Probabilidade e Teoria da Informação

Bartolomeu F. Uchôa Filho, Ph.D

Grupo de Pesquisa em Comunicações – GPqCom
Departamento de Engenharia Elétrica
Universidade Federal de Santa Catarina
E-mail: `uchoa@eel.ufsc.br`

Teoria da Informação

1. Introdução
2. Sistema de Comunicação Digital
   - Codificação de Fonte
   - Codificação de Canal
3. Fontes Discretas sem Memória
4. Medidas de Informação
   - A Medida de Hartley
   - A Medida de Shannon
5. Entropia, Entropia Conjunta e Entropia Condicionada

6. Canais Discretos sem Memória
7. Informação Mútua
8. Capacidade de um Canal Discreto sem Memória - o Caso do Canal (Ruidoso) BSC
9. Teorema da Codificação de Canal (2^o Teorema de Shannon)

Introdução

- Claude E. Shannon, 1948 $\Leftrightarrow$ Teoria Matemática das Comunicações
- Medida quantitativa de informação
- Limites da compressão e comunicação confiável
- A Teoria de Códigos Corretores de Erro, a partir da década de 50

- Aplicações práticas em comunicações
  - ligações interurbana e internacional
  - telefonia celular
  - TV digital
  - MODEMS
  - Meios ópticos e magnéticos de armazenagem de informação

Sistema de Comunicação Digital

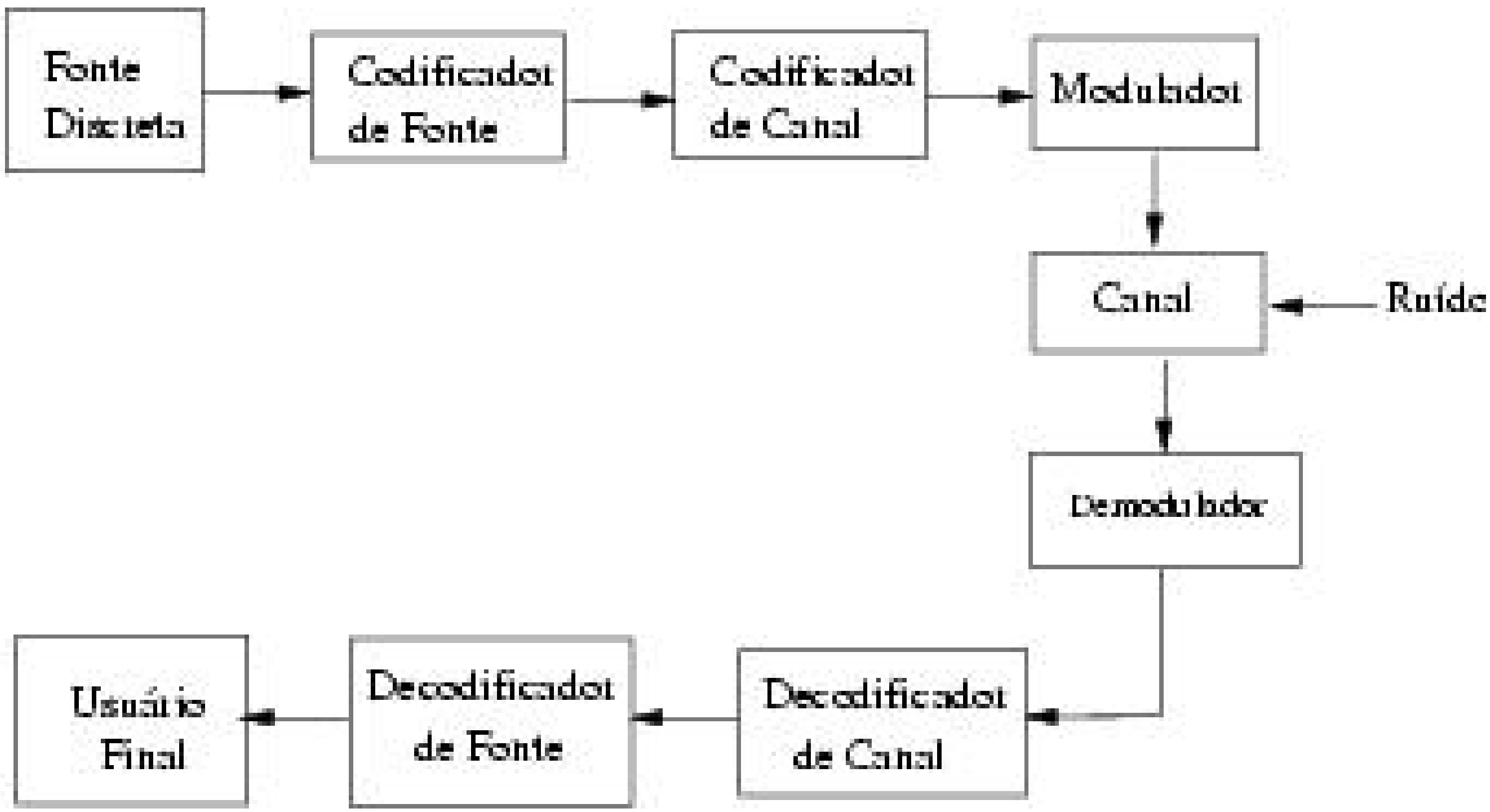


Figura 1: Sistema de comunicação digital.

Codificação de Fonte

Figura 2: Diagrama de blocos para a codificação de fonte.

OBJETIVO: O papel deste codificador é o de retirar a informação redundante (ou inútil, desnecessária) da fonte, explorando a estatística da fonte.

Exemplo: Código Morse (1837)

Cada letra do alfabeto é representada por uma seqüência formada dos símbolos "·" e "—"

Na língua inglesa,

$$\text{A letra } E \Leftrightarrow 10,3\%$$

$$\text{A letra } Q \Leftrightarrow 0,08\%$$

Morse escolheu as representações:

$$E \Leftrightarrow \text{“}\cdot\text{”}$$

$$Q \Leftrightarrow \text{“—,—,}\cdot\text{,—”}$$

Codificação de Canal

Figura 3: Diagrama de blocos para a codificação de canal.

OBJETIVO: O papel da codificação de canal é o de minimizar a probabilidade de erro dos símbolos transmitidos através da adição de redundância estruturada à informação a ser transmitida pelo canal ruidoso.

Exemplo: código detector de erro

Considere a fonte discreta com alfabeto $S = \{s_1, s_2, s_3, s_4\}$. Podemos ter a seguinte representação:

$s_1 \leftrightarrow 00$, $s_2 \leftrightarrow 01$, $s_3 \leftrightarrow 10$ e $s_4 \leftrightarrow 11$.

Adicionando-se um bit de redundância, podemos ter o seguinte código detector de 1 erro de bit:

$$\begin{array}{ccc} s_1 & \leftrightarrow & 000 \\ s_2 & \leftrightarrow & 011 \\ s_3 & \leftrightarrow & 101 \\ s_4 & \leftrightarrow & 110 \end{array}$$

Exemplo: código corretor de erro

Considere a fonte discreta com alfabeto $S = \{s_1, s_2\}$. Podemos ter o seguinte código corretor de 1 erro de bit:

$$\begin{array}{ccc} s_1 & \leftrightarrow & 000 \\ s_2 & \leftrightarrow & 111 \end{array}$$

Em síntese, a teoria da informação nos fornece dois limites fundamentais:

1. **Sobre a codificação de fonte:** O número médio de dígitos por símbolo de uma fonte S é sempre $\geq H(s)$, a **entropia** da fonte.

2. **Sobre a codificação de canal:** Para se ter uma comunicação confiável, ou seja, com uma probabilidade de erro tão pequena quanto se queira, a taxa de transmissão R deve ser $< C$, a **capacidade de canal**.

Fontes Discretas sem Memória

Uma **fonte discreta** é qualquer dispositivo que emita seqüências de símbolos pertencentes a um alfabeto fixo e finito, digamos

$$S = \{s_1, s_2, \ldots, s_n\}.$$

Para especificar plenamente a fonte discreta, necessitamos de uma distribuição de probabilidade para os símbolos emitidos pela fonte.

Se

$$P(s_i^k | s_j^{k-1}) = P(s_i^k)$$

então a fonte é dita ser uma **fonte sem memória**.

Uma fonte discreta sem memória nada mais é do que uma *variável aleatória discreta*.

Medidas de Informação

Vamos considerar uma fonte discreta sem memória como sendo uma variável aleatória discreta X, cujo espaço amostral é também denotado por

$$X = \{x_1, x_2, \ldots, x_L\}$$

com L eventos elementares. Denotaremos por

$$I(x)$$

a quantidade de informação contida no símbolo $x \in X$.

A Medida de Hartley

Em 1928, Hartley, na tentativa de estabelecer uma medida quantitativa para informação, fez as seguintes considerações:

1. Um símbolo contém informação apenas se existir mais de um valor possível para este símbolo, isto é, se $L \geq 2$.

   Se $L = 1$, então $I(x) = 0$.

