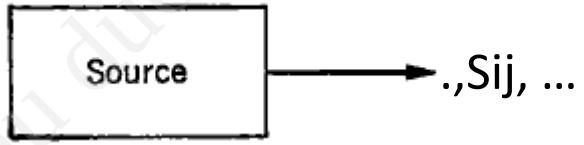
Chapter 3.4: Nguồn tin

3.4.1. Nguồn tin là gì?

- Thông tin là khái niệm trừu tượng. Để nói về thông tin, lý thuyết thông tin gán cho mỗi tin một ký hiệu của một nguồn
- Tập ký hiệu của nguồn cũng được gọi là bảng chữ của nguồn thường là hữu hạn S = {s1, s2, …, sq}
- Nguồn phát một choỗi các ký hiệu (bản tin) từ bangr chữ cái (alphabet) m = {si1, si2, …} ; sij là ký hiệu si ϵ S, được tạo ra tại thời điểm j
- Mỗi ký hiệu được tạo ra tuân theo một luật phân bố xác suất
- Mô hình S



- Tại mỗi thời điểm, ký hiệu được phát ra được coi là 1 giá trị của một biến ngâu nghiên (ví dụ X)
  - Xác suất của giá trị của biến ngâu nhiên = xác suất của ký hiệu
- Nguồn là một biến ngẫu nhiên

3.4.2. Các loại nguồn

- Nguồn rời rạc
  - Tạo ra các chữ cái (ký hiệu nguồn) rời rạc
  - Bảng chữ cái thường là hữu hạn
  - Nguồn được mô tả bởi một biến ngẫu nhiên
  - Các loại nguồn rời rạc:
    - *Nguồn rời rạc không nhớ: các chữ được tạo ra độc lập nhau*.
      - Chữ tạo ra ở một thời điểm không phụ thuộc vào chữ tạo ra ở bất cứ thời điểm nào khác
      - Biến ngẫu nhiên mô tả nguồn này là
        - X = {x1, x2…xn}
        - P(X) = {P(x1), P(x2),…P(Xn)}
    - ***Nguồn rời rạc có nhớ: một ký hiệu nguồn (chữ) được tạo ra phụ thuộc vào một số chữ đã tạo ra trước đó***
      - Cấp của nguồn là thứ tự nguồn (tính các chữ đã tạo ra trước đó)
      - Nguồn có nhớ thường được mô hình hóa bởi chuỗi Markov và gọi là nguồn Markov.
    - Nguồn Ergodic là nguồn có đặc trưng không phụ thuộc gốc thời gian và trị trung bình theo thời gian bằng trị trung bình theo tập hợp

3.4.2. các loại nguồn (cont.)

- Nguồn liên tục:
  - Bản tin tạo ra là liên tục (theo cả thời gian và giá trị)
    - Bản tin tạo ra sẽ có dạng một hàm liên tục
    - Biến ngẫu nhiên mô tả nguồn liên tục
      - X = P{x} xmin <x < xmax
      - P{x}: Hàm mật độ xác suất

3.4.2. các loại nguồn (Cont.)

- Nguồn nhị phân:
  - Nguồn rời rạc
  - Bảng chữ hay tập tin của nguồn chỉ có 2 giá trị
  - Ví dụ: X = {0,1}; P(X)= {0.5, 0.5}
- Nguồn Markov:
  - Mỗi ký hiệu nguồn chỉ phụ thuộc vào 1 ký hiệu xuất hiện trước nó.

$$p(x_{i_n}|x_{j_{n-1}}, x_{k_{n-2}} \ldots) = p(x_{i_n}|x_{j_{n-1}})$$

  - Tại thời điểm n, đầu ra của nguồn là ký hiệu xj với xác suất pij = p(xj,n|xi,n-1) khi tại (n-1) đầu ra của nguồn là xi
  - $\sum_{j=1}^{L} p_{ij} = 1$ L: số lượng ký hiệu của nguồn

3.4.2. các loại nguồn (Cont.)

