CHƯƠNG 3: SINH MÃ TÁCH ĐƯỢC (Decypherable Coding)

Mục tiêu:

Phân này đề cập đến bài toán mã hóa (coding) các giá trị của một biến X. Khi mã các giá trị của X người ta phải sử dụng bảng ký tự mã (Coding Character Table) hay bảng chữ cái (Code Alphabet). Như vậy, một giá trị x của X sẽ được mã thành một từ mã (Code Word) w dưới dạng một dãy các ký tự mã với độ dài là n ký tự. Trong truyền tin, một dãy các giá trị của X được phát sinh và được mã thành một dãy liên tục các từ mã hay một dãy các ký tự mã lấy từ bảng ký tự mã. Vấn đề cần giải quyết là:

1. Khi nhận một dãy ký tự mã liên tục đó thì ta có thể giải mã thành một dãy các giá trị duy nhất của X hay không ? Nói cách khác, dãy ký tự mã này có tách được thành các từ mã một cách duy nhất hay không ?
2. Chỉ ra phương pháp xây dựng mã tách được tối ưu.

BÀI 3.1: KHÁI NIỆM VỀ MÃ TÁCH ĐƯỢC

Mục tiêu

Sau khi hoàn tất bài học này bạn có thể:

- Biết yêu cầu của bài toán sinh mã,
- Hiểu khái niệm về bảng mã tách được và bảng mã không tách được,
- Hiểu khái niệm về bảng mã tức thời,
- Hiểu giải thuật kiểm tra tính tách được của một bảng mã,
- Vận dụng giải thuật kiểm tra tính tách được của một bảng mã để kiểm tra xem một bảng mã có phải là bảng mã tách được hay không.

Đặt vấn đề bài toán sinh mã

Giả sử nguồn tin X xuất hiện và được ghi lại thông qua một thiết bị đặc biệt. Chẳng hạn như ảnh được ghi lại bằng máy ảnh, âm thanh được ghi lại bằng máy ghi âm, … Qua kênh truyền, những thông tin này cần phải được mã hóa cho phù hợp. Để có thể mã hóa người ta cần một bảng chữ cái gồm các chữ cái quy định trước (chẳng hạn bảng chữ cái la tinh, bảng mã nhị phân, …). Mỗi giá trị của X sau đó được mã dưới dạng một dãy hữu hạn các chữ cái và ta gọi dãy hữu hạn các chữ cái gán cho một giá trị của x là một từ mã.

Ta xét BNN $X=\{x_1, x_2, \ldots, x_n\}$ có phân phối $\{p_1, p_2, \ldots, p_n\}$ được quan sát liên tục và độc lập. Dãy các giá trị nhận được gọi là thông báo (Message) có dạng $x_{i1}x_{i2}\ldots x_{in}$. Tập hợp $A=\{a_1, a_2, \ldots, a_n\}$ là tập hợp ký tự mã (Code Characters) hay là bảng chữ cái (Code Alphabet) dùng để sinh mã. Một giá trị $x_i \in X$ được gán bởi một dãy hữu hạn các ký tự mã được gọi là từ mã (Code word). Tập hợp gồm tất cả các từ mã gán cho tất cả các giá trị của X được gọi là bộ mã hay bảng mã (Code). Các từ mã phải khác nhau từng đôi một.

Bộ mã được gọi là tách được nếu như từ một dãy các ký tự mã nhận được liên tục (được mã hóa từ bộ mã này), ta luôn luôn giải mã được với kết quả duy nhất là dãy các giá trị gốc của X.

Shannon (1948) lần đầu tiên đã đưa ra định lý cơ sở về sinh mã tách được. Mc Millan (1956) đã chứng minh định lý về điều kiện cần và đủ của bảng mã tách được. Nhưng vấn đề sinh mã tách được chỉ được xét một cách chuẩn mực bởi Feinstein (1958), Abramson (1963) và Fano (1961). Sardinas(1960) và Patterson (1963) đã đưa ra định lý về giải thuật kiểm tra tính tách được của một bảng mã. Abramson (1963) đã đưa ra khái niệm bảng mã tức thời.

Trong phạm vi bài giảng này, bài toán sinh mã tối ưu được đặt ra ở đây là tìm ra một phương pháp sinh mã sao cho độ dài trung bình của các từ mã trong bộ mã là nhỏ nhất. Nghĩa là, nếu giá trị x_i được gán bởi từ mã có độ dài n_i thì bài toán sinh mã phải thỏa:

$$\sum_{i=1}^{n} p_i n_i \rightarrow Min$$

Huffman (1950) đã đưa ra qui trình xây dựng một bảng mã tối ưu thỏa yêu cầu này.

Khái niệm về bảng mã không tách được

Bảng mã không tách được là bảng mã mà khi mã hóa thông báo **Msg** ta sẽ nhận được một dãy các từ mã **ws**, và khi giải mã dãy các từ mã **ws** thì ta có thể nhận được nhiều thông báo **Msg** khác nhau.

Ví dụ: Xét biến ngẫu nhiên $X=\{x_1, x_2, x_3, x_4\}$ có bảng mã $W=\{w_1=0, w_2=1, w_3=01, w_4=10\}$.

Giả sử thông báo nguồn có nội dung: $x_1x_2x_3x_4x_3x_2x_1$. Khi đó dãy mã tương ứng viết từ W có dạng: 0101100110.

Nếu giải mã tuần tự từ trái qua phải ta nhận kết quả: $x_1x_2x_1x_2x_2x_1x_1x_2x_2x_1$. Nhưng nếu bằng phương pháp khác ta có thể nhận được kết quả: $x_3x_3x_4x_3x_4$ và nhiều thông báo khác nữa.

Nhận xét: Bảng mã giải mã không tách được là bảng mã mà trong đó tồn tại ít nhất một từ mã này là mã khóa của một hay nhiều từ mã khác trong bộ mã (ví dụ từ mã $w_1=0$ hay $w_2=1$ là mã khóa của w_3).

Bảng mã tách được

Bảng mã tách được là bảng mã mà khi mã hóa thông báo **Msg** ta sẽ nhận được dãy các từ mã **ws**, và khi giải mã dãy các từ mã **ws** thì ta chỉ nhận được một thông báo duy nhất là **Msg** ban đầu.

Ví dụ: Xét biến ngẫu nhiên $X=\{x_1, x_2\}$ có bảng mã tương ứng $W=\{w_1=0, w_2=01\}$.

Phương pháp giải mã được sử dụng như sau: chỉ giải mã khi nào đã nhận được đoạn mã với độ dài bằng độ dài của từ mã dài nhất.