2. Se $L \geq 2$, a quantidade de informação de n símbolos é n vezes a quantidade de informação de um símbolo.

A Medida de Hartley (Cont.)

Assim, Hartley propôs:

$$I(x) = \log_b L$$

Note que esta medida de informãção satisfaz as duas condicões acima:

1. $L = 1 \Rightarrow I(X) = \log_b 1 = 0$

2. $\underbrace{- - - \ldots -}_{n \text{ símbolos}}$ Para L^n possíveis valores, temos que

$$I(x) = \log_b L^n = n \; I(x)$$

A Escolha da Base do Logaritmo é Arbitrária

Para $b = 2$, a unidade de $I(x)$ é **bits**.

Para $b = 10$, a unidade é **Hartley**.

Para $b = e = 2,71\ldots$, temos a unidade **nats**.

Exemplo

Considere o caso de uma moeda ($L = 2$), ou seja, X={cara, coroa}.

Para $b = 2$, temos:

$$I(x) = \log_2 2 = 1 \text{ bit de informação}$$

Para $b = e = 2,71\ldots$, temos:

$$I(x) = \log_e 2 = \ln 2 = 0,693 \text{ nats de informação}$$

Experimento com urnas

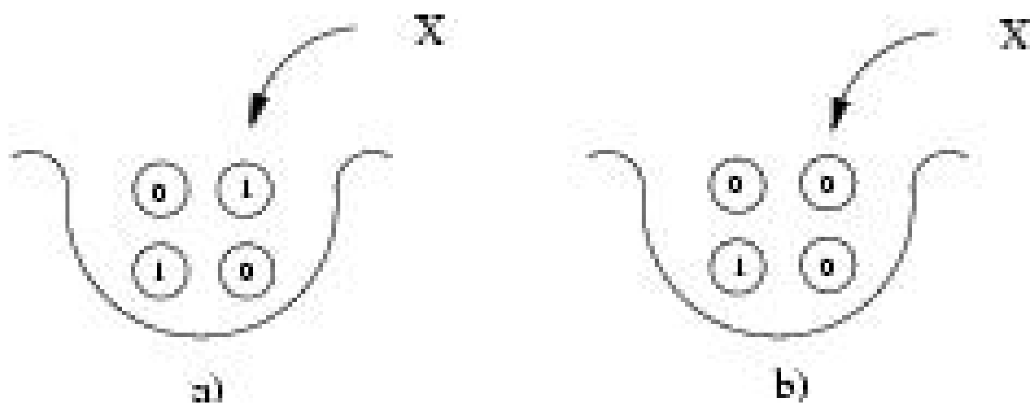


Figura 4: Experimento com duas urnas sobre a medida de informação de Hartley.

Se a bola $X = 0$ é retirada, qual a quantidade de informação revelada?

Como $L = 2$, segundo Hartley,

$$I(x) = \log_2 2 = 1 \;\; \text{bit}$$

para qualquer uma das duas urnas.

Porém, intuitivamente, diríamos que a quantidade de informação $I(X = 0)$ obtida quando uma bola 0 é retirada da urna (a) é maior do que quando uma bola 0 é retirada da urna (b).

Porquê?

A Medida de Informação de Shannon

A medida de Shannon leva em consideração a probabilidade de ocorrência do símbolo.

Shannon definiu a quantidade de informação de um símbolo como:

$$I(x_i) = \log_b \frac{1}{P(X = x_i)} = -\log P(X = x_i)$$

Exemplo

Voltando ao exemplo das urnas, teremos que:

1. Na urna (a), $P(X = 0) = 1/2 \Rightarrow I(0) = \log_2 \frac{1}{\frac{1}{2}} = 1$ bit

2. Na urna (b), $P(X = 0) = 3/4 \Rightarrow I(0) = \log_2 \frac{1}{\frac{3}{4}} = 0,415$ bit.

Portanto, a medida de Shannon satisfaz, além das duas condições de Hartley (verifique isso), a nossa intuição.

Exemplo

Neste exemplo $X = \{x_1, x_2, \ldots, x_L\}$ é uma variável aleatória com distribuição uniforme, ou seja, $P(x_i) = \frac{1}{L}$, para $i = 1, 2, \ldots, L$.

Assim, para todo $i = 1, 2, \ldots, L$, temos que $I(x_i) = \log_b \frac{1}{1/L} = \log_b L$.

Note que neste caso, e somente neste caso, as medidas de Hartley e de Shannon fornecem o mesmo resultado.

Exemplo

Consideremos agora a variável aleatória X tal que $P(X = x_i) = 1$ para algum i, e $P(X = x_j) = 0$ para todo $j \neq i$.

Assim,

$$I(x_i) = \log_b \frac{1}{1} = \log_b 1 = 0$$

e

$$I(x_j) = \log_b \frac{1}{0} = \log_b \infty = \infty$$

Fração de bit !!!???

Devemos perceber a diferença entre um *bit* e um *dígito binário*.

Note que o evento 101, com probabilidade $3/4$, é composto por 3 dígitos binários, mas contém $0,415$ bit de informação.

$I(X)$ é uma Variável Aleatória

Devemos observar que $I(X)$ é uma função da variável aleatória X, uma vez que para cada evento elementar $x_i \in X$ obtemos um valor $I(x_i)$ a partir da função $I(\cdot)$.

Conseqüentemente, $I(X)$ também é uma variável aleatória, com espaço amostral

$$\{I(x_1), I(x_2), \ldots, I(x_L)\},$$

onde $I(x_i)$ ocorre com probabilidade $P(X = x_i)$.

Entropia

A entropia (ou incerteza a priori) de uma variável aleatória X (ou de uma fonte X) é definida como:

$$\begin{aligned} H(X) &\triangleq E\{I(X)\} = \sum_{x_i \in X} P(x_i) I(x_i) \\ &= \sum_{x_i \in X} P(x_i) \log_b \frac{1}{P(x_i)} \end{aligned}$$

O que $H(X)$ representa?

$H(X)$ pode ser entendida como a quantidade média de informação produzida por um símbolo da fonte X.

Numa outra leitura, $H(X)$ é considerada como a **incerteza a priori** sobre o valor da variável aleatória X.

Incerteza, ..., Surpresa, ..., Informação

Antes de se observar X, tem-se uma incerteza a priori igual a $H(X)$ bits.

Ao se observar X, tem-se em média uma surpresa de $H(X)$ bits.

Depois de X ter sido revelada, ganha-se em média $H(X)$ bits de informação e a incerteza a respeito de X é reduzida a zero.

> **"Informação é a diferença entre as incertezas antes e depois da revelação de uma variável aleatória."**

Exemplo

Seja $X = \{x_1, x_2, x_3\}$ uma variável aleatória com $P(x_1) = 1/2$, $P(x_2) = P(x_3) = 1/4$.

Assim,

$$H(X) = \frac{1}{2}\log_2\frac{1}{1/2} + \frac{1}{4}\log_2\frac{1}{1/4} + \frac{1}{4}\log_2\frac{1}{1/4} = 1,5 \text{ bits}$$

Exemplo

Seja N uma variável aleatória de Bernoulli representando o ruído binário num canal binário simétrico, descrita por:

$$N = \begin{cases} 0, & \text{com probabilidade } 1-p \\ 1, & \text{com probabilidade } p \end{cases}$$

A entropia da variável aleatória binária N é obtida por:

$$H(N) = p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p} \triangleq \mathcal{H}(p)$$

A função $\mathcal{H}(p)$ é conhecida como a entropia de uma variável aleatória binária com parâmetro p

Teorema 1

Seja X uma variável aleatória com L possíveis valores, segundo uma distribuição de probabilidades $P(x)$.

Então,

$$0 \leq H(X) \leq \log L$$

A igualdade da esquerda ocorre se e somente se $P(x_i) = 1$ para algum i, com $1 \leq i \leq L$, e $P(x_j) = 0$ para todo $j \neq i$.

A igualdade da direita ocorre se e somente se $P(x_i) = 1/L$ para todo $1 \leq i \leq L$.

Lema 1: A Desigualdade da Teoria de Informação

Seja z um número real e positivo. Então,

$$\frac{\log_b z}{\log_b e} = lnz \leq z - 1$$

com igualdade se e somente se $z = 1$.

Assim,

$$\log_b z \leq (z-1)\log_b e.$$

Provar que $H(X) \leq \log L$**:**

$$
\begin{aligned}
H(X) - \log L &= \sum_x P(x) \log \frac{1}{P(x)} - \log L \left(\sum_x P(x) \right) \\
&= \sum_x P(x) \left[\log \left(\frac{1}{P(x)} \right) - \log L \right] \\
&= \sum_x P(x) \log \left(\frac{1}{L\, P(x)} \right) \\
&\leq \sum_x P(x) \left[\frac{1}{L\, P(x)} - 1 \right] \log e \\
&= \log e - \log e \\
&= 0
\end{aligned}
$$

Note que a desigualdade acima segue do Lema 1, fazendo-se a substituição

$$z = \frac{1}{L\ P(x)}.$$

Do Lema 1, temos a igualdade se e somente se

$$z = \frac{1}{L\ P(x)} = 1 \Rightarrow P(x) = \frac{1}{L}, \forall x \in X$$

ou seja, se e somente se X tem distribuição uniforme.

