Lecture 6: Entropy Rate

- Entropy rate $H(\mathcal{X})$
- Random walk on graph

Coin tossing versus poker

- Toss a fair coin and see and sequence

Head, Tail, Tail, Head $\cdots$

$$(x_1, x_2, \ldots, x_n) \approx 2^{-nH(X)}$$

- Play card games with friend and see a sequence

A ♣ | K ♥ | Q ♦ | J ♠ | 10 ♣ | $\cdots$

$$(x_1, x_2, \ldots, x_n) \approx ?$$

How to model dependence: Markov chain

- A stochastic process $X_1, X_2, \cdots$
  - State $\{X_1, \ldots, X_n\}$, each state $X_i \in \mathcal{X}$
  - Next step only depends on the previous state

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

  - Transition probability

$$p_{i,j}: \text{ the transition probability of } i \rightarrow j$$

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

Hidden Markov model (HMM)

- Used extensively in speech recognition, handwriting recognition, machine learning.
- Markov process $X_1, X_2, \ldots, X_n$, unobservable
- Observe a random process $Y_1, Y_2, \ldots, Y_n$, such that

$$Y_i \sim p(y_i|x_i)$$

- We can build a probability model

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

Time invariance Markov chain

- A Markov chain is time invariant if the conditional probability $p(x_n|x_{n-1})$ does not depend on n

$$p(X_{n+1} = b|X_n = a) = p(X_2 = b|X_1 = a), \text{ for all } a, b \in \mathcal{X}$$

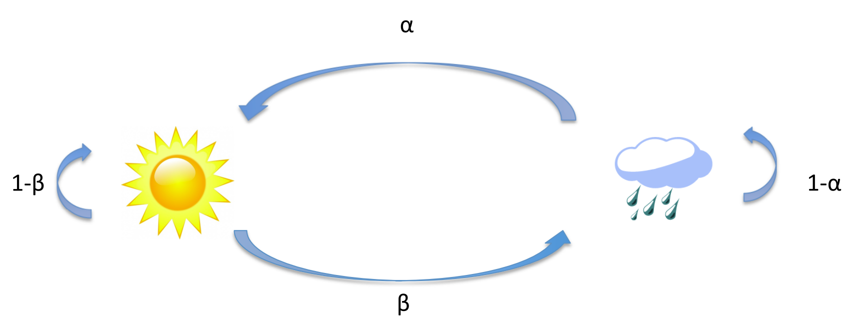
- For this kind of Markov chain, define transition matrix

$$P = \begin{bmatrix} P_{11} & \cdots & P_{1n} \\ & \cdots & \\ P_{n1} & \cdots & P_{nn} \end{bmatrix}$$

Simple weather model

- $\mathcal{X}$ = { Sunny: S, Rainy R }
- $p(\text{S}|\text{S}) = 1 - \beta$, $p(\text{R}|\text{R}) = 1 - \alpha$, $p(\text{R}|\text{S}) = \beta$, $p(\text{S}|\text{R}) = \alpha$

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



- Probability of seeing a sequence SSRR:

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

What will this sequence behave, after many days of observations?

- What sequences of observations are more typical?

- What is the probability of seeing a typical sequence?

Stationary distribution

- Stationary distribution: a distribution μ on the states such that the distribution at time $n+1$ is the same as the distribution at time n.

- Our weather example:
  - If $\mu(\text{S}) = \frac{\alpha}{\alpha+\beta}$, $\mu(\text{R}) = \frac{\beta}{\alpha+\beta}$

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

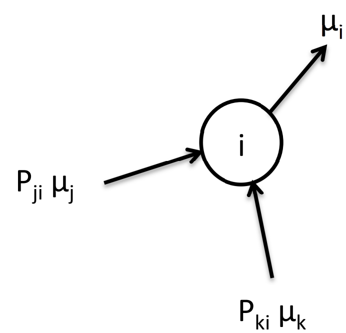
  - Then

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

- How to calculate stationary distribution
  - Stationary distribution μ_i, $i = 1, \cdots, |\mathcal{X}|$ satisfies

$$\mu_i = \sum_j \mu_j p_{ji}, (\mu = \mu P), \text{and } \sum_{i=1}^{|\mathcal{X}|} \mu_i = 1.$$

  - "Detailed balancing":



Stationary process

- A stochastic process is stationary if the joint distribution of any subset is invariant to time-shift

$$p(X_1 = x_1, \cdots, X_n = x_n) = p(X_2 = x_1, \cdots, X_{n+1} = x_n).$$

- Example: coin tossing

$$p(X_1 = \text{head}, X_2 = \text{tail}) = p(X_2 = \text{head}, X_3 = \text{tail}) = p(1 - p).$$