Giả sử dãy mã nhận được (cần giải mã) là: 0010000101001.
Sử dụng phương pháp giải mã trên ta nhận được duy nhất dãy thông báo gốc:

$x_1x_2x_1x_1x_1x_2x_2x_1x_2$.

Có thể chi tiết hóa các bước giải mã dãy từ mã trên như sau:

Nhận được đoạn 00 -> Giải ra x_1 , còn lại 0.
Nhận tiếp 1 ->01 -> Giải ra x_2.
Nhận tiếp 00 -> Giải ra x_1, còn lại 0.

Nhận tiếp 0 -> 00 -> Giải ra x_1, còn lại 0.
Nhận tiếp 0 -> 00 -> Giải ra x_1, còn lại 0.
Nhận tiếp 1 -> 01 -> Giải ra x_2.
Nhận tiếp 01 -> Giải ra x_2.
Nhận tiếp 00 -> Giải ra x_1, còn lại 0.
Nhận tiếp 1 -> 01 -> Giải ra x_2.

Kết quả dãy thông báo là: $x_1x_2x_1x_1x_1x_2x_2x_1x_2$.

Kết luận: Bảng mã tách được là bảng mã mà trong đó không tồn lại từ mã này là mã khóa từ mã khác, tuy nhiên vẫn có thể tồn tại từ mã này là tiền tố (phần đầu) của từ mã kia.

Khái niệm bảng mã tức thời

Bảng mã tức thời là bảng mã mà khi mã hóa thông báo **Msg** ta sẽ nhận được dãy các từ mã **ws**, và khi giải mã dãy các từ mã **ws** thì ta chỉ nhận được một thông báo duy nhất là **Msg** ban đầu. **Abramson đã chứng minh được kết quả sau: Bảng mã tức thời là bảng mã không tồn tại từ mã này là tiền tố của từ mã khác**.

Ví dụ 1: Bảng mã W={w_1=10; w_2=101; w_3=100} không phải bảng mã tức thời vì w_1 là tiền tố của w_2 và w_3.
Ví dụ 2: Bảng mã W={w_1=0, w_2=100, w_3=101, w_4=11} là bảng mã tức thời vì không tồn tại từ mã này là tiền tố của từ mã khác.

Giải thuật kiểm tra tính tách được của bảng mã

Thủ tục sau đây do Sardinas (1960), Patterson (1963) và Abramson (1963) đưa ra nhằm kiểm tra xem một bảng mã nào đó có phải là bảng mã tách được (bảng mã cho phép giải mã duy nhất) hay không.

Input: Bảng mã W
Output: Kết luận bảng mã tách được hay không tách được.

Giải thuật:

Bước khởi tạo: Gán tập hợp S_0=W.
Bước 1: xác định tập hợp S_1 từ S_0:
- Khởi tạo S_1={}
- Với $\forall$ w_i, $w_j \in S_0$, ta xét: nếu w_i=w_jA (w_j là tiền tố của w_i) hoặc w_j=w_i A (w_i là tiền tố của w_j) thì thêm A (phần hậu tố) vào S_1.

Bước k: xác định tập hợp S_k ($k\geq2$) từ tập hợp S_0 và S_{k-1}:
- Khởi tạo: S_k={}
- Với $\forall$ $w_i \in S_0$ và $\forall$ $v_j \in S_{k-1}$, ta xét: nếu w_i=v_jA (v_j là tiền tố của w_i) hoặc v_j=w_i A (w_i là tiền tố của v_j) thì thêm A (phần hậu tố) vào S_k.

Điều kiện để dừng vòng lặp:
Nếu S_k={} thì dừng và kết luận bảng mã tách được ($k\geq1$).
Nếu tồn tại từ mã w_i trong S_k hay $S_k \cap S_0 \neq \varnothing$ thì dừng và kết luận bảng mã không tách được.
Nếu S_k=$S_{t<k}$ thì dừng và kết luận bảng mã tách được ($k\geq1$).

Bài toán 1- yêu cầu

Bài toán: Kiểm tra xem bảng mã W={a, c, ad, abb, bad, deb, bbcde} có phải là bảng mã tách được hay không?

Áp dụng Giải thuật kiểm tra tính tách được của một bảng mã:
Bước khởi tạo: S0={a, c, ad, abb, bad, deb, bbcde}
Bước 1: Tính S1
Khởi tạo S1={}
Vì a là tiền tố của ad nên đưa phần hậu tố "d" vào S1 => S1={d}.
Vì a là tiền tố của abb nên đưa phần hậu tố "bb" vào S1 => S1={d, bb}.
Kiểm tra điều kiện dừng: không thỏa -> qua bước 2.

Bước 2: Tính S2 từ S0 và S1.
Khởi tạo S2={}.
Vì d ∈ S1 là tiền tố của deb ∈ S0 nên đưa phần hậu tố "eb" vào S2
=> S2={eb}
Vì bb∈ S1 là tiền tố của bbcde ∈ S0 nên đưa phần hậu tố "cde" vào S2
=> S2={eb, cde}
Kiểm tra điều kiện dừng: không thỏa -> qua bước 3.
Bài toán 1 - Áp dụng giải thuật
Bước 3: Tính S3 từ S0 và S2.
Khởi tạo S3={}.
Vì c∈ S0 là tiền tố của cde ∈ S2 nên đưa phần hậu tố "de" vào S3
=> S3={de}
Kiểm tra điều kiện dừng: không thỏa -> qua bước 4.
Bước 4: Tính S4 từ S0 và S3.
Khởi tạo S4={}.
Vì de∈ S3 là tiền tố của deb ∈ S0 nên đưa phần hậu tố "b" vào S4
=> S4={b}
Kiểm tra điều kiện dừng: không thỏa -> qua bước 5.

Bước 5: Tính S5 từ S0 và S4.
+ khởi tạo S5={}.
+ Vì b∈ S4 là tiền tố của bad ∈ S0 nên đưa phần hậu tố "ad" vào S5 => S5={ad}
+ Vì b∈ S4 là tiền tố của bbcde ∈ S0 nên đưa "bcde" vào S5
=> S5={ad, bcde}
Kiểm tra điều kiện dừng: Vì S5 có chứa từ mã ad nên dừng lại và kết luận đây là bảng mã không tách được.
Bài toán 2
Bài toán: Kiểm tra xem bảng mã W={010, 0001, 0110, 1100, 00011, 00110, 11110, 101011} có phải là bảng mã tách được không?