Isso conclui a prova do Teorema 1.

Exemplo

Considere um conjunto de n bolas que, aos nossos olhos, têm a mesma aparência e mesmo peso, porém uma delas possivelmente é diferenciada. Ou seja, as n bolas podem ser idênticas ou exatamente uma delas pode ser mais leve ou mais pesada. As três situações para a bola diferenciada são eqüiprováveis. Devemos descobrir a bola diferenciada (se existir) através do uso de uma balança de feira.

Dado o número médio K de pesagens, e supondo que a balança nos forneça 3 possíveis resultados (-, $+$, ou 0), devemos encontrar um limitante superior para o número de bolas, n, de modo que a bola diferenciada possa ser determinada com este número médio K de pesagens.

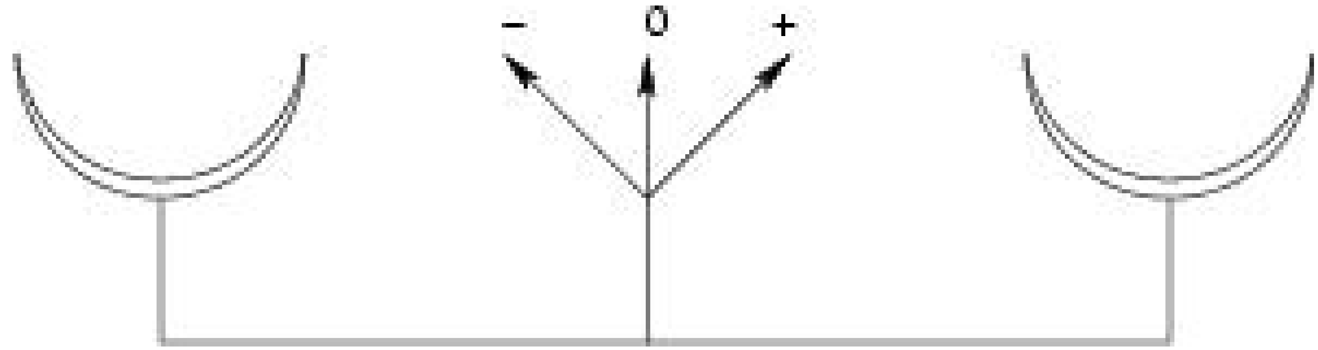


Figura 5: Balança para o problema das n bolas com uma bola possivelmente diferenciada.

A incerteza do problema é:

$$H(X) = \log_2(2n+1) \text{ bits.}$$

Supondo que a balança seja carregada de modo que os 3 resultados sejam eqüiprováveis, e supondo que os resultados de sucessivas pesagens sejam estatisticamente independentes, então a quantidade média de informação obtida da balança é:

$$K \log_2 3 \text{ bits.}$$

Daí, obtemos

$$K \log_2 3 \geq H(X) = \log_2(2n+1) \Rightarrow n \leq \frac{2^{K \log_2 3} - 1}{2}$$

Exemplo

Para $K = 3$, usamos

$$n \leq \frac{2^{K \log_2 3} - 1}{2}$$

e teremos que $n \leq 13$.

De fato, é possível resolver o problema com $n = 12$ bolas e 3 pesagens.

É impossível resolver o problema com 3 pesagens se o número de bolas for 14.

Entropia Conjunta

A *entropia conjunta* $H(X,Y)$ (ou $H(XY)$) de um par de variáveis aleatórias (X,Y) com distribuição de probabilidades conjunta $P(x,y)$ é definida como:

$$H(X,Y) \triangleq \sum_x \sum_y P(x,y) \log \left[\frac{1}{P(x,y)} \right]$$

Note que não há nada de novo nessa definição, uma vez que (X,Y) pode ser considerado como uma única variável aleatória (na verdade um vetor aleatório) $Z = (X,Y)$.

Naturalmente, podemos ter a entropia conjunta para n variáveis aleatórias:

$$H(X_1, X_2, \ldots, X_n)$$

Entropia Condicionada: $H(X|Y=y)$

A entropia condicionada de uma variável aleatória X, dado que $Y=y$, é definida como:

$$H(X|Y=y) = \sum_{x \in X} P(x|y) \log \frac{1}{P(x|y)}$$

Entropia Condicionada: $H(X|Y)$

A entropia condicionada de X, dado Y, é definida como:

$$\begin{aligned} H(X|Y) &= \sum_{y \in Y} P(y) H(X|Y=y) \\ &= \sum_{y \in Y} P(y) \left(\sum_{x \in X} P(x|y) \log \frac{1}{P(x|y)} \right) \\ &= \sum_{x \in X} \sum_{y \in Y} P(x|y) P(y) \log \frac{1}{P(x|y)} \\ &= \sum_{x \in X} \sum_{y \in Y} P(x,y) \log \frac{1}{P(x|y)} \end{aligned}$$

Exemplo

Considere o canal BSC. Digamos que os eventos $X = 0$ e $X = 1$ sejam eqüiprováveis. Devemos calcular $H(X|Y = y)$ para $y = 0$ e $y = 1$ e, em seguida, obter $H(X|Y)$.

Como X é uma variável aleatória binária uniformemente distribuída, teremos que $H(X) = 1$ bit.

Calculando $H(X|Y = 0)$, temos que:

$$H(X|Y = 0) = \sum_{x \in X} P(x|0) \log \frac{1}{P(x|0)}$$

Note que não dispomos da probabilidade $P(x|y)$.

Mas, pelo teorema de Bayes:

$$\begin{aligned} P(x|y) &= P(X=x|Y=y) \\ &= \frac{P(Y=y|X=x)P(X=x)}{P(Y=y|X=0)P(X=0)+P(Y=y|X=1)P(X=1)} \end{aligned}$$

onde todos as probabilidades envolvidas são conhecidas.

Analogamente, podemos calcular $H(X|Y=1)$, e em seguida:

$$H(X|Y) = \sum_{y \in Y} P(y)H(X|Y=y)$$

Note que $P(y)$ pode ser obtida pelo teorema da probabilidade total:

$$P(Y=y) = P(Y=y|X=0)P(X=0)+P(Y=y|X=1)P(X=1)$$

Teorema 2

Para quaisquer variáveis aleatórias discretas X e Y,

$$H(X|Y) \leq H(X)$$

com igualdade se e somente se X e Y são estatisticamente independentes.

Prova Devemos provar que $H(X|Y) - H(X) \leq 0$. Assim,

$$\begin{aligned}
H(X|Y) - H(X) &= \sum_x \sum_y P(x,y) \log \frac{1}{P(x|y)} - \sum_x P(x) \log \frac{1}{P(x)} \\
&= \sum_x \sum_y P(x,y) \left[\log \frac{1}{P(x|y)} - \log \frac{1}{P(x)} \right] \\
&= \sum_x \sum_y P(x,y) \log \left(\frac{P(x)}{P(x|y)} \right) \\
&= \sum_x \sum_y P(x,y) \log \left(\frac{P(x)P(y)}{P(x,y)} \right) \\
&\leq \sum_x \sum_y P(x,y) \left[\frac{P(x)P(y)}{P(x,y)} - 1 \right] \log e \\
&= (1-1) \log e \\
&= 0
\end{aligned}$$

onde a desigualdade acima segue do Lema 1, fazendo-se a substituição:

$$z = \frac{P(x)P(y)}{P(x,y)}.$$

Portanto, $H(X|Y) \leq H(X)$. Do Lema 1, a igualdade ocorre se e somente se:

$$z = \frac{P(x)P(y)}{P(x,y)} = 1$$

Ou seja, se e somente se $P(x,y) = P(x)P(y)$, que equivale a dizer que X e Y são estatisticamente independentes.

Comentário

Esse resultado é bastante intuitivo. Note que se X e Y são estatisticamente independentes, então o conhecimento de Y não deveria reduzir a incerteza a respeito de X. Logo, $H(X|Y) = H(X)$, e ficamos com a mesma incerteza a respeito de X observando Y ou não.

Exemplo

Sejam X e Y variáveis aleatórias com distribuição conjunta $P(x, y)$ dada por $P(1,1) = 0$, $P(1,2) = 1/8$, $P(2,1) = 3/4$ e $P(2,2) = 1/8$. Devemos calcular $H(X)$, $H(X|Y = 1)$, $H(X|Y = 2)$ e $H(X|Y)$, e comentar os resultados.

