Chương 4. Truyền tin trên kênh nhiễu

ntnhut@hcmus.edu.vn

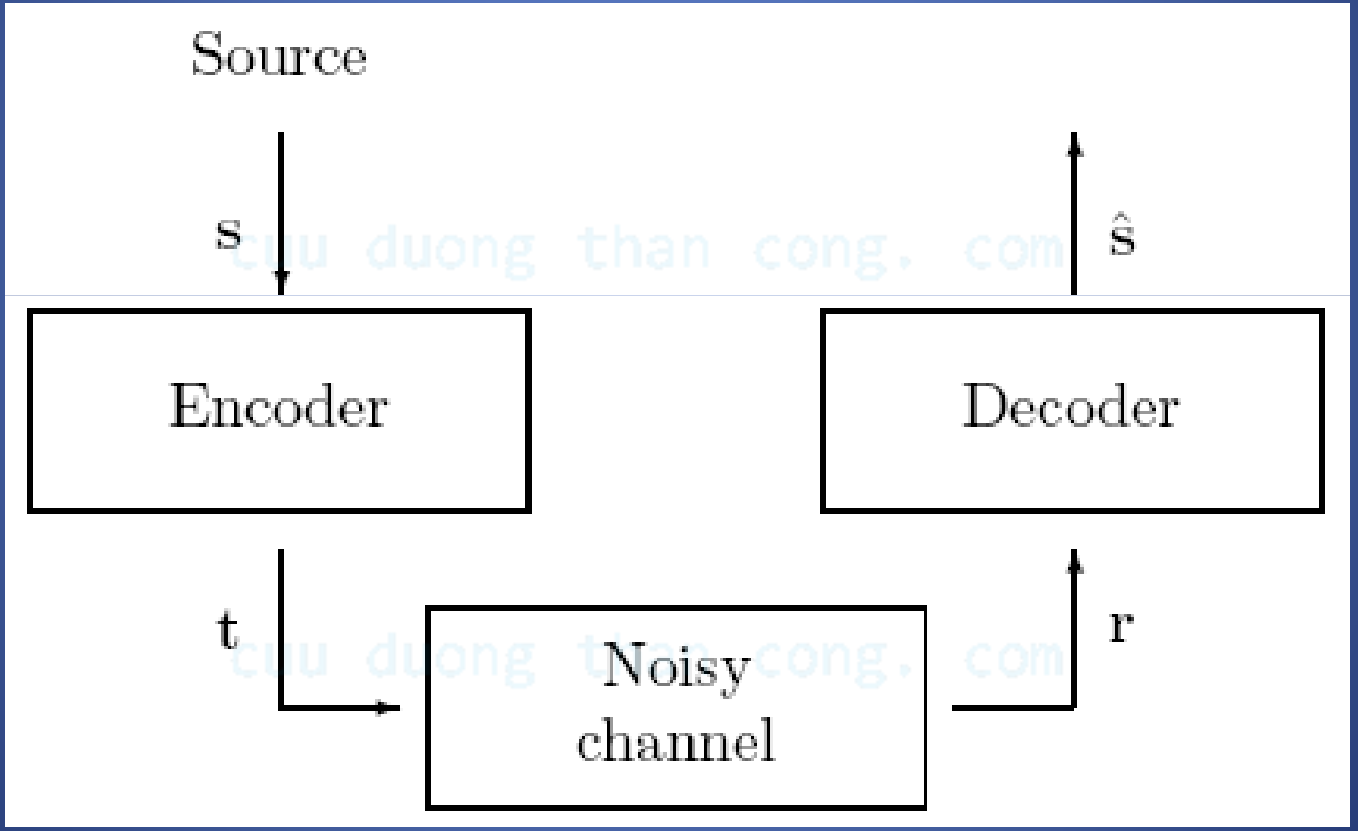
Mã dùng để phát hiện và tự sửa lỗi

- Kênh truyền hay thiết bị lưu trữ thông tin không tránh khỏi bị nhiễu/lỗi.