Áp dụng Giải thuật kiểm tra tính tách được của một bảng mã:
Bước khởi tạo và bước 1
- Tập hợp S_0={010, 0001, 0110, 1100, 00011, 00110, 11110, 101011}
- Tập hợp S_1 ={1}
Dành cho sinh viên tự làm các bước tiếp theo.

Kết quả gợi ý:
Tập hợp S_2 ={100, 1110, 01011}
Tập hợp S_3={11}

Tập hợp S_4={00, 110}
Tập hợp S_5={01, 0, 011, 110}
Tập hợp S_6={0, 10, 001, 110, 0011, 0110}
Tập hợp S_6 chứa từ mã 0110 nên bảng mã này không phải là bảng mã tách được.

Bài tập

1. Hãy cho biết bảng mã sau có phải là bảng mã tách được hay không?
   W={w_1=00, w_2=01, w_3=0010, w_4=0111, w_5=0110}
2. Hãy lấy ví dụ một bảng mã tách được, và chứng minh nó là bảng mã tách được.

TaiLieu.vn

BÀI 3.2: QUAN HỆ GIỮA MÃ TÁCH ĐƯỢC VÀ ĐỘ DÀI MÃ

Mục tiêu

Sau khi hoàn tất bài học này bạn có thể hiểu:

- Định lý Kraft (1949),
- Định nghĩa cây bậc D cỡ K,
- Vấn đề sinh mã cho cây bậc D cỡ K,
- Vận dụng định lý Kraff để kiểm tra sự tồn tại bảng mã tách được và sinh bảng mã tách được.

Định lý Kraftn(1949).

Gọi $X=\{x_1, x_2,\ldots, x_M\}$ là biến ngẫu nhiên chứa các giá trị cần truyền có phân phối là $P=\{p_1, p_2, \ldots, p_M\}$.

$A=\{a_1, a_2,\ldots,a_D\}$ là bộ ký tự sinh mã có D chữ cái (D được gọi là cơ số sinh mã).
Giá trị x_i được mã hóa thành từ mã w_i có độ dài là n_i.
Đặt $N=\{n_1, n_2,\ldots,n_M\}$ là tập hợp độ dài các từ mã.

Định lý (Kraft- 1949):

Điều kiện cần và đủ để tồn tại bảng mã tức thời với độ dài $N=\{n_1,n_2,\ldots,n_M\}$ là

$$\sum_{i=1}^{M} D^{-n_i} \le 1$$

Ví dụ 1: Bộ mã $W=\{w_1, w_2, w_3\}$ với $M=3$; $n_1=1$; $n_2=2$; $n_3=3$; $D=2$

$$\sum_{i=1}^{M} D^{-n_i} = \frac{1}{2^1}+\frac{1}{2^2}+\frac{1}{2^3}=\frac{7}{8}<1$$

=> Tồn tại bảng mã tức thời.

Ví dụ 2: Bộ mã $W=\{w_1, w_2, w_3\}$ với $M=3$; $n_1=n_2=1$; $n_3=2$; $D=2$

$$\sum_{i=1}^{M} D^{-n_i} = \frac{1}{2^1}+\frac{1}{2^1}+\frac{1}{2^2}=\frac{5}{4}>1$$

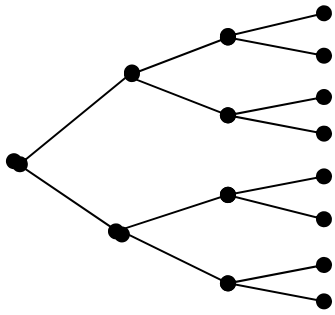
=> Không tồn tại bảng mã tức thời.
<u>Đề nghị</u>: sinh viên tìm hiểu nội dung tiếp theo và trở lại giải thích 2 ví dụ trên.

Định nghĩa cây bậc D cỡ k.

Định nghĩa: Cây bậc D cỡ k là cây có hệ thống nút, cạnh thỏa điều kiện:

- Từ 1 nút có số cạnh đi ra không vượt quá D hay một nút có không quá D nút con.
- Nút cuối cùng (Nút lá) cách nút gốc không vượt quá k cạnh.

Ví dụ: cây bậc D=2 và cỡ k=3

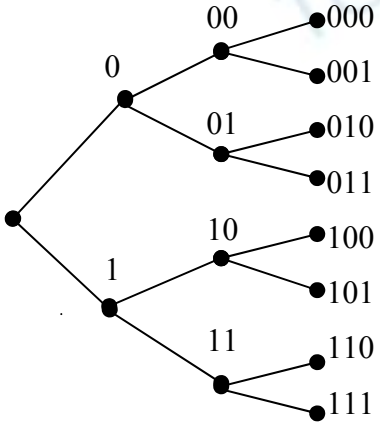


Vấn đề sinh mã cho cây bậc D cỡ k

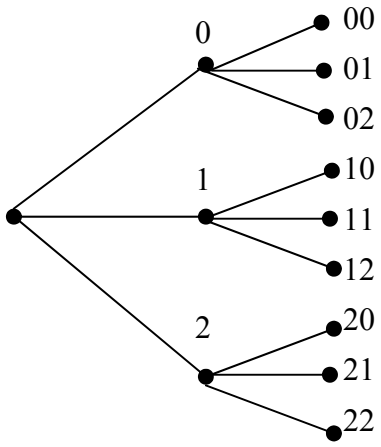
Sinh mã cho các nút của cây bậc D cỡ K (trừ nút gốc):

Để đơn giản hóa: mỗi nút (trừ nút gốc) được ký hiệu bởi dãy ký hiệu của nút cha làm tiền tố + một ký tự bổ sung lấy từ tập hợp {0, 1, 2, …, D-1} thay cho bảng chữ cái A={a_1, a_2, …, a_D}.

Ví dụ 1: Cây bậc D=2 cỡ k=3



Ví dụ 2: Cây bậc D=3 cỡ k=2.



Tính chất:

+ Các nút (trừ nút gốc) của cây đều được mã hóa từ bảng chữ cái {0, 1, 2,…, D-1}
+ Mỗi nút (đã mã hóa) có mã của nút kề trước là tiền tố.
+ Tổng số các nút lá bằng D^k = tổng số các mã tức thời có thể có.

Chứng minh định lý Kraft (Điều kiện cần)

Giả sử, cho trước bảng mã tức thời W={w_1, w_2,…, w_M} với N={$n_1 \le n_2 \le \ldots \le n_M$}. Ta cần c/m:

$$\sum_{i=1}^{M} D^{-n_i} \le 1$$

Xây dựng cây bậc D cỡ n_M và sinh mã cho các nút trừ nút gốc với các ký tự mã lấy từ bảng chữ cái A = {0, 1, 2,…, D-1}. Mã tại mỗi nút (trừ nùt gốc) đều có khả năng được chọn là từ mã.