Para calcularmos $H(X)$, precisamos da distribuição de probabilidade marginal $P(x)$:

$$P(x) = \sum_y P(x, y)$$

Assim,

$$P(X = 1) = \sum_y P(1, y) = P(1, 1) + P(1, 2) = 0 + \frac{1}{2} = \frac{1}{8}$$

e

$$P(X = 2) = \sum_y P(2, y) = \frac{7}{8}.$$

A entropia $H(X)$ pode agora ser obtida:

$$H(X) = \sum_{x} P(x) \log_2 \frac{1}{P(x)} = \frac{1}{8} \log_2 8 + \frac{7}{8} \log_2 \left(\frac{8}{7} \right) = 0,544 \text{ bits}$$

A seguir, as entropias condicionadas podem ser obtidas. Precisamos da probabilidade condicionada $P(x|y)$, que pode ser facilmente obtida usando-se a regra de Bayes. Assim,

$$H(X|Y=1) = \sum_{x} P(x|Y=1) \log_2 \frac{1}{P(x|Y=1)} = 0$$

e

$$H(X|Y=2) = \sum_{x} P(x|Y=2) \log_2 \frac{1}{P(x|Y=2)} = 1 \text{ bit}$$

Note que, curiosamente, $H(X|Y=2) > H(X)$. Ou seja, uma informação privilegiada aumentou a incerteza sobre X.

Evidências reduzem incerteza . . .

O conhecimento particular de um evento $Y = y$ pode eventualmente aumentar a incerteza sobre X, mas em média o conhecimento de Y reduz a incerteza sobre X, como demonstramos no Teorema 2, ou seja, *na média, informação privilegiada reduz incerteza*.

Note que ao tentar desvendar um crime, um detetive pode ter a sua incerteza aumentada a partir de uma evidência particular.

Mas, na média,

> **"Evidências reduzem incerteza"**

Exemplo Anterior

Podemos verificar mais uma vez essa afirmação no cálculo de $H(X|Y)$:

$$
\begin{aligned}
H(X|Y) &= \sum_y P(y)H(X|Y=y) \\
&= \frac{3}{4} \times 0 + \frac{1}{4} \times 1 = 0,25 < H(X)
\end{aligned}
$$

Teorema 3: [Regra da cadeia para $H(X,Y)$]

Sejam X e Y variáveis aleatórias discretas quaisquer. Então:

$$H(X,Y) = H(X) + H(Y|X)$$

ou

$$H(X,Y) = H(Y) + H(X|Y)$$

Prova

Escrevemos

$$
\begin{aligned}
H(X,Y) &= -\sum_x \sum_y P(x,y) \log P(x,y) \\
&= -\sum_x \sum_y P(x,y) \log \left(P(y|x)P(x)\right) \\
&= -\sum_x \sum_y P(x,y) \log P(x) - \sum_x \sum_y P(x,y) \log(P(y|x)) \\
&= -\sum_x P(x) \log P(x) - \sum_x \sum_y P(x,y) \log(P(y|x)) \\
&= H(X) + H(Y|X)
\end{aligned}
$$

Analogamente, podemos mostrar que $H(X,Y) = H(Y) + H(X|Y)$.

Exemplo

Sejam X e Y variáveis aleatórias discretas descritas pela distribuição de probabilidades conjunta mostrada na tabela abaixo.

P(x,y)	X=1	X=2	X=3	X=4
Y=1	1/8	1/16	1/32	1/32
Y=2	1/16	1/8	1/32	1/32
Y=3	1/16	1/16	1/16	1/16
Y=4	1/4	0	0	0

Verificamos que $P(x) = (1/2, 1/4, 1/8, 1/8)$, e que $P(y) = (1/4, 1/4, 1/4, 1/4)$. Assim, $H(X) = 7/4$ bits e $H(Y) = 2$ bits.

A entropia conjunta é obtida por:

$$
\begin{aligned}
H(X|Y) &= \sum_{i=1}^{4} P(Y=i)H(X|Y=i) \\
&= \frac{1}{4}\sum_{i=1}^{4} H(X|Y=i) \\
&= \frac{1}{4}\Bigg\{ \underbrace{H_{X|Y}(1/2,1/4,1/8,1/8)}_{\text{para } Y=1} + \underbrace{H_{X|Y}(1/4,1/2,1/8,1/8)}_{\text{para } Y=2} + \\
&\quad + \underbrace{H_{X|Y}(1/4,1/4,1/4,1/4)}_{\text{para } Y=3} + \underbrace{H_{X|Y}(1,0,0,0)}_{\text{para } Y=4} \Bigg\}
\end{aligned}
$$

Calculando as entropias condicionadas, obtemos:

$$
\begin{aligned}
H(X|Y) &= \frac{1}{4}\left\{\frac{7}{4} + \frac{7}{4} + 2 + 0\right\} \\
&= \frac{11}{8} \text{ bits}
\end{aligned}
$$

Note que fizemos uso da seguinte relação:

$$
P(X=1|Y=1) = \frac{P(X=1, Y=1)}{P(Y=1)} = \frac{1/8}{1/4} = \frac{1}{2}
$$

Também, $H(Y|X) = 13/8$ e $H(X,Y) = 27/8$.

Assim, constatamos que:

1. $H(X,Y) = 27/8 = H(X) + H(Y|X) = 7/4 + 13/8 = 27/8$

2. $H(X,Y) = 27/8 = H(Y) + H(X|Y) = 2 + 11/8 = 27/8$

3. $H(Y|X) < H(Y) \Leftrightarrow 13/8 < 2 = 16/8$

4. $H(X|Y) \neq H(Y|X) \Leftrightarrow 11/8 \neq 13/8$

Veremos mais adiante que $H(X) - H(X|Y) = H(Y) - H(Y|X)$.

Para o exemplo anterior, temos

$$H(X) - H(X|Y) = 7/4 - 11/8 = 3/8 = H(Y) - H(Y|X) = 2 - 13/8 = 3/8$$

Teorema 4: [Regra da cadeia para n variáveis aleatórias]

Sejam $X_1, X_2, \ldots, X_n$ n variáveis aleatórias discretas quaisquer. Então:

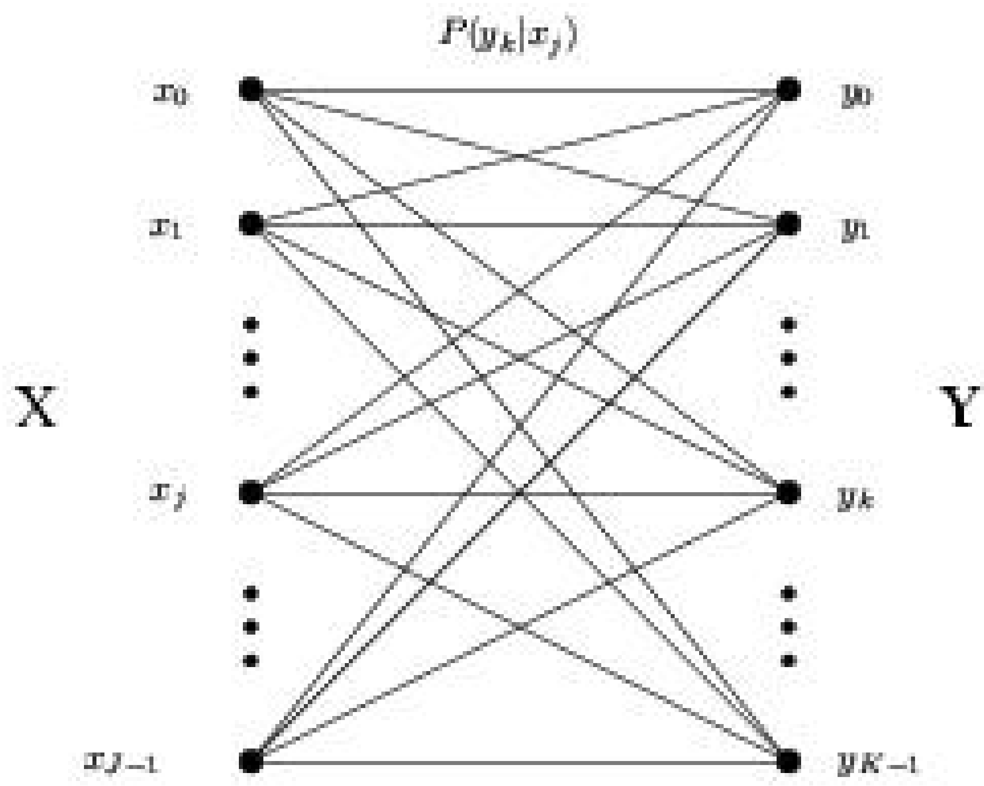
$$H(X_1, X_2, \ldots, X_n) = H(X_1) + H(X_2|X_1) + H(X_3|X_1, X_2) + \ldots + H(X_n|X_1, X_2, \ldots$$

Prova Segue facilmente por indução matemática.