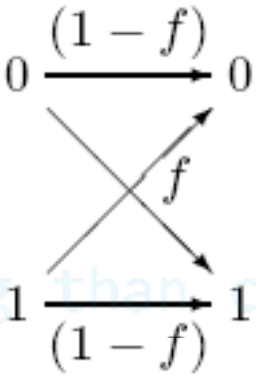
- Lý thuyết mã có thể dùng để phát hiện lỗi và tự sửa lỗi.

→ ***Mã tự sửa (error-correcting code)***

Giải pháp

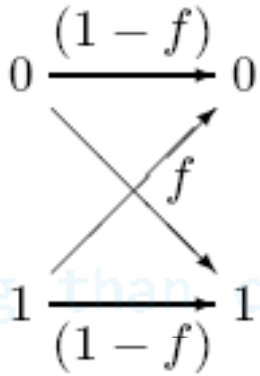


Kênh đối xứng nhị phân



1. Inputs = ouputs = {0,1}
2. P(nhận 1| truyền 0) = P(nhận 0| truyền 1) = f
   → ***error probability***.
3. P(nhận 1| truyền 1) = P(nhận 0| truyền 0) = 1 – f.

Tỷ lệ nhiễu/lỗi



Hình 1. Ảnh nhị phân kích thước 10.000 bits truyền trên kênh nhị phân đối xứng với tỷ lệ nhiễu/lỗi f = 0.1 (cứ khoảng 10 bit thì có 1 bit bị lỗi 0 ⇔ 1).

Bài toán tính xác suất nhận đúng tín hiệu

- **Tín hiệu nhị phân:**
  - Một tín hiệu được tạo thành từ những bit 0,1. Qua thống kê, ta biết do nhiễu, bình quân 1/5 số bit 0 và ¼ số bit 1 bị lỗi.
  - Biết rằng tỷ số số bit 0 và 1 truyền đi là 5:3. Tính xác suất nhận đúng tín hiệu phát đi.
- **Ký hiệu:**
  - T_0 = "phát bit 0", T_1 = "phát bit 1"
  - N_0 = "nhận bit 0", N_1 = "nhận bit 1".
- **Tính:**
  - $P(T_0 \mid N_0) = ?$
  - $P(T_1 \mid N_1) = ?$

Cách giải bài toán tín hiệu nhị phân

- Tính các xác suất:
  - $P(T_0)$, $P(T_1)$
  - $P(N_0 \mid T_0)$, $P(N_0 \mid T_1)$, $P(N_1 \mid T_1)$, $P(N_1 \mid T_0)$,
- Dùng công thức Bayes

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}.$$

- Tính
  - $P(T_0 \mid N_0) = ?$
  - $P(T_1 \mid N_1) = ?$

Bài tập

VD cách khắc phục lỗi: Mã lặp

0	000
1	111

Repetition code K_3

- Ví dụ mã hoá

Nhiễu

s	0	0	1	0	1	1	0
t	000	000	111	000	111	111	000
n	000	001	000	000	101	000	000
r	000	001	111	000	010	111	000