Entropy rate

- When X_i are i.i.d., entropy $H(X^n) = H(X_1, \cdots, X_n) = \sum_{i=1}^n H(X_i) = nH(X)$
- With dependent sequence X_i, how does $H(X^n)$ grow with n? Still linear?
- Entropy rate characterizes the growth rate
- Definition 1: average entropy per symbol

$$H(\mathcal{X}) = \lim_{n\to\infty} \frac{H(X^n)}{n}$$

- Definition 2: rate of information innovation

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

- $H'(X)$ exists, for X_i stationary

$$H(X_n|X_1, \cdots, X_{n-1}) \le H(X_n|X_2, \cdots, X_{n-1}) \quad (1)$$
$$\le H(X_{n-1}|X_1, \cdots, X_{n-2}) \quad (2)$$

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

- $H(\mathcal{X}) = H'(\mathcal{X})$, for X_i stationary

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

- Each $H(X_n|X_1, \cdots, X_{n-1}) \to H'(\mathcal{X})$

- Cesaro mean:

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

- So

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

AEP for stationary process

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

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

Entropy rate for Markov chain

- For Markov chain

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

- By definition

$$p(X_2 = j|X_1 = i) = P_{ij}$$

- Entropy rate of Markov chain

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

Calculate entropy rate is fairly easy

1. Find stationary distribution μ_i

2. Use transition probability P_{ij}

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

Entropy rate of weather model

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

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

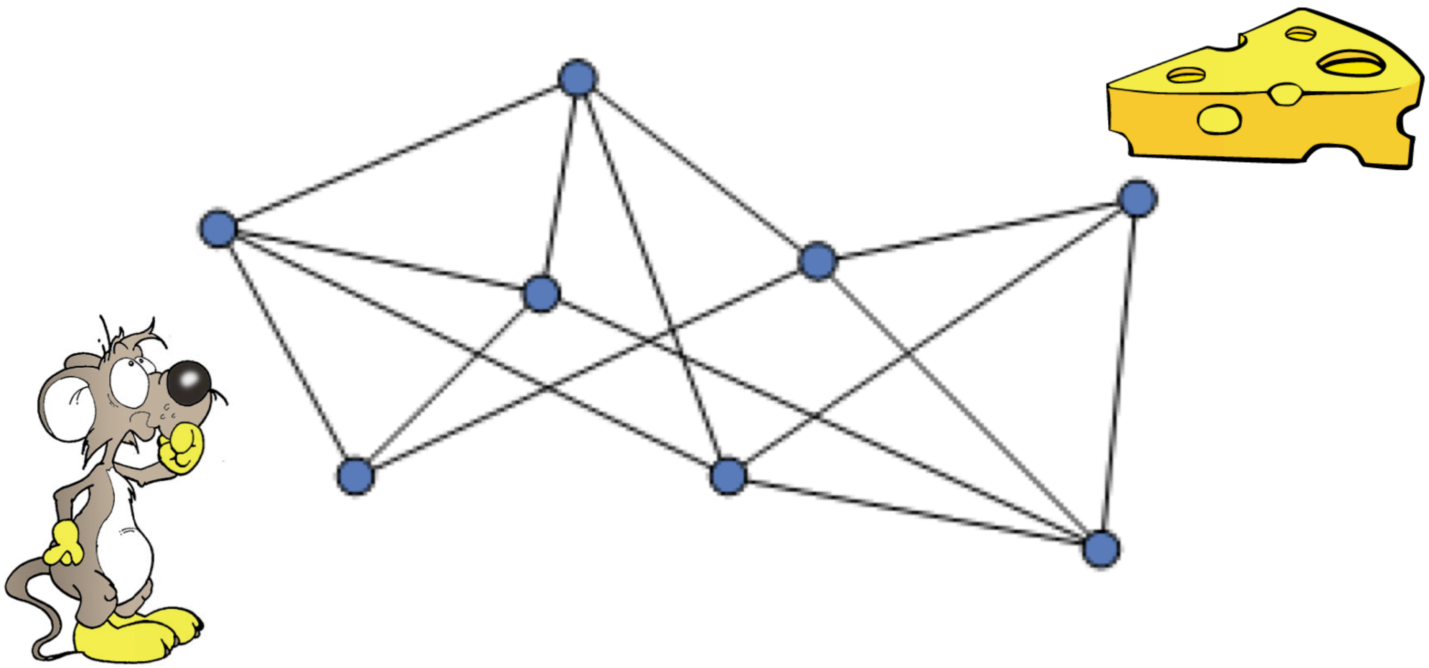
$$\begin{aligned} H(\mathcal{X}) &= \frac{\beta}{\alpha+\beta}[\alpha \log \alpha + (1-\alpha)\log(1-\alpha)] \\ &= \frac{\alpha}{\alpha+\beta}H(\beta) + \frac{\beta}{\alpha+\beta}H(\alpha) \\ &\leq H\left(2\frac{\alpha\beta}{\alpha+\beta}\right) \leq H\left(\sqrt{\alpha\beta}\right) \end{aligned}$$