Canais Discretos sem Memória (DMC)

Um DMC (do inglês *Discrete Memoryless Channel*) é um canal ruidoso através do qual qualquer um dos J símbolos de $X = \{x_0, x_1, \ldots, x_{J-1}\}$ pode ser transmitido. O símbolo recebido pode ser qualquer um dos K símbolos de $Y = \{y_0, y_1, \ldots, y_{K-1}\}$, segundo a probabilidade condicionada: $P(Y = y_k | X = x_j) = P(y_k | x_j)$, para $i = 0, 1, 2, \ldots, J - 1$ e $j = 0, 1, 2, \ldots, K - 1$.

Representação Gráfica de um DMC



Especificação do DMC

Um DMC é especificado por:

$$
\begin{aligned}
X &= \text{alfabeto de entrada} \\
Y &= \text{alfabeto de saída} \\
p(y_k|x_j) &= \text{probabilidade condicionada de } y_k \text{ dado } x_j.
\end{aligned}
$$

Esse canais são ditos sem memória pois o resultado da transmissão de um símbolo num dado instante não depende do resultado das últimas transmissões, ou seja, a variável aleatória representando o ruído no instante discreto k é estatisticamente independente da variável aleatória representando o ruído no instante discreto $j \neq k$. Daí se dizer que o canal não tem memória.

Exemplo

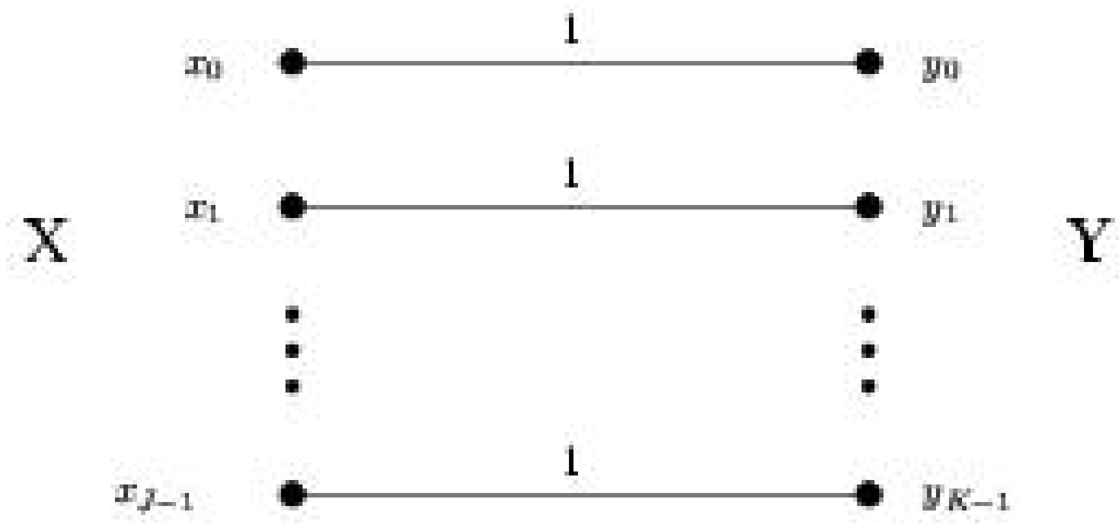


Figura 6: Canal sem ruído com J entradas e K saídas.

Exemplo

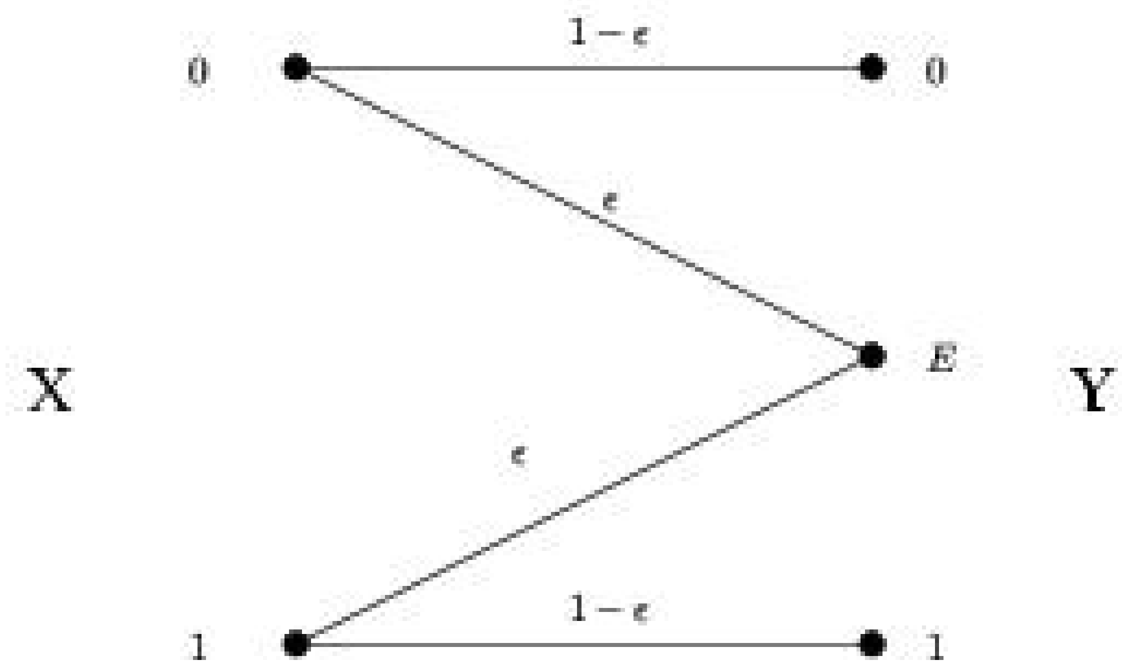


Figura 7: Canal binário com apagamento (BEC) com parâmetro ϵ.

Especificação de um DMC Genérico

Um DMC é plenamente especificado pelos alfabetos X e Y e pela matriz de canal:

$$P = \begin{bmatrix} P(y_0|x_0) & P(y_1|x_0)\ldots & P(y_{K-1}|x_0) \\ P(y_0|x_1) & P(y_1|x_1)\ldots & P(y_{K-1}|x_1) \\ \vdots & \vdots & \vdots \\ P(y_0|x_{J-1}) & P(y_1|x_{J-1})\ldots & P(y_{K-1}|x_{J-1}) \end{bmatrix}$$

Note que:

$$\sum_{k=0}^{K-1} P(y_k|x_j) = 1, \quad \forall j \in \{1, 2, \ldots, J-1\}$$

Também, dado $P(x_j)$ $\forall j$, temos que

$$P(y_k) = \sum_{j=0}^{J-1} P(y_k|x_j)P(x_j), \quad k = 0, 1, \ldots, K-1$$

Probabilidade de Erro em um DMC

Podemos calcular a probabilidade de erro dado que x_j tenha sido transmitido (fazemos $J = K$):

$$P_{e|x_j} = Prob\{erro|x_j\} = \sum_{\substack{k=0 \\ k \neq j}}^{K-1} P(y_k|x_j) = 1 - P(y_j|x_j)$$

onde $P(y_j|x_j)$ é a probabilidade de acerto.

A probabilidade de erro média é dada por:

$$\begin{aligned} P_e &= \sum_{j=0}^{J-1} P_{e|x_j} P(x_j) \\ &= \sum_{j=0}^{J-1} \sum_{\substack{k=0 \\ k \neq j}}^{K-1} P(y_k|x_j) P(x_j) \end{aligned}$$

A probabilidade de acerto é dada por:

$$P_c = 1 - P_e$$

Exemplo

Para o canal BSC ($J = K = 2$), temos que

$$
\begin{aligned}
P_e &= \sum_{k=0}^{1} \sum_{\substack{j=0 \\ j \neq k}}^{1} P(y_k|x_j)P(x_j) \\
&= P(y_0|x_1)P(x_1) + P(y_1|x_0)P(x_0) \\
&= p\,P(x_0) + p\,P(x_1) \\
&= p\,[P(x_0) + P(x_1)] \\
&= p \quad \text{(canal BSC)}
\end{aligned}
$$

Informação Mútua

Consideremos o canal discreto sem memória mostrado na figura abaixo.

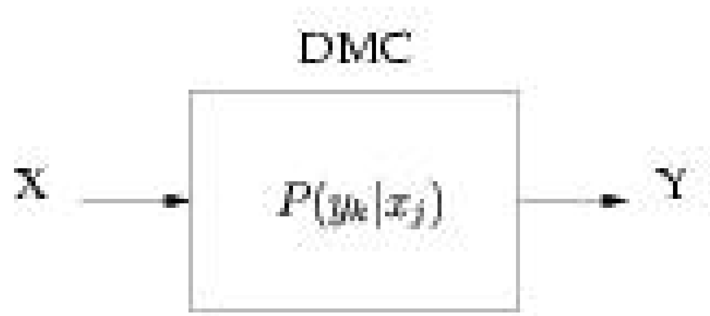


Figura 8: Canal discreto sem memória.