- Nguồn Markov cấp m:
  - Mỗi ký hiệu phụ thuộc vào m ký hiệu xuất hiện trước nó
- nguồn Markov gồm:
  - Alphabet
  - Tập xác suất trạng thái
  - Tập phép chuyển trạng thái
  - Tập các nhãn (label) cho mỗi phép chuyển trạng thái
  - Hai tập xác suất
    - Phân bố xác suất ban đầu của các trạng thái xác định xác suất của các choỗi bắt đầu với từng ký tự .
    - Tập các xác suất chuyển với mỗi cặp trạng thái
      - Nhãn trên chuyển là ký tự được tạo ra

5.2. Các loại nguồn (Cont.)

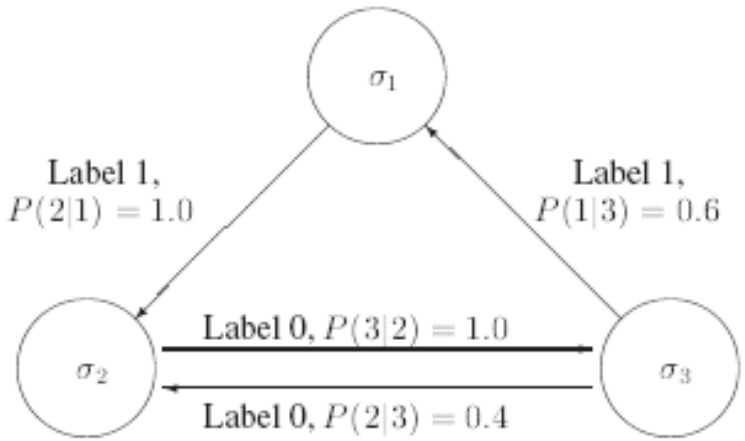
- Ví dụ về nguồn Markov:
- Alphabet {0,1} và tập trạng thái $\{\sigma_1, \sigma_2, \sigma_3\}$
- Các chuyển trạng thái có thể:

$\sigma_1 \rightarrow \sigma_2$ với nhãn (label 1) và P(2|1) = 1

$\sigma_2 \rightarrow \sigma_3$ với label 0 và P(3|2) = 1

$\sigma_3 \rightarrow \sigma_1$ với label 1 và P(1|3) = 0.6

$\sigma_3 \rightarrow \sigma_2$ với label 0 và P(2|3) = 0.4



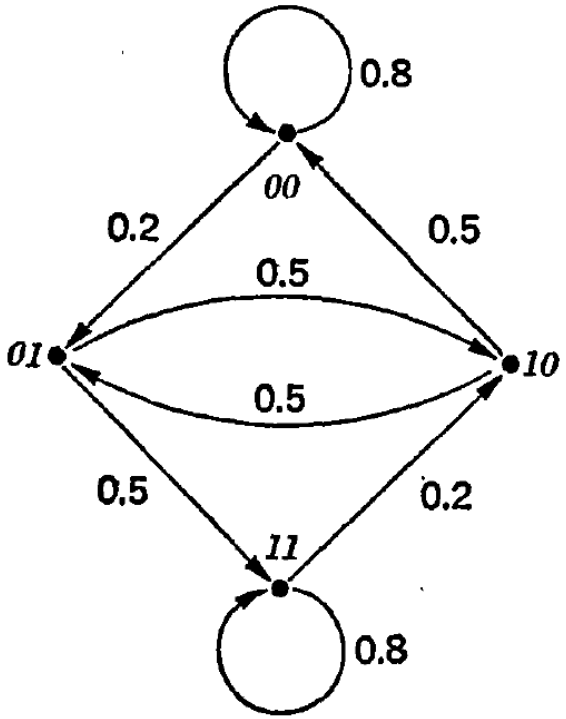
- Phân bố xác suất ban đầu: $P(\sigma_1) = 1/3$, $P(\sigma_2) = 1/3$, $P(\sigma_3) = 1/3$