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

Random walk on graph

- An undirected graph with m nodes $\{1, \ldots, m\}$
- Edge $i \to j$ has weight $W_{ij} \geq 0$ ($W_{ij} = W_{ji}$)
- A particle walks randomly from node to node
- Random walk $X_1, X_2, \cdots$: a sequence of vertices
- Given $X_n = i$, next step chosen from neighboring nodes with probability

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



Entropy rate of random walk on graph

- Let

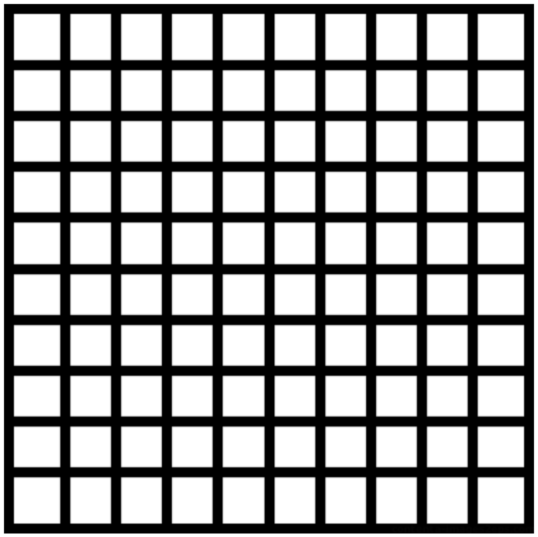
$$W_i = \sum_j W_{ij}, \qquad W = \sum_{i,j:i>j} W_{ij}$$

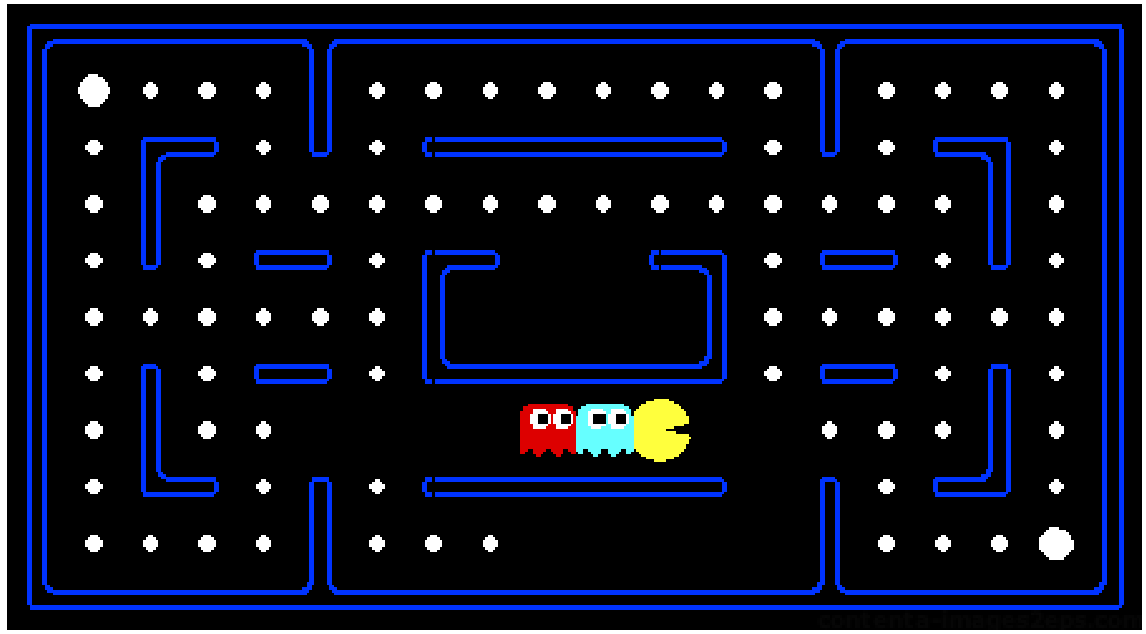
- Stationary distribution is

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

- Can verify this is a stationary distribution: $\mu P = \mu$

- Stationary distribution $\propto$ weight of edges emanating from node i (locality)





Summary

- AEP Stationary process $X_1, X_2, \cdots, X_i \in \mathcal{X}$: as $n \to \infty$

$$p(x_1, \cdots, x_n) \approx 2^{-nH(\mathcal{X})}$$

- Entropy rate

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

- Random walk on graph

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