Informação Mútua (cont.)

Estamos interessados nas quantidades:

$$H(X) = \text{ incerteza sobre } X \text{ antes de se observar } Y$$

$$H(X|Y) = \text{ incerteza sobre } X \text{ depois de se observar } Y$$

e na diferença:

$$H(X) - H(X|Y)$$

que é a parte da incerteza sobre X que é “resolvida” ao se observar Y.

Definição de Informação Mútua

A informação mútua entre duas variáveis aleatórias X e Y, denotada por $I(X,Y)$, é definida como:

$$I(X,Y) \triangleq H(X) - H(X|Y)$$

Informação Mútua e Comunicações

É muito importante observar que $I(X,Y)$ representa a quantidade de informação sobre X que obtemos ao observarmos Y. Essa observação passará a ter um significado extramente importante no contexto de comunicações, quando X representar os dados a serem transmitidos por um canal ruidoso e Y representar o sinal recebido na saída deste canal. Neste contexto, $I(X,Y)$ representa a quantidade de informação transmitida pelo canal.

Teorema 5: [Simetria da Informação Mútua]

$$I(X, Y) = I(Y, X)$$

Teorema 6: [Não-negatividade]

$$I(X, Y) \geq 0$$

com igualdade se e somente se X e Y são estatisticamente independentes.

Prova:

Da definição de informação mútua, temos que

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

Do Teorema 2, temos que

$$H(X|Y) \leq H(X)$$

Logo

$$I(X,Y) = H(X) - H(X|Y) \geq H(X) - H(X) = 0$$

Teorema 7

$$H(X,Y) = H(X) + H(Y) - I(X,Y)$$

Prova:

Pela regra da cadeia,

$$H(X,Y) = H(X) + H(Y|X)$$

Do Teorema 5, temos que

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

ou seja,

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

Assim,

$$\begin{aligned} H(X,Y) &= H(X) + H(Y|X) \\ &= H(X) + H(Y) - I(X,Y) \end{aligned}$$

e o teorema está provado.

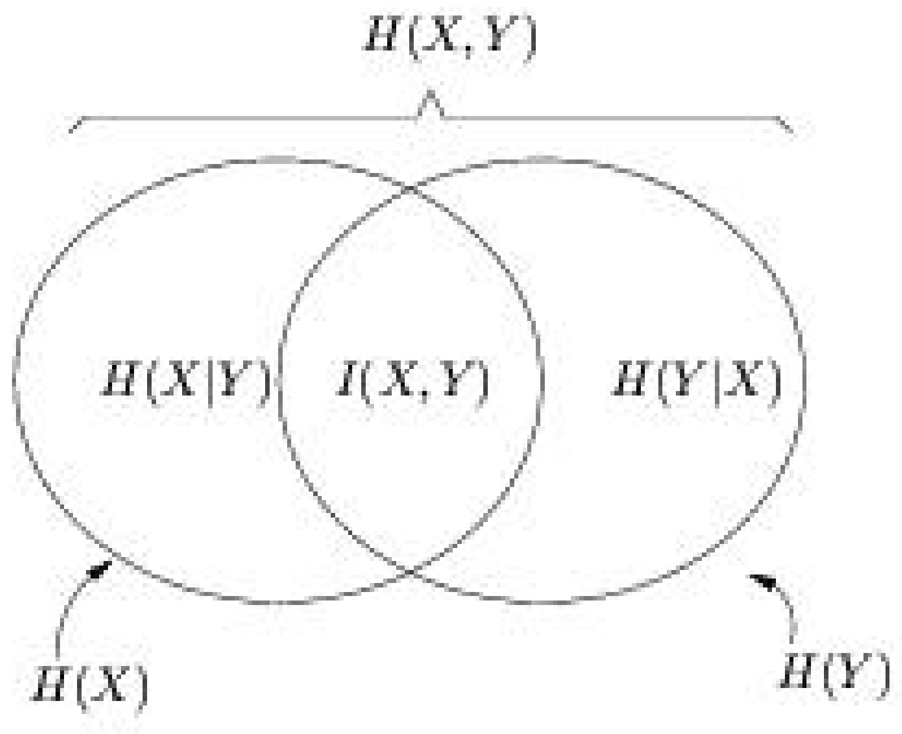


Figura 9: Relação entre as entropias e a informação mútua.

Exemplo

Consideremos mais uma vez o canal BSC, com $P(X = 0) = P(X = 1) = 1/2$, e parâmetro p. Calcular a informação mútua $I(X, Y)$.

Solução

Podemos usar qualquer uma das duas equações abaixo:

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

ou

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

A segunda equação é mais apropriada. Assim, obtemos $H(Y|X)$:

$$H(Y|X) = H(Y|X=0)P(X=0) + H(Y|X=1)P(X=1)$$

Dado que $X = 0$, temos que

$$P(Y=0|X=0) = 1 - p$$

e

$$P(Y = 1|X = 0) = p$$

Logo,

$$H(Y|X = 0) = (1 - p)\log\frac{1}{1 - p} + p\log\frac{1}{p} \triangleq \mathcal{H}(p)$$

Similarmente, podemos obter

$$H(Y|X = 1) = \mathcal{H}(p)$$

Deste modo, temos que

$$H(Y|X) = \mathcal{H}(p)$$

Pelo teorema da probabilidade total, temos que

$$\begin{aligned} P(Y=0) &= P(Y=0|X=0)P(X=0) + P(Y=0|X=1)P(X=1) \\ &= (1-p)\,\frac{1}{2} + p\,\frac{1}{2} \\ &= \frac{1}{2} \end{aligned}$$

Assim,

$$P(Y=0) = P(Y=1) = \frac{1}{2}$$

e conseqüentemente:

$$H(Y) = 1 \ \text{ bit}$$

Finalmente, a informação mútua é dada por:

$$I(X,Y) = H(Y) - H(Y|X) = 1 - \mathcal{H}(p)$$

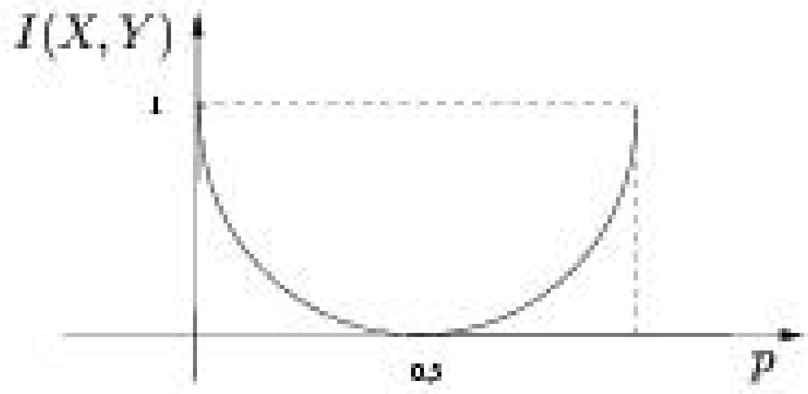


Figura 10: Informação mútua para o canal BSC com $P(X=0)=P(X=1)=1/2$, em função do parâmetro p.

$H(Y|X)$ **não depende de** $P(x)$**, mas** $I(X,Y)$ **depende . . .**

Devemos notar que, devido à simetria do canal BSC, temos que $H(Y|X) = \mathcal{H}(p)$, independentemente da distribuição de probabilidade da entrada do canal, ou seja, $P(x)$.

Mas $I(X,Y) = H(Y) - H(Y|X)$ depende de $P(x)$.

Vejamos a seguir o caso do canal BSC com distribuição $P(x)$ não uniforme.