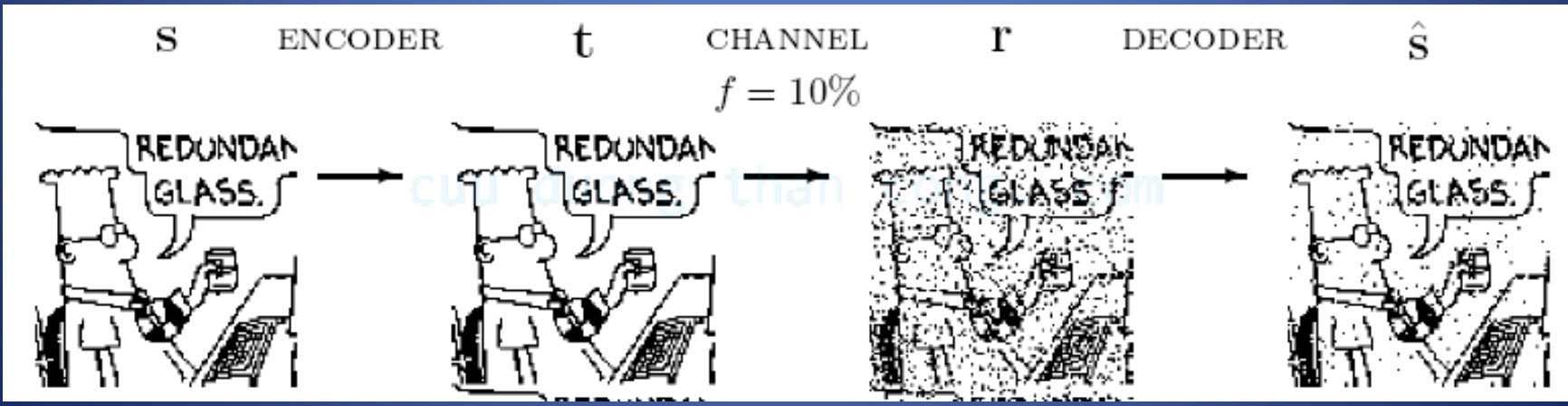
Giải mã mã lặp theo luật đa số

- ***Luật đa số (Majority vote rule)***

Received sequence r	Decoded sequence $\hat{s}$
000	0
001	0
010	0
100	0
101	1
110	1
011	1
111	1

Giải mã, phát hiện và tự sửa lỗi

s	0	0	1	0	1	1	0
t	000	000	111	000	111	111	000
n	000	001	000	000	101	000	000
r	000	001	111	000	010	111	000
$\hat{s}$	0	0	1	0	0	1	0



Lỗi khi tự sửa

s	0	0	1	0	1	1	0
t	000	000	111	000	111	111	000
n	000	001	000	000	101	000	000
r	000	001	111	000	010	111	000
$\hat{s}$	0	0	1	0	0	1	0
corrected errors		⋆					
undetected errors					⋆		

Xác suất giải mã bị lỗi $p_{err}(K)$

- t = "$a_1a_2a_3$" → r = "$b_1b_2b_3$".
- Giải mã theo "luật đa số" bị lỗi khi:
  - $b_i \neq a_i$, $\forall i$. Xác suất: f^3.
  - Có 2 vị trí bị lỗi.

    3 trường hợp

  → xác suất = $3f^2(1 - f)$.

- $p_{err}(K_3) = f^3 + 3f^2(1 - f) = 3f^2 - 2f^3$

VD giảm lỗi: Mã lặp K_5

0	00000
1	11111

Repetition code K_5

- Xác suất giải mã bị lỗi

$$P_{err}(K_5) = p^5 + 5p^4q + \binom{5}{3}p^3q^2$$

- Ít lỗi hơn K_3.

Bài tập

- Tính xác suất giải mã bị lỗi của mã lặp K_N $p_{err}(K_N)$

- **BT Thực hành:** tạo hàm mã hoá và giải mã nhị phân mã lặp:
  - Mã hoá t = EncodeK(N,x),
  - Tạo nhiễu ngẫu nhiên tỷ lệ f r = AddNoise(t,f)
  - Giải mã y =DecodeK(N,r).
  - Tính tỷ lệ lỗi p sau khi so sánh x và y.

Tỷ lệ thông tin

- Mã khối nhị phân K có các từ mã dài n bits:
  - Chỉ có k bits "mang thông tin".
  - n – k bits kiểm tra lỗi.
  - Có tất cả 2^k từ mã.
- VD: Mã lặp K_n:
  - Chỉ có 1 bit ý nghĩa
  - n – 1 bits còn lại lặp lại bit đầu tiên.

$R(K_n) = 1/n$

Định nghĩa: ***Tỷ lệ thông tin (information rate)*** của mã khối K độ dài n của nguồn r ký tự có r^k từ mã là

$$R(K) = k/n.$$

Tính chất của tỷ lệ thông tin

- $0 \leq R(K) = k/n \leq 1$.
- $R(K) = 0$ khi $k = 0$: không có thông tin.
- $R(K) = 1$ khi $k = n$: mã không phát hiện + sửa được lỗi.