5.2. Các loại nguồn (Cont.)

- Ví dụ: second-order Markov source

{0,1}
P(0|00) = P(1|11) = 0.8
P(1|00) = P(0|11) = 0.2
P(0|01) = P(0|10) = P(1|01) = P(1|10) = 0.5

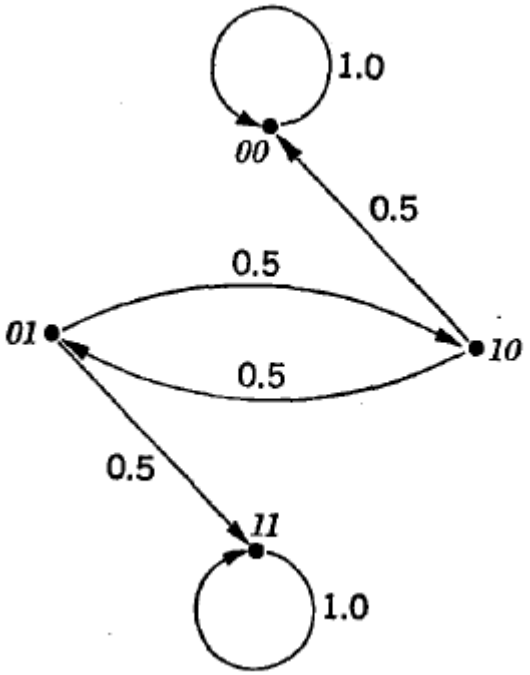
Xác suất chuyển từ 01 đến 10, được biểu diễn bởi P(10|01),sẽ được biểu diễn bởi xác suất tạo ký hiệu 0 khi ở trạng thái 01, nó là P(0|01)



3.4.2. Các loại nguồn (Cont.)

- Nguồn Markov không ergodic

{0,1}
P(0|00) = P(1|11) =1.0
P(1|00) = P(0|11) = 0
P(0|01) = P(0|10) = P(1|01) = P(1|10) = 0.5



3.4.2. Các loại nguồn (Cont.)

- n trạng thái {... } có
  - Matrix chuyển:

$$\Pi = \begin{bmatrix} P(1|1) & P(1|2) & \cdots & P(1|N) \\ P(2|1) & P(2|2) & \cdots & P(2|N) \\ \vdots & \vdots & \ddots & \vdots \\ P(N|1) & P(N|2) & \cdots & P(N|N) \end{bmatrix}$$

  - là xác suất ở trạng thái tại thời điểm t

$$W^t = \begin{bmatrix} w_1^t \\ w_2^t \\ \vdots \\ w_N^t \end{bmatrix}$$

  - Và :

$$W^{t+1} = \Pi W^t$$

$$W^t = \Pi^t W^0$$

3.4.2. các loại nguồn (Cont.)

- Nguồn dừng: Hàm phân bố xác suất W trên các trạng thái của nguồn Markov với ma trận chuyên Π thỏa mãn Π W= W
- Ví dụ
  - $\sum w_i = 1$

$$W = \{w_1, w_2, w_3\}$$

$$\Pi = \begin{bmatrix} 0.25 & 0.50 & 0.00 \\ 0.50 & 0.00 & 0.25 \\ 0.25 & 0.50 & 0.75 \end{bmatrix}$$

3.4.3. Lượng tin riêng của nguồn

- Nguồn không nhớ:
- Lượng tin riêng của ký hiệu Si

$$I(s_i) = \log \frac{1}{P(s_i)}$$

- Lượng tin trung bình của các tin hay lượng tin riêng của nguồn

$$\sum_S P(s_i) I(s_i)$$

- Entropy của nguồn

$$H(S) \triangleq \sum_S P(s_i) \log \frac{1}{P(s_i)}$$

- H(S) max = log |S| khi nguồn S có phân bố đều (các ký hiệu có cùng xác suất)

3.4.3. (Cont.)

- Ví dụ:
  - Nguồn S = {s1, s2, s3} Có P(s1) = 1/2 và P(s2) = P(s3) = 1/4.
  - Lúc này:

$$H(S) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{4}\log_2\frac{1}{4}$$

$$= \frac{1}{2}\log_2 2 + \frac{1}{4}\log_2 4 + \frac{1}{4}\log_2 4$$

$$= \frac{3}{2} \text{ bits/tin}$$

3.4.3. (Cont.)

- Nguồn Markov:
  - : Phân bố xác suất của tập các trạng thái ở thời điểm
  - : entropy của mỗi trạng thái ở thời điểm thứ
  - M:

$$H(P_i) = -\sum_{j=1}^{N} P(j|i) \log(P(j|i)$$

$$H(M) = \sum_{i=1}^{N} w_i H(P_i) = -\sum_{i=1}^{N} \sum_{j=1}^{N} w_i P(j|i) \log(P(j|i))$$