Exemplo

Considere o canal BSC com parâmetro p, e com $P(X=0)=1-\gamma$ e $P(X=1)=\gamma$. Neste caso,

$$I(X,Y)=H(Y)-H(Y|X)=H(Y)-\mathcal{H}(p)$$

Para obtermos $H(Y)$, devemos calcular:

$$\begin{aligned} P(Y=0) &= P(Y=0|X=0)P(X=0)+P(Y=0|X=1)P(X=1) \\ &= (1-p)\,(1-\gamma)+p\,\gamma \\ &= 1-p-\gamma+2p\gamma \end{aligned}$$

Temos também que

$$P(Y=1)=1-P(Y=0)=p+\gamma-2p\gamma$$

Finalmente, a entropia $H(Y)$ é dada por:

$$
\begin{aligned}
H(Y) &= (1-p-\gamma+2p\gamma)\log\frac{1}{(1-p-\gamma+2p\gamma)} + (p+\gamma-2p\gamma)\log\frac{1}{(p+\gamma-} \\
&= \mathcal{H}(p+\gamma-2p\gamma)
\end{aligned}
$$

A informação mútua para o canal BSC com parâmetro p, e com $P(X=0)=1-\gamma$ e $P(X=1)=\gamma$, é portanto:

$$I(X,Y)=\mathcal{H}(p+\gamma-2p\gamma)-\mathcal{H}(p)$$

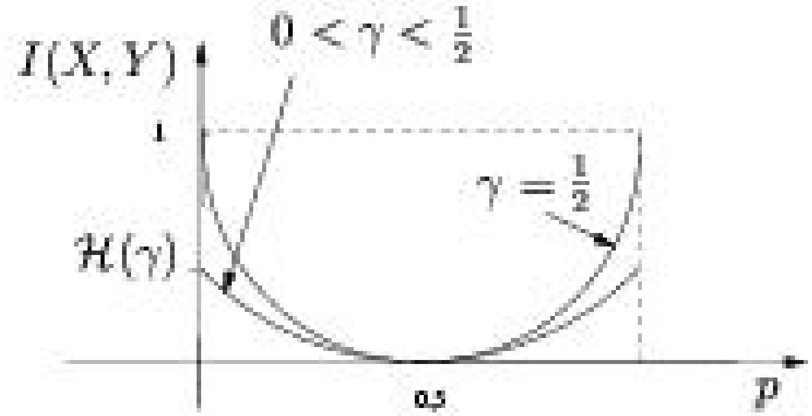


Figura 11: Informação mútua em função do parâmetro p para o canal BSC, com $P(X = 0) = 1 - \gamma$ e $P(X = 1) = \gamma$. Note que, para qualquer valor de p, $I(X, Y)$ é máxima para $\gamma = 1/2$.

Observações Importantes

É importante observarmos que $\gamma = 1/2$ maximiza $I(X, Y)$ para qualquer valor fixo de p, ou seja, para qualquer canal BSC fixo.

Isto significa que a quantidade máxima de informação que pode ser transmitida pelo canal BSC com parâmetro p é obtida quando a distribuição $P(x)$ for uniforme ($\gamma = 1/2$), ou seja, $P(X = 0) = P(X = 1) = 1/2$.

Isso nos leva ao conceito de **capacidade de canal**, apresentado a seguir.

Capacidade de Canal de um Canal Discreto Sem Memória - o Caso do Canal (Ruidoso) BSC

Lembremos de que um canal discreto sem memória (ou seja, um DMC) especifica a probabilidade condicionada $P(y|x)$.

Porém, a distribuição de probabilidades da entrada do canal, ou seja $P(x)$, pode ser escolhida a fim de maximizar a quantidade de informação transmitida pelo canal, como no caso da seção anterior.

Definição

A capacidade de canal de um canal DMC, denotada por C, é definida por:

$$C \triangleq \max_{P(x)} I(X,Y) \text{ bits por uso do canal}$$

Equivalentemente,

$$C \triangleq \max_{P(x)} [H(Y) - H(Y|X)]$$

Exemplo

Para o canal BSC com parâmetro p, a partir da Figura 11 concluimos que a capacidade de canal é dada por:

$$C_{\mathsf{BSC}} = 1 - \mathcal{H}(p)$$

onde $P(x)$ uniforme (ou seja, $\gamma = 1/2$) maximiza $I(X, Y)$.

Exemplo

Consideremos o canal binário com apagamento (BEC) ilustrado na Figura 7. Pode-se mostrar que a capacidade de canal I_{BEC} do canal BEC é como mostrada na figura abaixo.

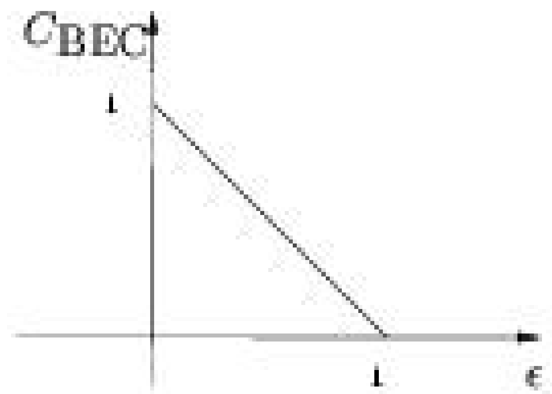
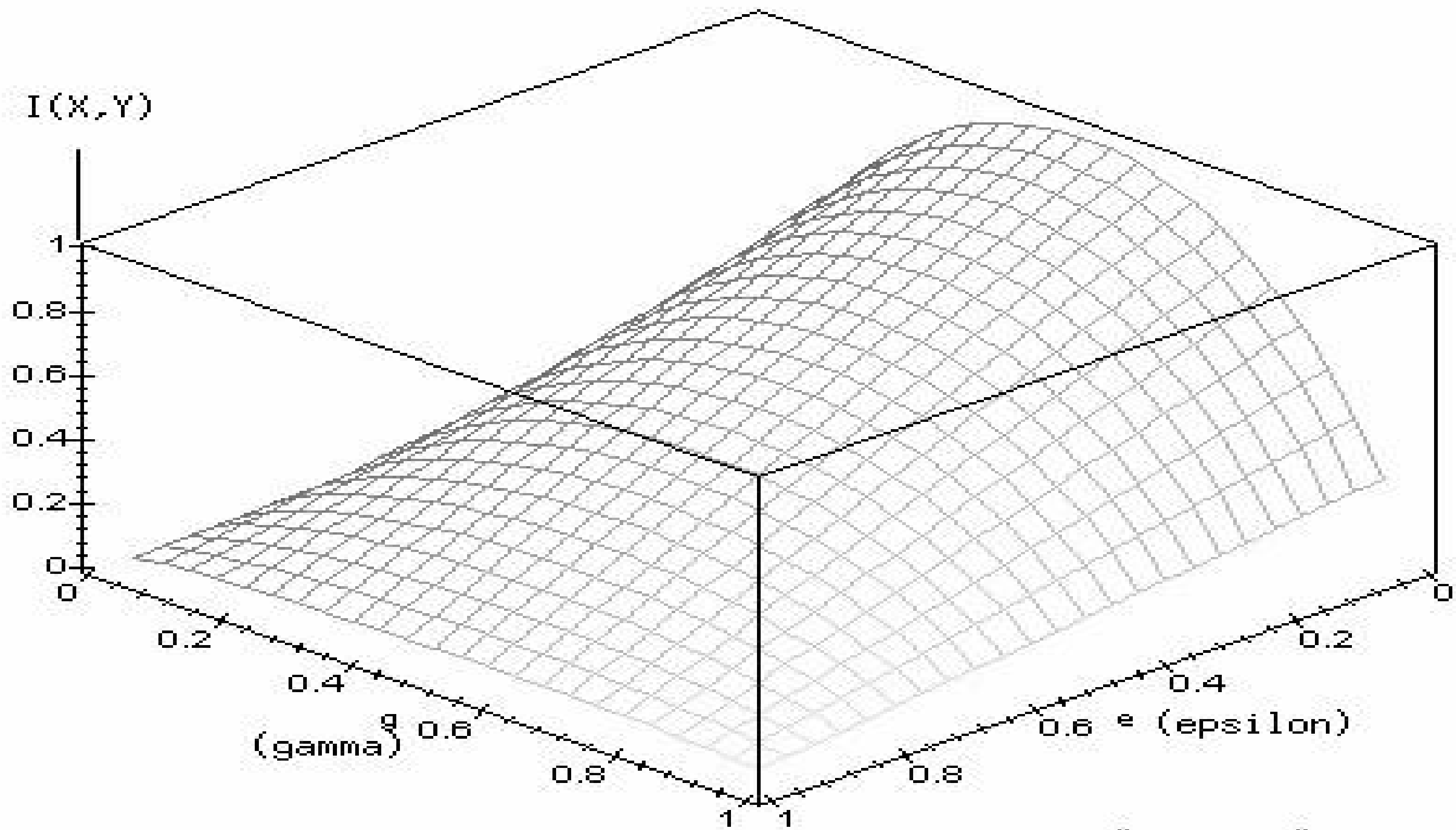


Figura 12: Capacidade de canal do canal binário com apagamento (BEC) com parâmetro ϵ.



Informacao Mutua para o canal BEC com parametro "epsilon" e
com distribuicao de entrada P(X=0) = 1 - gamma, e P(X=1) = gamma.

Codificação de Canal

Figura 14: Diagrama de blocos para a codificação de canal.

OBJETIVO: O papel da codificação de canal é o de minimizar a probabilidade de erro dos símbolos transmitidos através da adição de redundância estruturada à informação a ser transmitida pelo canal ruidoso.

Teorema da Codificação de Canal (2^o Teorema de Shannon)

Vamos admitir que a fonte binária produza bits u (com probabilidades $P(u = 0) = P(u = 1) = 1/2$) e que blocos $(u_1, u_2, \ldots, u_k)$ de k bits sejam formados e apresentados ao codificador de canal.

Note que, como u é uniformemente distribuída, temos que $H(u) = 1$ bit, e assim 1 bit equivale a um dígito binário.

O codificador de canal, por sua vez, produz uma palavra binária $(v_1, v_2, \ldots, v_n)$ com n bits para cada bloco de informação.