- $R(K) \rightarrow 0$: không hiệu quả về mặt chi phí nhưng phát hiện + sửa lỗi hiệu quả.
- $R(K) \rightarrow 1$: hiệu quả về mặt chi phí nhưng phát hiện + sửa lỗi không hiệu quả.

Ví dụ tỷ lệ thông tin cao: Mã kiểm chẵn lẻ

Information Bits	Code Word	Information Bits	Code Word
000	0000	110	1100
100	1001	101	1010
010	0101	011	0110
001	0011	111	1111

Figure 2: The even-parity code of length 4

- R(K) = (n – 1)/n
- Có thể phát hiện 1 bit lỗi.

Bài toán lập mã tự sửa

Lập mã tự sửa K thoả:

1. Biết trước xác suất lỗi khi truyền mỗi bit = p.
2. Xác suất giải mã bị lỗi không quá $p_{err}(K)$.
3. Tỷ lệ thông tin không dưới R(K).
4. Phát hiện/sửa lỗi được càng nhiều càng tốt.

Ví dụ mã K_4*

Cho p = 0.001, $p_{err}(K) \leq 0.0002$, $R(K) \geq 0.5$.

VD với mã K_4* (lặp 3 lần bit thứ hai)

- Giải mã chính xác khi:
  - cả 4 bit đều đúng (xác suất $q^4 = (1 - p)^4$) hoặc
  - bit đầu + 2 bit còn lại đúng (xác suất $q3pq^2 = 3pq^3$).
- $p_{err}(K_4^*) = 1 - (q^4 + 3pq^3)$

$= 1 - (0.999^4 + 3(0.001)0.999^3)$

$\cong 0.0003$.

($> p_{err}(K)$) → K_4* chưa tốt

Information Bits	Code Word
00	0000
01	0111
10	1000
11	1111

Ví dụ mã K_6*

- Các từ mã của K_6* khác nhau đôi một ít nhất 3 bit.

→ phát hiện + tự sửa được chính xác 1 bit lỗi.

VD: nhận "010100", không có trong bộ từ mã → có lỗi.

- "010100" Chỉ khác 1 bit đối với từ mã "010101".

→ giải mã thành "010".

- Giải mã chính xác khi:
  - Cả 6 bit đều đúng (q^6), hoặc
  - Chỉ có 1 bit sai ($6pq^5$)

Information Bits	Code Word
000	000000
100	100011
010	010101
001	001110
011	011011
101	101101
110	110110
111	111000

$p_{err}(K_6^*) = 1 - (q^6 + 6pq^5)$
$\cong 0.00015 \; (< p_{err}(K)$ ☺$)$.

Khoảng cách Hamming

Định nghĩa: ***Khoảng cách Hamming*** của hai từ $a=a_1a_2\ldots a_n$ và $b=b_1b_2\ldots b_n$ là số vị trí mà a và b khác nhau, ký hiệu d(a,b).

$$d(a,b) = \#\{i \mid a_i \neq b_i\}.$$

Ví dụ:

- Mã lặp K_N có khoảng cách Hamming giữa các từ mã bằng N.
- Mã K_6* có khoảng cách Hamming giữa các từ mã là 3.

000		000000
100		100011
010	K_6*	010101
001		001110
011		011011
101		101101
110		110110
111		111000

Tính chất của d(a,b)

Với mọi từ a, b, c cùng độ dài n trong một bảng ký tự Σ, d(a,b) là một ***metric****:*

1. $d(a,a) = 0$; *và với* $a \neq b$ *thì* $d(a,b) > 0$.
2. $d(a,b) = d(b,a)$.
3. $d(a,b) + d(b,c) \geq d(a,c)$.

Chứng minh: (bài tập)

Giải mã hợp lý nhất

- Nếu từ nhận được, r, không có trong bộ mã

→ chọn từ mã có khoảng cách Hamming nhỏ nhất để giải mã

→ ***maximum likelihood decoding***.

