\infty} a$, $b_n = \frac{1}{n}\sum_{i=1}^{n} a_i$, then $b_n \xrightarrow{n\to\infty} a$.

$\underbrace{\frac{1}{n} H(X_1, \cdots, X_n)}_{b_n} = \frac{1}{n}\sum_{i=1}^{n} \underbrace{H(X_i|X_1 \cdots X_{i-1})}_{a_i}$.

$X_1\ X_2 \cdots X_n \cdots$ S.M.C.

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

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

$= H(X_2|X_1)$

平稳随机过程

### Theorem 4.2.2

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

### Proof.

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

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

# Entropy Rate

**Theorem 4.2.1**

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

**Proof.**

By the chain rule,

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

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

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

□

# AEP for Stationary Ergodic Process (chap 16)

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

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

# Entropy Rate for Markov Chain

- For a stationary Markov chain, the entropy rate is

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

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

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

## To Calculate Entropy Rate

1. Find *stationary distribution* $\mu_i$

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

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

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

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

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

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

# Textbook

Related Sections : Whole Chapter 4