3.4.3. Lượng tin riêng (Cont.)

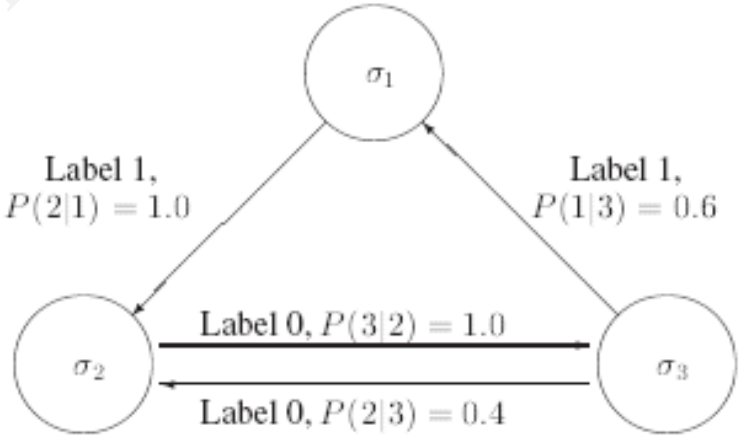
$\sigma_1 \rightarrow \sigma_2$ with label 1 and P(2|1) = 1

$\sigma_2 \rightarrow \sigma_3$ with label 0 and P(3|2) = 1

$\sigma_3 \rightarrow \sigma_1$ with label 1 and P(1|3) = 0.6

$\sigma_3 \rightarrow \sigma_2$ with label 0 and P(2|3) = 0.4

H(P_i) =? w_i = ? H(M)=?



3,4.3. (Cont.)

- Nguồn liên tục:
  - Entropy của nguồn dừng:

$$H(X) = -\int_{-\infty}^{\infty} p(x) \log p(x) dx$$

  - H(X) max:
    - Nguồn có công suất đỉnh hữu hạn: Pmax, Pmin là các giá trị hữu hạn
      - xmax = ; xmin =
      - H(X) max = log (xmax – xmin) khi nguồn có phân bố đều (P(x) = 1/(xmax- xmin)) cho mọi x)
    - Nguồn có công suất trung bình hữu hạn: Pav is là giá trị hữu hạn
      - H(X) max = log
        - e: cơ số tự nhiên

3.4.3. (Cont.)

- Nguồn liên tục:
  - Entropy của nguồn dừng:

$$H(X) = -\int_{-\infty}^{\infty} p(x) \log p(x) dx$$

  - H(X) max:
    - Nguồn có công suất đỉnh hữu hạn r: Pmax là giá trị hữu hạn
      - xmax = ; xmin = -xmax (hữu hạn)
      - H(x) = log (x max - x min)
      - H(X) max = log (2xmax) khi nguồn có phân bố đều (P(x) = 1/(2xmax) với mọi x)
    - Nguồn có công suất trung bình hạn chế: Pav là giá trị hữu hạn
      - H(X) max = ln $\sqrt{2\pi e P x}$
        - e: cơ số tự nhiên

$$\int_{\infty}^{\infty} x^2 p(x) dx = P_{av}^2$$

3.4.4. Độ dư của nguồn

- Nguồn có H(X)max:
  - Lượng tin mang bởi mỗi tin của nguồn là max
- Nguồn có H(X) < H(X)max:
  - Lượng tin mang bởi mỗi tin của nguồn chưa đạt max
- Số lượng tin trong choỗi của nguồn có H(X)max là min để mạng lượng tin xác định
  - Tạo ra nguồn tin cho trước: Nguồn có H(X) < H(X)max cần tạo nhiều tin hơn nguồn H(X) = H
  - Nguồn có H(X) < H(X)max có sự dư thừa (tin tạo ra)
- Độ dư của nguồn định nghĩa bởi H(X)max – H(X)
  - Miền xác định của nguồn có H(X)max và H(X) giống nhau
- Nguồn có độ dư = 0: Mỗi ký hiệu mạng lớn tin lớn nhất
- Nguồn có độ dư >0: Cần nén để giảm bớt số ký hiệu
  - Nén tốt nhất sẽ đạt được khi làm cho H(X) = H(X)max

3.4.4. Độ dư của nguồn(Cont.)

- Ví dụ:
  - Nguồn S1 = {0,1} với P(S1) = {1/2,1/2}
    - H(S1)max = – log1/2 – log1/2 = 2 = 1 bit/ ký hiệu
  - Nguồn S2 ={0,1} với P(S2) = {3/4,1/4}
    - H(S2) = – 3/4log3/4 – 1/4log1/4 ≈ 2 – 1.19 ≈ 0.81 bits/ký hiệul

  →Để tạo lượng tin 810 bits
    - S1 cần tạo 810 ký hiệu
    - S2 cần tạo 1000 ký hiệu
  - S2 có độ dư: H(X)max – H(X) = 1 – 0.81 = 0.19 bits/ký hiệu