Như vậy, ta tiến hành chọn các từ mã cho bảng mã tức thời với qui tắc là: một nút nào đó được chọn để gán một từ mã thì tất cả các nút kề sau nút gán từ mã phải được xóa. Cụ thể như sau:

Chọn một nút có mã với độ dài mã là n_1 gán cho nó một từ mã w_1.

=> Tổng số nút lá được xóa tương ứng là $D^{n_M - n_1}$

Chọn một nút có mã với độ dài mã là n_2 gán cho nó một từ mã w_2.

=> Tổng số nút lá được xóa tương ứng là $D^{n_M - n_2}$

……

Chọn một nút có mã với độ dài mã là n_n gán cho nó một từ mã w_n.

=> số nút lá được gán từ mã là $D^{n_M - n_M}$

Vậy số nút lá bị xóa hoặc được gán từ mã là:

$$=> D^{n_M - n_1} + D^{n_M - n_1} + \cdots + D^{n_M - n_M} = \sum_{i=1}^{M} D^{n_M - n_i} \le D^{n_M} = \text{tổng số nút lá.}$$

$$=> \sum_{i=1}^{M} D^{-n_i} \le 1 \text{ (đpcm)}$$

Chứng minh định lý Kraft (Điều kiện đủ)

Giả sử: $\sum_{i=1}^{M} D^{-n_i} \le 1$, để cần chứng minh tồn tại bảng mã tức thời với $N=\{n_1, n_2, \ldots, n_M\}$, ta chỉ cần chỉ ra thủ tục xây dựng bảng mã tức thời như sau:

Thủ tục tạo mã tức thời:

Xét $N=\{n_1, n_2, \ldots, n_M\}$ và cơ số sinh mã là D**:**

Bước 1: Ta xếp thứ tự $n_1 \le n_2 \le \ldots \le n_M$, xây dựng cây bậc D cỡ $k=n_M$ và sinh mã cho các nút .

Bước 2: Chọn nút bất kỳ trên cây có độ dài n_1 gán cho từ mã w_1 và xóa tất cả các nút kề sau nó.

Bước 3: Lặp lại các bước 2 đối với việc chọn các từ mã còn lại $w_2, \ldots, w_M$ ứng với $n_2, \ldots, n_M$.

=> Bảng mã $W=\{w_1, w_2, \ldots, w_M\}$ là bảng mã tức thời.

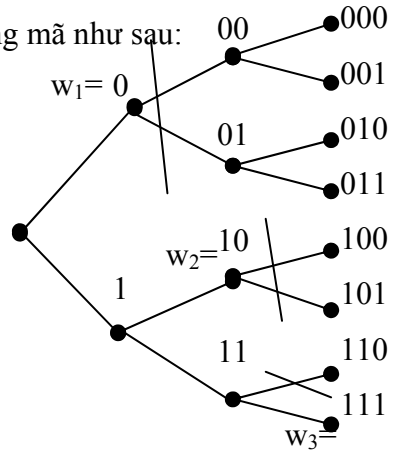
Ví dụ minh họa định lý Kraft

Ví dụ 1: Xét bảng mã thỏa M=3, D=2, $n_1=1$, $n_2=2$, $n_3=3$. Vậy ta kiểm tra xem có tạo được bảng mã tức thời hay không?

Ta có $\sum_{i=1}^{3} 2^{-n_i} = 2^{-1} + 2^{-2} + 2^{-3} = \frac{7}{8} < 1$

=> $W= \{w_1, w_2, w_3\}$ là bảng mã tức thời

Ta Xây dựng bảng mã như sau:



- Chọn $w_1=0$, cắt bỏ các nút con của nút w_1.
- Chọn $w_2=10$, cắt bỏ các nút con của nút w_2.
- Chọn $w_3=111$

Chú ý: ngoài bảng mã tức thời chọn được ở trên, ta còn có thể sinh được nhiều bảng mã tức thời khác. *Đề nghị sinh viên đưa ra bảng mã tức thời khác.*

Bài tập

1. Tìm 1 bảng mã tách được thỏa tính chất D = 2, k = 4?
2. Tìm tất cả các bảng mã tách được thỏa tính chất D=2, k=3?
3. Hãy chỉ ra bảng mã sau đây là bảng mã không tách được:
   W={w1=00, w2=1, w3=100, w4=110, w5=111}
4. Hãy tìm một bảng mã nhị phân tách được có ít nhất 5 từ mã thỏa điều kiện
   $$\sum_{i=1}^{M} D^{-n_i} = 1$$

TaiLieu.vn

BÀI 3.3: TÍNH TỐI ƯU CỦA ĐỘ DÀI MÃ

Mục tiêu

Sau khi hoàn tất bài học này bạn có thể:

- Hiểu định lý Shannon (1948),
- Biết được các tiêu chuẩn đánh giá bảng mã tối ưu tuyệt đối và bảng mã tối ưu tương đối,
- Điều kiện nhận biết một bảng mã tối ưu,
- Hiểu Định lý Huffman,
- Biết Phương pháp sinh mã Huffman,
- Vận dụng phương pháp sinh mã Huffman để sinh mã Huffman cho một thông báo,
- Vận dụng phương pháp sinh mã Huffman để viết chương trình nén.

Định lý Shannon (1948)

Phát biểu định lý:

Đặt $\bar{n} = \sum_{i=1}^{M} p_i n_i$ là độ dài trung bình của bảng mã.

Khi đó $\bar{n} \geq \dfrac{H(X)}{\log_2 D}$

Dấu đẳng thức xảy ra khi và chỉ khi $p_i = D^{-n_i}$ hay $\sum_{i=1}^{M} D^{-n_i} = 1$

Diễn giải: Đối với mã tách được độ dài trung bình của mã sẽ có cận dưới là $\dfrac{H(X)}{\log_2 D}$. Nếu mã không tách được độ dài trung bình của nó có thể nhỏ hơn cận dưới. Nếu mã tách được không tối ưu thì độ dài của nó sẽ lớn hơn nhiều so với cận dưới, còn nếu mã tách được tối ưu thì độ dài trung bình của nó gần với cận dưới.

Bài toán đặt ra sẽ là tìm phương pháp xây dựng bảng mã tách được tối ưu.

Chú ý:

$$H_D(X) = -\sum p_i \log_D p_i$$

$$H_D(X) = \frac{H(X)}{\log_2 D} = \frac{-\sum p_i \log_2 p_i}{\log_2 D}$$

là entropy của X với cơ số D.