Khoảng cách nhỏ nhất

Định nghĩa: ***khoảng cách nhỏ nhất d(K)*** của mã khối K là khoảng cách Hamming nhỏ nhất giữa hai từ mã khác nhau bất kỳ,

$$d(K) = \min\{d(a,b) \mid a \neq b \in K\}.$$

Ví dụ:

- Mã lặp có $d(K_N) = N$.
- Mã K_6^* có $d(K_6^*) = 3$.
- Mã kiểm chẵn lẻ có $d(K) = 2$.

Phát hiện lỗi

Mệnh đề: *Mã K phát hiện được t lỗi khi và chỉ khi d(K) > t.*

Chứng minh: (bài tập)

Ví dụ:

- Mã lặp K_N phát hiện được tối đa N-1 lỗi.
- Mã K_6^* phát hiện được tối đa 2 lỗi.
- Mã kiểm chẵn lẻ phát hiện được tối đa 1 lỗi.

Sửa lỗi

- Ý tưởng:
  - **Mã K sửa được t lỗi** nếu mỗi chuỗi ký tự a' nhận được từ việc gây lỗi ở 1, hoặc 2, …, hoặc t ký tự của từ mã a thì vẫn giải mã được duy nhất là a.
  - Nói cách khác:

$$\forall a \in K, \forall a' :$$

$$1 \le d(a,a') \le t \Rightarrow d(a,a') \le d(b,a'), \forall b \ne a \in K$$

Sửa lỗi

Mệnh đề: *Một mã sửa được t lỗi khi và chỉ khi khoảng cách nhỏ nhất của nó lơn hơn 2t.*

Chứng minh: (Bài tập)

Ví dụ:

- Mã lặp K_5 có thể sửa được 2 lỗi.
- Mã K_6* có d(K_6*) = 3, chỉ sửa được 1 lỗi.

Tóm tắt

- Tỷ lệ nhiễu f
- Bài toán tín hiệu nhị phân
- Mã lặp K_n
- Xác suất giải mã bị lỗi p_{err}
- Tỷ lệ thông tin R(K) = k/n.
- Khoảng cách Hamming d(a,b)
- Khoảng cách nhỏ nhất d(K)

Dung lượng kênh truyền

Kênh truyền

- **Kênh truyền rời rạc:**
  - Inputs: $x_1, x_2, \ldots, x_n$.
  - Outputs: $y_1, y_2, \ldots, y_m$.
- **Không nhớ:**
  - Output tại một thời điểm y_t chỉ phụ thuộc input x_t tại thời điểm đó. Không phụ thuộc các inputs trước.

- **Biết tính năng của kênh truyền như thế nào?**
  - thông qua thử nghiệm việc truyền mỗi x_j, với mọi $j = 1, 2, \ldots, n$.
  - Kết quả của việc thử cho ta các phân phối xác suất $P(y_1 \mid x_j), P(y_2 \mid x_j), \ldots, P(y_m \mid x_j)$ với mỗi input x_j.

Kênh thông tin rời rạc không nhớ

ĐN: ***Kênh thông tin rời rạc không nhớ*** gồm tập các ký tự input $\{x_1, x_2, \ldots, x_n\}$, tập các ký tự outputs $\{y_1, y_2, \ldots, y_m\}$ cùng với các phân phối xác suất $P(y_i \mid x_j)$ ứng với mỗi j = 1..nthoả

$$\sum_{i=1}^{m} P(y_i \mid x_j) = 1$$

VD: 1. Với n = m = 2 và $P(y_2 \mid x_1) = P(y_1 \mid x_2) = p$ là kênh đối xứng nhị phân.