3.4.5. Mở rộng nguồn

- Mở rộng nguồn cho nguồn S:
  - S^n là nguồn mà mỗi ký hiệu của nó s_i^n là chuỗi n ký hiệu của nguồn S (s_{ij} là một ký hiệu của nguồn S nằm ở vị trí j trong ký hiệu thứ I của nguồn mở rộng)
    - $s_i^n = s_{i1}s_{i2}s_{i3} \dots s_{in}$
  - Các ký hiệu của nguồn S trong s_i^n là độc lập
    - $P(s_i^n) = P(s_{i1})\, P(s_{i2}) \dots P(s_{in})$
- Entropy của S^n:
  - $H(S^n) = n\, H(S)$

3.4.5. Extension source

- Nguồn không nhớ S{0,1}
- $P_0 = 0.2$, $P_1 = 0.8$
- Nguồn mở rộng?

E.g: P_{00}, P_{001}? H(S^2)?

3.4.6. Tốc độ tạo tin của nguồn

- Tốc độ tạo tin của nguồn (R): Là lượng tin trung bình mà nguồn tạo được trong một đơn vị thời gian
- R = $n_o \times H(X)$
  - n_o: Số tin nguồn có thể tạo trong một đơn vị thời gian
  - H(X): Lượng tin trung bình chưa trong mỗi tin (entropy)
- Trong trường hợp của lý thuyết thông tin, n_o là tham số vật lý nên n_o có thể coi có giá trị đơn vị ($n_o = 1$)
- Trong trường hợp rời rạc
  - n_o = F
    - F: là số tin nguồn tạo ra trong một đơn vị thời gian hay nhịp tạo tin của nguồn
    - R = F × *H(X)*
    - *Rmax = F × log |X|*

3.4.6. Tốc độ tạo tin (Cont.)

- Nguồn tạo 9.6 kbaud:

  (baud = tin/ s)

X_i	$P(X_i)$	BCD word
A	0.30	000
B	0.10	001
C	0.02	010
D	0.15	011
E	0.40	100
F	0.03	101

- Tốc đọ tạo tin R =?

3.4.6. Tốc độ tạo tin (Cont.)

$$
\begin{aligned}
H &= -\sum_{i=1}^{6} P(X_i) \cdot \log_2 P(X_i) = -0.30 \cdot \log_2 0.30 - 0.10 \cdot \log_2 0.10 - 0.02 \cdot \log_2 0.02 \\
&\quad -0.15 \cdot \log_2 0.15 - 0.40 \cdot \log_2 0.40 - 0.03 \cdot \log_2 0.03 \\
&= 0.52109 + 0.33219 + 0.11288 + 0.41054 + 0.52877 + 0.15177 \\
&= 2.05724 \text{ bits/symbol}
\end{aligned}
$$

Information rate: $R = H \cdot R_s = 2.05724$ [bits/symbol] $\cdot$ 9600 [symbols/s] $= 19750$ [bits/s]

3.4.6. Tốc độ tạo tin (Cont.)

- Trong trường hợp liên tục
  - Nguồn liên tục tương đương với nguồn được rời rạc từ nó với chu kỳ lấy mẫu là 1/(2 fmax) hay số mẫu trong 1 đơn vị thời gian là 2fmax
- n_o là số mẫu nguồn tạo ra được trong một đơn vị thời gian
  - n_o = 2 Fmax
    - Fmax: tần số lớn nhất có trong bản tin nguồn tạo ra
    - R = 2 Fmax x H(x)
    - R = 2 Fmax x log (xmax – xmin) khi nguồn có công suất đỉnh hạn chế
    - R = 2 Fmax x log $\sqrt{2\Pi eP}av$ khi nguồn có công suất trung bình hạn chế

Bài tập 1.

- Cho một nguồn chỉ tạo ra một bản tin có nội dung là “công nghệ thông tin” viết ở dạng chữ Việt không dấu, không phân biệt chữ thường chữ hoa, không có dấu cách giữa các từ. Mỗi ký tự trong bản tin là một tin được tạo ra từ nguồn. Xác suất xuất hiện của mỗi tin bằng tỷ số của số lần xuất hiện tin chia cho độ dài bản tin (tần suất xuất hiện của tin trong bản tin)

    a. Hãy viết đúng bản tin được tạo ra

    b. Hãy xác định mô hình nguồn (hai tập giá trị: tập tin của nguồn X và tập xác suất xuất hiện mối tin của nguồn p(X).

c. Tinh Entropy của nguồn?

d. Tính lượng tin của bản tin?