Bảng mã tối ưu tuyệt đối

Định lý: Bảng mã được gọi là tối ưu tuyệt đối khi $\bar{n} = \dfrac{H(X)}{\log_2^D}$ hay $p_i = D^{-n_i}$

Ví dụ: xét biến ngẫu nhiên X={x_1, x_2, x_3, x_4}

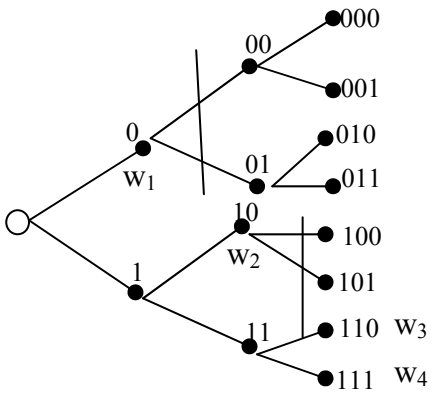
Có phân phối: P={1/2, 1/4, 1/8, 1/8}

Có bảng mã W={w_1= 0, w_2=10, w_3=110, w_4=111}

Ta tính được độ dài trung bình từ mã: $\bar{n} = \frac{1}{2}*1 + \frac{1}{4}*2 + \frac{1}{8}*3 + \frac{1}{8}*3 = \frac{12}{8} = 1.75$

Tính Entropy của X: H(X)= H(0.5, 0.25, 0.125, 0.125) = 0.5 +0.5 + 0.375 + 0.375 =1.75

Log_2D=1.



W= $\{w_1, w_2, w_3, w_4\}$ là bảng mã tối ưu tuyệt đối vì thỏa điều kiện:

$$\bar{n} = \frac{H(X)}{\log_2^D}$$

Bảng mã tối ưu tương đối

Định lý: Bảng mã được gọi là tối ưu tương đối khi: $\frac{H(X)}{\log_2 D} \le \bar{n} < \frac{H(X)}{\log_2 D} + 1$

Điều kiện nhận biết một bảng mã tối ưu

Định lý (với D=2):

- Xác suất p_i càng lớn thì độ dài n_i của từ mã w_i càng nhỏ.
- Giả sử $p_1 \ge p_2 \ge \ldots \ge p_M$. Nếu $p_i \ge p_{i+1} \ge p_{i+r}$ thì $n_i \le n_{i+1} \le n_{i+r}$ thì 2 từ mã tương ứng với 2 giá trị có xác suất nhỏ nhất có độ dài mã bằng nhau $n_{M-1} = n_M$.
- Trong các từ mã có độ dài bằng nhau và cùng bằng n_M (dài nhất) thì tồn tại ít nhất 2 từ mã w_{M-1} và w_M có M-1 ký tự đầu giống nhau và ký tự thứ M khác nhau.

Ví dụ: xét bảng mã $W=\{w_1=0, w_2=100, w_3=101, w_4=1101, w_5=1110\}$
Bảng mã trên không phải là bảng mã tối ưu vì 2 từ mã w_4, w_5 có độ dài lớn nhất =4 ký tự nhưng 3 ký tự đầu khác nhau.

Ghi chú: D > 2 được xét tương tự.

Định lý Huffman

Định lý: Giả sử X có phân phối xác suất với thứ tự giảm dần sau:

X	x_1	x_2	…	x_M
P	$p_1 \ge$	$p_2 \ge$	…	$\ge p_M$

Giả sử bảng mã của X là $W=\{w_1, w_2, \ldots, w_{M-1}, w_M\}$.
Đặt $x_{M-1,M}=\{x_{M-1}, x_M\}$ có xác suất là $p_{M-1,M}=p_{M-1}+p_M$.
và $X^* = \{x_1, x_2, \ldots, x_{M-1,M}\}$ có phân phối sau:

X*	x_1	x_2	…	x^*_{M-2}	$x^*_{M-1,M}$
P	P_1	p_2	…	p^*_{M-2}	$p^*_{M-1,M}$

Giả sử $W^* = \{w_1, w_2, \ldots, w_{M-2}, x^*_{M-1,M}\}$ là bảng mã tối ưu của X*. Khi đó:
- $w_{M-1} = w^*_{M-1,M}$ + "0".
- $w_M = w^*_{M-1,M}$ + "1".

Phương pháp sinh mã Huffman

Giả sử X có phân phối xác suất với thứ tự giảm dần sau:

X	x_1	x_2	…	x_M
P	$p_1 \geq$	$p_2 \geq$	…	$\geq p_M$

Thủ tục lùi (D=2):

Khởi tạo: Đặt M_0=M

Bước 1:

- Đặt $x_{M_0-1,M_0} = \{x_{M_0-1}, x_{M_0}\}$ có xác suất $p_{M_0-1,M_0} = p_{M_0-1} + p_{M_0}$
- Sắp xếp lại theo tứ tự giảm dần của xác suất ta nhận được dãy phân phối mới có M_0-1 phần tử như sau: $p_1,\ p_2, \cdots, p_{M_0-2},\ p_{M_0-1,M_0}$

Bước 2: Lặp lại bước 1 với sự lưu vết

$$w_{M_0-1} = w_{M_0-1,M_0} + "0"$$

$$w_{M_0} = w_{M_0-1,M_0} + "1"$$

Giảm M_0: M_0=M_0-1, vòng lặp kết thúc khi M_0=2

(**Chú ý:** trong trường hợp tổng quát, vong lặp kết thúc khi $M_0 \leq$ D.)

Thủ tục tiến:

Đi ngược lại so với thủ tục lùi đồng thời xác định từ mã ở mỗi bước từ sự lưu vết ở thủ tục lùi.