2. Một kênh truyền theo thống kê thấy 1% input cho output ERROR và 99% truyền đúng.

y_i	0	1	E
$P(y_i \mid 0)$	0.99	0	0.01
$P(y_i \mid 1)$	0	0.99	0.01

Nhắc lại một số công thức xác suất

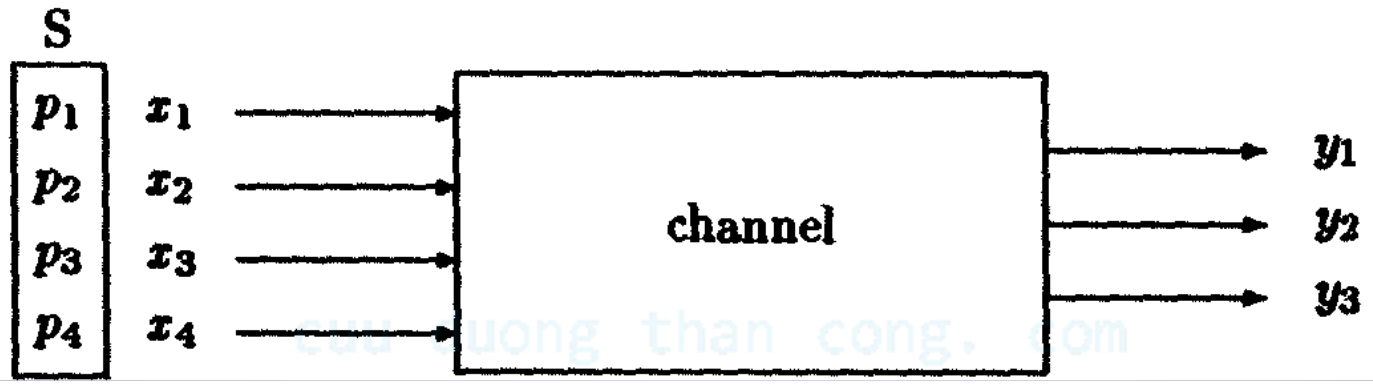


Figure 9: An information channel with an information source

$$P(x_j, y_i) = P(x_j)P(y_i \mid x_j).$$

$$P(y_i) = \sum_{j=1}^{n} P(x_j, y_i).$$

$$P(x_j \mid y_i) = \frac{P(x_j, y_i)}{P(y_i)} = \frac{P(x_j)P(y_i \mid x_j)}{\sum_{k=1}^{n} P(x_k)P(y_i \mid x_k)}.$$

VD Tính các xác suất

y_i	0	1	E
$P(y_i \mid 0)$	0.99	0	0.01
$P(y_i \mid 1)$	0	0.99	0.01

$p_0 = P(0)$ and $p_1 = P(1) = 1 - p_0$.

$$P(x_j, y_i) = P(x_j)P(y_i \mid x_j).$$

y_i	0	1	E
$P(0, y_i)$	$0.99p_0$	0	$0.01p_0$
$P(1, y_i)$	0	$0.99p_1$	$0.01p_1$
$P(y_i)$	$0.99p_0$	$0.99p_1$	0.01

$$P(x_j \mid y_i) = \frac{P(x_j, y_i)}{P(y_i)} = \frac{P(x_j)P(y_i \mid x_j)}{\sum_{k=1}^{n} P(x_k)P(y_i \mid x_k)}.$$

y_i	0	1	E
$P(0 \mid y_i)$	1	0	p_0
$P(1 \mid y_i)$	0	1	p_1

Entropy điều kiện và thông tin chung

ĐN: Cho một nguồn thông tin S và một kênh thông tin với inputs là các ký tự của S.

- ***Entropy có điều kiện***, ký hiệu H(X|Y), là giá trị

$$H(X|Y) = -\sum_{i=1}^{m}\sum_{j=1}^{n} P(x_j, y_i)\log_2 P(x_j|y_i) \quad \text{(bits)}$$

- ***Lượng thông tin***, ký hiệu I(X,Y), là giá trị

$$I(X,Y) = H(S) - H(X|Y)$$

Ví Dụ

y_i	0	1	E
$P(0, y_i)$	$0.99p_0$	0	$0.01p_0$
$P(1, y_i)$	0	$0.99p_1$	$0.01p_1$
$P(y_i)$	$0.99p_0$	$0.99p_1$	0.01

y_i	0	1	E
$P(0 \mid y_i)$	1	0	p_0
$P(1 \mid y_i)$	0	1	p_1

$$H(X \mid Y) = -\sum_{i=1}^{m}\sum_{j=1}^{n} P(x_j, y_i) \log_2 P(x_j \mid y_i)$$