Essas são chamadas de *palavras código*. Assim, serão 2^k palavras código.

O conjunto de todas as palavras código é chamado de *código de bloco*. Assim, um código de bloco binário é qualquer subconjunto com 2^k n-uplas binárias do conjunto formado por todas as 2^n n-uplas binárias possíveis.

O canal discreto sem memória a ser considerado é o canal BSC. Note que, como k bits de informação são transmitidos através de n "usos do canal", a taxa de transmissão, R, é dada por:

$$R = \frac{k}{n} \text{ bits/uso}$$

É importante ressaltar que, na apresentação do próximo resultado, nos restringiremos à transmissão binária via um canal BSC, embora o conceito de capacidade de canal a ser apresentado seja válido para qualquer canal.

Teorema [Teorema da Codificação de Canal ou 2^o Teorema de Shannon]

Se os bits de informação de uma fonte binária forem transmitidos a uma taxa R (em bits por uso do canal) através de um canal (DMC) com capacidade C (em bits por uso do canal), então o seguinte é dito a respeito da probabilidade de erro.

Teorema da Codificação de Canal: Transmissão com Erros

1. Seja P_b a probabilidade de erro de bit. Então,

$$P_b \geq \mathcal{H}^{-1}(1 - C/R), \quad \text{se } R > C$$

onde $\mathcal{H}^{-1}(\alpha)$ é o valor de p, com $0 \leq p \leq 1/2$, tal que $\mathcal{H}(p) = p \log \frac{1}{p} + (1-p) \log(\frac{1}{1-p})$ seja igual a α.

Ou seja, se $R > C$, então não existe nenhum esquema de codificação de canal que proporcione uma probabilidade de erro menor do que um certo valor mínimo.

Teorema da Codificação de Canal: Transmissão Confiável

2. Seja P_E a probabilidade de erro de palavra código. Então,

$$P_E < 2^{-n(C-R)} + O(1/n), \quad \text{se } R < C$$

onde $O(1/n)$ representa uma expressão que tende a zero quando n tende a infinito.

Ou seja, se $R < C$, então P_E pode ser escolhida tão pequena quanto se queira. Basta usarmos um código de bloco com comprimento de palavra código n muito grande.

Exemplo: Transmissão com Erros

Suponha que um DMC tenha capacidade $C = 1/4$ bit/uso do canal, e que a taxa de transmissão seja $R = 1/2 > C$ bit/uso do canal. Então:

$$\begin{aligned} P_b &\geq \mathcal{H}^{-1}\left(1 - \frac{1/4}{1/2}\right) \\ &= \mathcal{H}^{-}1(1/2) \\ &= 0,11 \end{aligned}$$

Assim, não existe uma maneira de se transmitir a uma taxa $R = 1/2$ bit por uso do canal por este canal, e obtendo-se uma probabilidade de erro de bit inferior a 0,11.

Exemplo: Transmissão Confiável

Suponha agora que um DMC tenha capacidade $C = 1/2$ bit/uso do canal, e que a taxa de transmissão seja $R = 1/4 < C$ bit/uso do canal. Considere também a exigência de que a probababilidade de erro de palavra código satisfaça a desigualdade:

$$P_E < 10^{-10}$$

ou seja, em média haja uma palavra código errada a cada 10 trilhões de palavras código transmitidas. Então, segundo Shannon (e ignorando o termo $O(1/n)$), existe pelo menos um código tal que

$$P_E < 2^{-\frac{n}{4}} < 10^{-10} \Rightarrow n \geq 133$$

Ou seja, existe um código de bloco com comprimento de palavra código $n = 133$ que satisfaz as exigências do problema.

PROBLEMAS

1. Considere o canal binário ($J = K = 2$) não simétrico definido por $P(Y = 0|X = 0) = 2/3$, $P(Y = 1|X = 0) = 1/3$, $P(Y = 0|X = 1) = 1/10$ e $P(Y = 1|X = 1) = 9/10$. Faça a representação gráfica deste canal. Supondo que $P(X = 0) = 3/4$ e $P(X = 1) = 1/4$, encontre:

   (a) $P(y)$
   (b) $P(x|y)$
   (c) $H(X)$
   (d) $H(X|Y = 0)$ e $H(X|Y = 1)$
   (e) $H(X|Y)$
   (f) $I(X, Y)$

   Comente os resultados.

2. Calcule a capacidade de canal do canal da Figura 15. (**DICA:** Use $I(X,Y) = H(X) - H(X|Y)$).

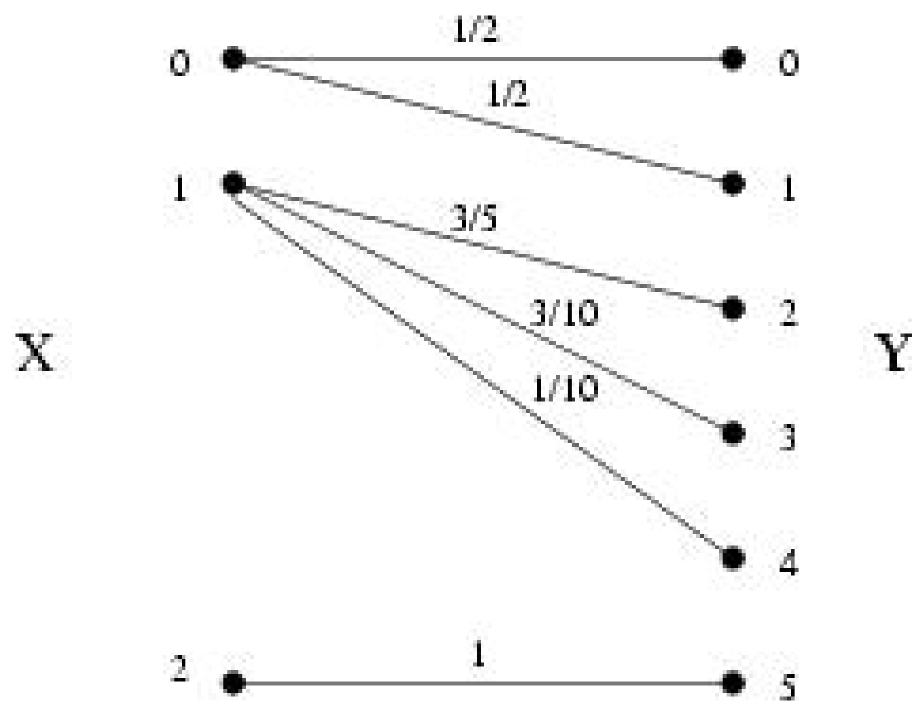


Figura 15: Capacidade discreto sem memória.

3. Calcule a capacidade de canal do canal da Figura 16. (**DICA:** Use $I(X,Y) = H(Y) - H(Y|X)$).

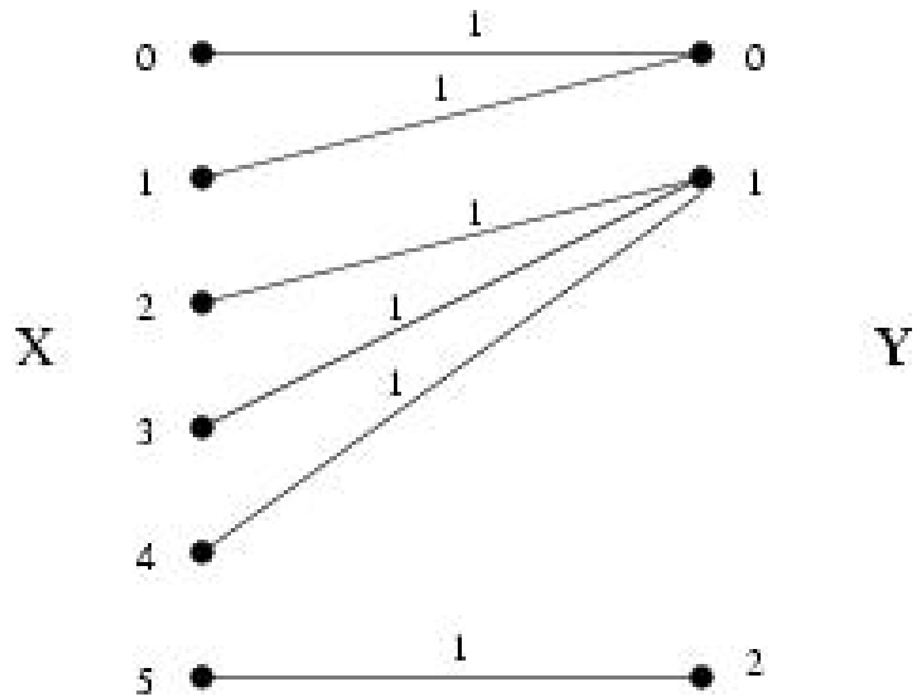


Figura 16: Capacidade discreto sem memória.