Minh họa phương pháp sinh mã Huffman

Ví dụ 1: sinh bảng mã nhị phân Huffman cho X có phân phối sau:

X	x_1	x_2	x_3	x_4	x_5	x_6
P	0.3	0.25	0.2	0.1	0.1	0.05

Thủ tục lùi:

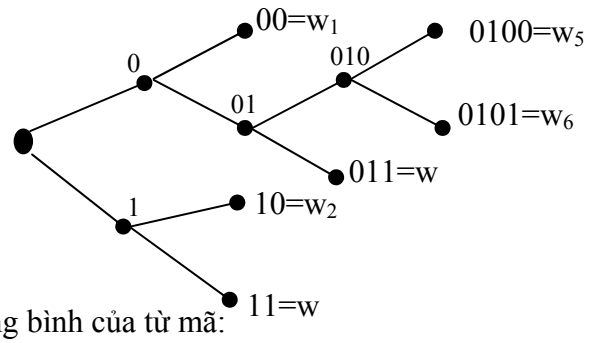
Bước 1		**Bước 2**		**Bước 3**		**Bước 4**		**Bước 5**		
X	**P**	**X**	**P**	**X**	**P**	**X**	**P**	**X**	**P**	
x_1	0.3	x_1	0.3	x_1	0.3	x_{23}	0.45	x_{1564}	0.55	0 ←
x_2	0.25	x_2	0.25	x_{564}	0.25	x_1	0.3	x_{23}		0.45 ← 1
x_3	0.2	x_3	02	x_2	0,25	x_{564}	0.25			
x_4	0.1	x_{56}	0.15	x_3	0,2					
x_5	0.1	x_4	0.1							
x_6	0.05									

Thủ tục tiến:

Bước 1		**Bước 2**		**Bước 3**		**Bước 4**		**Bước 5**	
X	**W**	**X**	**W**	**X**	**W**	**X**	**W**	**X**	**W**
x_{1564}	0	x_{23}	1	x_1	00	x_1	00	x_1	00 = w_1
x_{23}	1	x_1	00	x_{564}	01	x_2	10	x_2	10 = w_2
		x_{564}	01	x_2	10	x_3	11	x_3	11 = w_3
				x_3	11	x_{56}	010	x_4	011 = w_4
						x_4	011	x_5	0100 = w_5
								x_6	0101 = w_6

Nhận xét tính tối ưu của bảng mã Huffman

Vẽ cây Huffman của bảng mã trên:



Độ dài trung bình của từ mã:

$\bar{n}$ =(0.3 x 2)+ (0.25 x 2)+ (0.2 x 2) + (0.1 x 3) +(0.1 x 4) + (0.05 x 4) = 2.4

Entropy của X: H(X) = H(0.3, 0.25; 0.2, 0.1,0.1, 0.05)
= 2.4

Nhận xét: Do D = 2 và $\log_2 D=1$, Ta có $\bar{n}$ = H(X) nên bảng mã trên tối ưu tuyệt đối.

Bài tập

1. Cho biến ngẫu nhiên X có phân phối sau:

X	x_1	x_2	x_3	x_4
P	0.4	0.3	0.2	0.1

2. Cho biến ngẫu nhiên Y có phân phối sau:

Y	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9
P	0.3	0.2	0.2	0.1	0.05	0.05	0.04	0.03	0.03

3. Cho đoạn văn bản "thoi the thi thoi thi the thoi thi the". Tìm bảng mã nhị phân Huffman dùng để mã hóa đoạn văn bản trên.
4. Thay từng ký tự trong đoạn văn bản trên thành một từ mã, cắt từng đoạn 8 bits đổi thành số thập phân. Cho biết dãy số thập phân kết quả.

TaiLieu.vn

CHƯƠNG 4: KÊNH TRUYỀN

Mục tiêu:

Trình bày mô hình truyền tin rời rạc từng ký tự mã độc lập lẫn nhau (phù hợp với đặc điểm của kênh). Mô hình này còn gọi là kênh truyền rời rạc không nhớ (Memoryless Discret Channel). Từ mô hình này người ta có thể xây dựng cách tính dung lượng kênh truyền và phương pháp phân loại đầu nhận để có thể giải mã tốt nhất.

BÀI 4.1: KÊNH TRUYỀN RỜI RẠC KHÔNG NHỚ

Mục tiêu

Sau khi hoàn tất bài học này bạn có thể:

- Biết mô hình kênh truyền tin rời rạc không nhớ ở 2 khía cạnh vật lý và toán học.
- Khái niệm về lượng tin trên kênh truyền
- Định nghĩa dung lượng kênh truyền

Giới thiệu

Trước hết, ta có thể hiểu khái niệm kênh truyền rời rạc và không nhớ ở bài học này như sau: khái niệm truyền rời rạc ở đây là truyền tuần tự các ký tự độc lập nhau (hay truyền từng ký tự một), còn khái niệm không nhớ ở đây là chỉ xét mối quan hệ giữa ký tự truyền và ký tự nhận được tương ứng, không xét đến mối quan hệ giữa ký tự nhận được với ký tự nhận được trước đó.

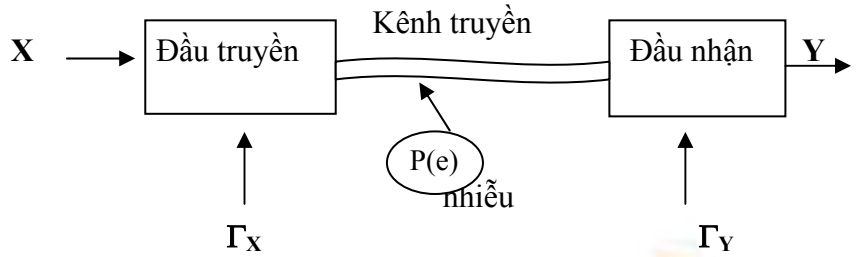
Khái niệm về một kênh truyền rời rạc dựa vào phân bố xác suất của tín hiệu ra phụ thuộc vào tín hiệu vào và trạng thái của kênh truyền đã được chuẩn hóa bởi Feinstein (1958) và Wolfowitz (1961). Dung lượng kênh (Channel Capacity) được xác định chính xác nhờ Muroya (1953) và Fano (1961). Giải thuật và chương trình tính dung lượng kênh đã được viết bởi Eisenberg (1963).

Định lý cơ bản về truyền tin đã chỉ ra rằng "với dung lượng kênh cho trước luôn có thể tìm ra một phương pháp truyền tin với lượng tin nhỏ hơn dung lượng kênh và đạt sai số nhỏ hơn sai số cho phép bất kỳ". Định lý cơ bản về truyền tin đã được Feinstein (1954, 1958) khảo sát. Các nhà khoa học Blackwell, Breinan (1958, 1959) và Thomasian (1961) đã lần lượt chỉnh lý để đạt chuẩn tốt hơn. Trong các nội dung tiếp theo của bài học, các bạn sẽ tìm hiểu về mô hình kênh truyền tin rời rạc không nhớ ở khia cạnh vật lý và toán học. Đặc biệt ở mô hình toán học sẽ chỉ ra cách xác định phân phối ở đầu ra dựa vào phân phối ở đầu vào.