$$\begin{aligned} H(X \mid Y) &= 0.99p_0 \log_2 1 + 0 + 0.01p_0 \log_2 p_0 \\ &\quad + 0 + 0.99p_1 \log_2 1 + 0.01p_1 \log_2 p_1 \\ &= 0.01(p_0 \log_2 p_0 + p_1 \log_2 p_1) \\ &= 0.01H(S). \end{aligned}$$

$$I(X, Y) = H(S) - 0.01H(S) = 0.99H(S).$$

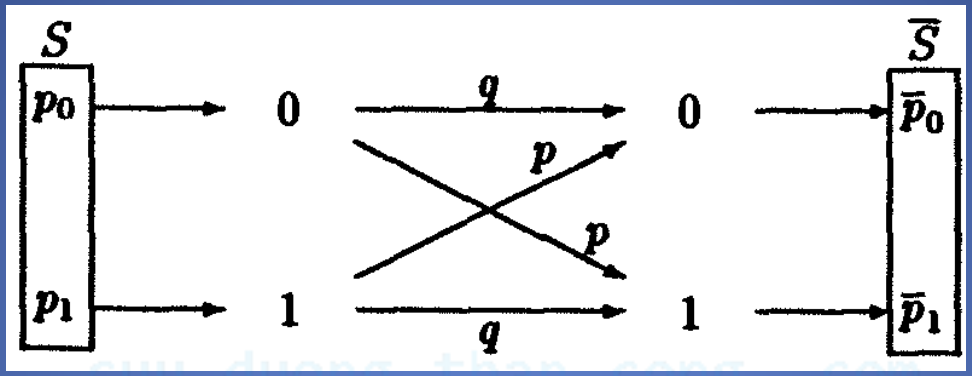
Dung lượng kênh truyền

ĐN: *Dung lượng (capacity) kênh truyền* C là giá trị cực đại của lượng thông tin

$$H(X|Y) = -\sum_{i=1}^{m}\sum_{j=1}^{n} P(x_j, y_i)\log_2 P(x_j|y_i)$$

$$I(X,Y) = H(S) - H(X|Y)$$

Dung lượng kênh truyền nhị phân đối xứng



$$I(X,Y) = H(S) - H(X\,|\,Y) = H(\overline{S}) - H(Y\,|\,X),$$

Định lý: Dung lượng của một kênh nhị phân đối xứng là

$$C = 1 - H(p, 1-p) = 1 + p\log_2 p + (1-p)\log_2(1-p) \quad \text{(bits)}.$$

Ví Dụ 1

y_i	0	1
$P(y_i \mid 0)$	q	p
$P(y_i \mid 1)$	p	q

y_i	0	1
$P(0, y_i)$	$p_0 q$	$p_0 p$
$P(1, y_i)$	$p_1 p$	$p_1 q$

Figure 12: Binary symmetric channel

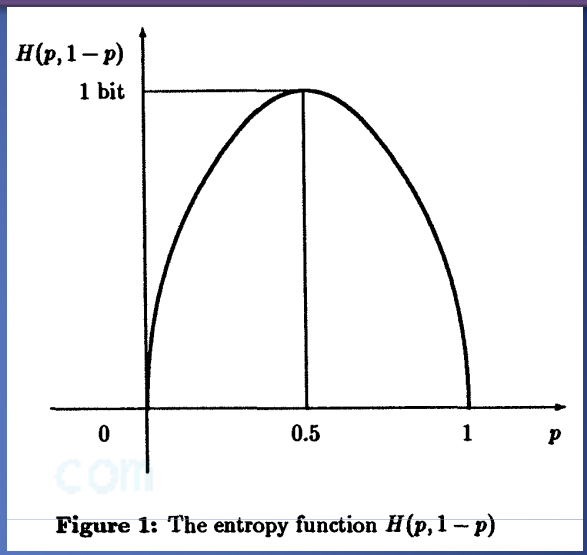


Figure 1: The entropy function $H(p, 1-p)$

- Kênh nhị phân đối xứng với tỷ lệ lỗi p = 0.001 có dung lượng

$$C = 1 + 0.001 \log_2 0.001 + 0.999 \log_2 0.999 \approx 0.989 \text{ bits.}$$

- Kênh nhị phân đối xứng với tỷ lệ lỗi p = 0.5 có dung lượng

$$C = 1 - H(0.5, 0.5) = 0 \text{ bits.}$$

Ví Dụ 2

y_i	0	1	E
$P(0, y_i)$	$0.99p_0$	0	$0.01p_0$
$P(1, y_i)$	0	$0.99p_1$	$0.01p_1$
$P(y_i)$	$0.99p_0$	$0.99p_1$	0.01