Mô hình vật lý

Một thông báo được cấu tạo từ các ký hiệu của một bảng chữ cái ở đầu truyền (input) và được truyền trên kênh. Thông báo được nhận ở cuối kênh (hay đầu nhận-output) và được giải mã theo bảng chữ cái ở đầu truyền. Mặt khác, từng ký tự ở đầu nhận có thể quan hệ với các ký tự ở đầu nhận trước đó, các ký tự ở đầu truyền và trạng thái của kênh truyền. Để đơn giản, ở đây chúng ta chỉ xét mô hình vật lý như sau:

Xét từng ký tự ở đầu nhận chỉ phụ thuộc vào ký tự ở đầu truyền tương ứng với nó, nếu kênh truyền có nhiễu thì một ký tự ở đầu truyền có thể được diễn giải (nhiễu) ra nhiều ký tự khác nhau ở đầu nhận và do đó tạo ra một phân phối xác suất có điều kiện cho ký tự ở đầu nhận. *Mô hình truyền tin rời rạc không nhớ là mô hình truyền tin chỉ xét mối quan hệ giữa ký tự truyền và ký tự nhận được tương ứng, không xét mối quan hệ giữa ký tự nhận được và ký tự nhận được trước đó.*
Mô hình:



Các qui ước:

- X: là biến ngẫu nhiên có giá trị cần truyền ở đầu truyền.
- Y: là biến ngẫu nhiên chứa giá trị có thể nhận được ở đầu nhận.
- Γ_X: là bảng chữ cái sinh mã ở đầu truyền.
- Γ_Y: là bảng chữ cái giải mã ở đầu nhận.
- X, Y, Γ_X, Γ_Y: đều hữu hạn và rời rạc.
- Truyền rời rạc từng ký tự và nhận cũng rời rạc từng ký tự.
- Ký tự nhận sau không phụ thuộc vào ký tự nhận trước.

Mô hình toán học

Ta gọi:

- $\Gamma_X=\{x_1, x_2, \ldots, x_M\}$ là bộ ký tự sinh mã ở đầu truyền (input).
- $\Gamma_Y=\{y_1, y_2,\ldots,y_L\}$ là bộ ký tự giải mã ở đầu nhận (output).
- Biến ngẫu nhiên X lấy giá trị (đã mã hóa) trên Γ_X và có phân phối $p(X=x_i)=p(x_i)$ với $i=1,..,M$.
- Biến ngẫu nhiên Y lấy giá trị (giải mã) trên Γ_Y và có phân phối xác suất có điều kiện: $P(Y=y_j/X=x_i)=p(y_j/x_i)=p_{ij}$ với $j=1,..,L$.

Gọi $A=||p_{ij}||$ là ma trận truyền tin hay mô hình truyền tin của kênh truyền rời rạc không nhớ.
Với $i=\overline{1,M}$, $j=\overline{1,L}$ và $p_{ij} = p(Y=y_j/X=x_i) = p(y_j/x_i)$ là xác suất nhận được giá trị y_j khi đã truyền giá trị x_i.

Tính phân phối đầu nhận:

Ta có: $p(Y=y_j) = p(y_j) = \sum_{i=1}^{M} p(x_i).p(y_j / x_i)$

$$\Rightarrow p(yj) = \sum_{i=1}^{M} p(x_i).p(y_j / x_i)$$

$$= \sum_{i=1}^{M} p(x_i).p_{ij}$$

Vậy $\mathbf{p(y_j)= P_X^{'} .A_j}$ **(1)**

Một các tổng quát: $\mathbf{P^{'}_Y = P^{'}_X.A}$ **(2)**
Trong đó:

- A_j là cột thứ j của A
- $P'_X = [p(x_1), p(x_2),\ldots, p(x_M)]$.
- $P'_Y = [p(y_1), p(y_2),\ldots, p(y_M)]$.

Ví dụ xác định phân phối đầu nhận

Cho ma trận truyền tin như sau:

$$A = \begin{array}{c} x_1 \\ x_2 \\ x_3 \\ \\ \end{array}\begin{bmatrix} 0.5 & 0.2 & 0.3 \\ 0.3 & 0.5 & 0.2 \\ 0.2 & 0.3 & 0.5 \end{bmatrix}$$
$$\begin{array}{ccc} y_1 & y_2 & y_3 \end{array}$$

Xác suất truyền: $p(x_1)=0.5$ và $p(x_2)=p(x_3)= 0.25$.

Ta tìm phân phối của Y :
Ta có: $P_X' =(0.5, 0.25, 0.25)$
Áp dụng công thức (1) ở trên ta được:

$p(y_1) = P_X'.A_1 = 0.375$
$p(y_2) = P_X'.A_2 = 0.3$
$p(y_3) = P_X'.A_3 = 0.325$

$\Rightarrow PY' =(0.375, 0.3, 0.325)$

Lượng tin trên kênh truyền

Ví dụ: cho ma trận truyền tin như sau:

$$A = \begin{array}{c} x_1 \\ x_2 \\ x_3 \\ \\ \end{array}\begin{bmatrix} 0.5 & 0.2 & 0.3 \\ 0.3 & 0.5 & 0.2 \\ 0.2 & 0.3 & 0.5 \end{bmatrix}$$
$$\begin{array}{ccc} y_1 & y_2 & y_3 \end{array}$$

Xác suất truyền: $p(x_1)=0.5$ và $p(x_2)=p(x_3)= 0.25$.
$X = \{x_1, x_2, x_3\}$ được xem như tập các ký tự truyền và $Y =\{y_1, y_2, y_3\}$ là tập các ký tự nhận.

Tính lượng tin trên kênh truyền:
Ta tìm phân phối của Y :
Ta có: $P_X' =(0.5, 0.25, 0.25)$
Áp dụng công thức (1) ở trên ta được:

$p(y_1) = P_X'.A_1 = 0.375$
$p(y_2) = P_X'.A_2 = 0.3$
$p(y_3) = P_X'.A_3 = 0.325$

$\Rightarrow P_Y' =(0.375, 0.3, 0.325)$

Tính các Entropy:

$H(Y) = H(0.375, 0.3, 0.325) = 1.58$ (bit)
$H(Y/X=x_1) = H(0.5, 0.2, 0.3)= 1.49$ (bit)
$H(Y/X=x_2) = H(0.3, 0.5, 0.2)= 1.49$ (bit)
$H(Y/X=x_1) = H(0.2, 0.3, 0.5)= 1.49$ (bit)
$H(Y/X)= p(x_1).H(Y/X=x_1) + p(x_2).H(Y/X=x_2) + p(x_3).H(Y/X=x_3) = 1.49$ (bit)

Lượng thông tin truyền trên kênh: $I (X/Y)= I (Y/X)= H(Y) - H(Y/X) = 0,09$ (bit)

Định nghĩa dung lượng kênh truyền

Dựa vào ma trận truyền tin A, ta có thể dễ dàng tính lượng tin trên kênh truyền.