y_i	0	1	E
$P(0 \mid y_i)$	1	0	p_0
$P(1 \mid y_i)$	0	1	p_1

$$
\begin{aligned}
H(X \mid Y) &= 0.99p_0 \log_2 1 + 0 + 0.01p_0 \log_2 p_0 \\
&\quad + 0 + 0.99p_1 \log_2 1 + 0.01p_1 \log_2 p_1 \\
&= 0.01(p_0 \log_2 p_0 + p_1 \log_2 p_1) \\
&= 0.01H(S).
\end{aligned}
$$

$$I(X, Y) = H(S) - 0.01H(S) = 0.99H(S).$$

- Giá trị cực đại của H(S) là 1bit, nên có dung lượng C = 0.99 bits.

Định lý cơ bản của Shannon

ĐL: Mọi kênh nhị phân đối xứng với dung lượng $C > 0$ cho trước có thể được mã hoá với độ tin cậy cao và tỷ lệ thông tin gần bằng C. Nói cách khác, tồn tại dãy các mã $K_1, K_2, \ldots$ có độ dài mã tương ứng là 1, 2, … sao cho

$$\lim_{n\to\infty} P_{err}(K_n) = 0 \quad \text{and} \quad \lim_{n\to\infty} R(K_n) = C.$$

ĐL đảo của ĐL Shannon

ĐL: Trong mỗi kênh nhị phân đối xứng với dung lượng C cho trước, nếu mã K_n (độ dài mã bằng n) có tỷ lệ thông tin lớn hơn C, thì mã có xu hướng không tin cậy. Nói cách khác:

$$R(K_n) \geq C + \varepsilon \implies \lim_{n \to \infty} P_{err}(K_n) = 1.$$

Ví Dụ

Kênh nhị phân đối xứng với tỷ lệ lỗi p = 0.001 có dung lượng C = 0.989.

1. Tồn tại mã có độ tin cậy bất kỳ với tỷ lệ thông tin R = 0.9 (< C).
2. Các mã có tỷ lệ thông tin R = 0.99 (> C) không thể có độ tin cậy cao.

Tóm tắt

- Kênh rời rạc không nhớ
- Xác suất, xác suất có điều kiện
- Entropy có điều kiện H(X|Y)
- Lượng thông tin I(X,Y)
- Dung lượng kênh truyền I_{max}.
- ĐL của Shannon

Homework

- Đọc lại và làm bài tập:
  - Chương 4 [1]
  - Chương 1 [2]
- Đọc trước chương 5 [1]

Bài tập 1

Cho một nguồn thông tin nhị phân có P(0) = p, P(1) = q = 1 – p. Giả sử n ký tự được truyền đi.

- CM xác suất một từ nhị phân độ dài n xuất hiện ký tự 0 ở các vị trí $i_1, i_2, \ldots, i_k$, còn lại là các ký tự 1 là $p^k q^{n-k}$.
- Suy ra xác suất một từ xuất hiện ký tự 0 ở k vị trí bất kỳ là

Bài tập 2

Trong một kênh nhị phân đối xứng có tỷ lệ lỗi p = 0.1

1. Tìm độ dài mã lặp K thoả xác suất giải mã lỗi $P_{err}(K) < 10^{-4}$.
2. Tính $P_{err}(K_6^*)$

Information Bits	Code Word
000	000000
100	100011
010	010101
001	001110
011	011011
101	101101
110	110110
111	111000

Bài tập 3

CM lượng thông tin có các tính chất sau

- $I(X,Y) \geq 0$. Dấu = xảy ra khi nào.
- $I(X,Y) \leq H(S)$. Dấu = xảy ra khi nào.