I(X/Y)= H(X)-H(Y/X)

= H(Y)-H(X/Y)

= I(Y/X)

Ta có I(X/Y)= H(Y)-H(Y/X), trong đó:

H(Y)= H(P_X .A) phụ thuộc vào P_X.

H(Y/X) phụ thuộc vào P_X

Vậy: I(Y/X) phụ thuộc hoàn toàn vào P_X và do đó I(Y/X) có thể đạt Max với P_X xác định nào đó.

Ta định nghĩa: $C = \underset{\forall p(X)}{Max}\, I(X/Y)$ là dung lượng của kênh truyền (ĐVT: bit).

TaiLieu.vn

BAI 4.2: CÁC DẠNG KÊNH TRUYỀN

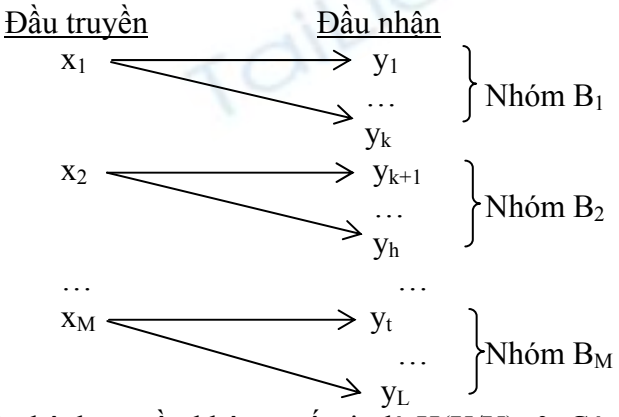
Mục tiêu

Sau khi hoàn tất bài học này bạn có thể:

- Biết kênh truyền không mất tin,
- Biết kênh truyền xác định,
- Biết kênh truyền không nhiễu,
- Biết kênh truyền không sử dụng được,
- Hiểu kênh truyền đối xứng,

Hiểu định lý về dung lượng kênh truyền,Kênh truyền không mất tin

Mô hình: từ tập hợp các giá trị có thể nhận được ở đầu nhận $Y=\{y_1, y_2, \ldots, y_L\}$ được phân thành M nhóm B_i tương ứng với các giá trị x_i ở đầu truyền và xác suất để truyền x_i với điều kiện đã nhận y_j là $p(X= x_i /Y=y_j \in B_i)=1$ (với $M < L$).

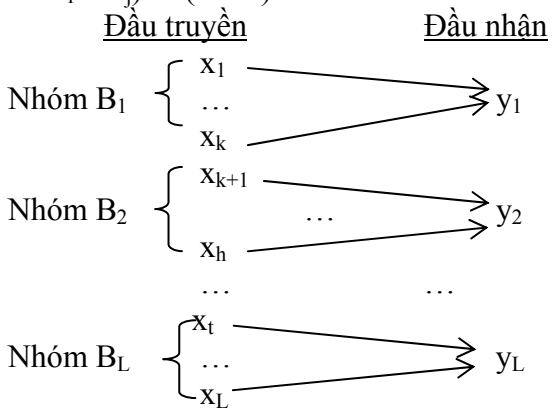


Đặc trưng của kênh truyền không mất tin là H(X/Y)=0. Có nghĩa là lượng tin chưa biết về X khi nhận Y là bằng 0 hay ta có thể hiểu khi nhận được Y thì ta hoàn toàn có thể biết về X.

Dung lượng: $C=\log_2 M$ (Sinh viên tự chứng minh, xem như bài tập)

Kênh truyền xác định

Mô hình: từ tập hợp các giá trị có thể truyền ở đầu truyền được phân thành L nhóm B_j tương ứng với các giá trị có thể nhận được y_j ở đầu nhận và xác suất để nhận y_j với điều kiện đã truyền x_i là $p(Y=y_j/X=x_i \in B_j)=1$ (M>L).



Đặc trưng: của kênh truyền xác định là H(Y/X)=0. Có nghĩa là lượng tin chưa biết về Y khi truyền X bằng 0 hay khi truyền X thì ta biết sẽ nhận được Y.

Dung lượng: $C=\log_2 L$ (Sinh viên tự chứng minh, xem như bài tập)

Kênh truyền không nhiễu

Mô hình: là sự kết hợp của kênh truyền xác định và kênh truyền không mất thông tin, truyền ký tự nào sẽ nhận được đúng ký tự đó.

Đầu truyền		Đầu nhận
x_1	$\longrightarrow$	x_1
x_2	$\longrightarrow$	x_2
…		…
x_M	$\longrightarrow$	x_M

Đặc trưng: H(X/Y)=H(Y/X)=0.

Dung lượng: $C=\log_2 L=\log_2 M$ *(Sinh viên tự chứng minh, xem như bài tập)*

Ví dụ: ma trận truyền tin của kênh truyền không nhiễu với M=L=3:

$$A=\begin{array}{c} x_1 \\ x_2 \\ x_3 \\ \\ \end{array}\begin{array}{c}\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \\ y_1 \ y_2 \ y_3 \end{array}$$

Kênh truyền không sử dụng được.

Mô hình: là kênh truyền mà khi truyền giá trị nào thì mất giá trị đó hoặc xác suất nhiễu thông tin trên kênh truyền lớn hơn xác suất nhận được.

Đặc trưng: H(X/Y)=H(Y/X)= max

Dung lượng: C=0 (Sinh viên tự chứng minh, xem như bài tập)

Ví dụ: kênh truyền có ma trận truyền tin như sau:

$$A=\begin{pmatrix} \varepsilon & 1-\varepsilon \\ \varepsilon & 1-\varepsilon \end{pmatrix}$$

Kênh truyền đối xứng

Mô hình: là kênh truyền mà ma trận truyền tin có đặc điểm sau:

+ Mỗi dòng của ma trận A là một hoán vị của phân phối $P=\{p'_1, p'_2, \ldots, p'_L\}$
+ Mỗi cột của ma trận A là một hoán vị của $Q=\{q'_1, q'_2, \ldots, q'_M\}$

Ví dụ: cho kênh truyền đối xứng có ma trận truyền tin như sau:

$$A = \begin{array}{c} x_1 \\ x_2 \\ x_3 \\ \\ \end{array}\begin{array}{c}\begin{bmatrix} 1/2 & 1/3 & 1/6 \\ 1/3 & 1/6 & 1/2 \\ 1/6 & 1/2 & 1/3 \end{bmatrix} \\ y_1 \quad y_2 \quad y_3 \end{array}$$