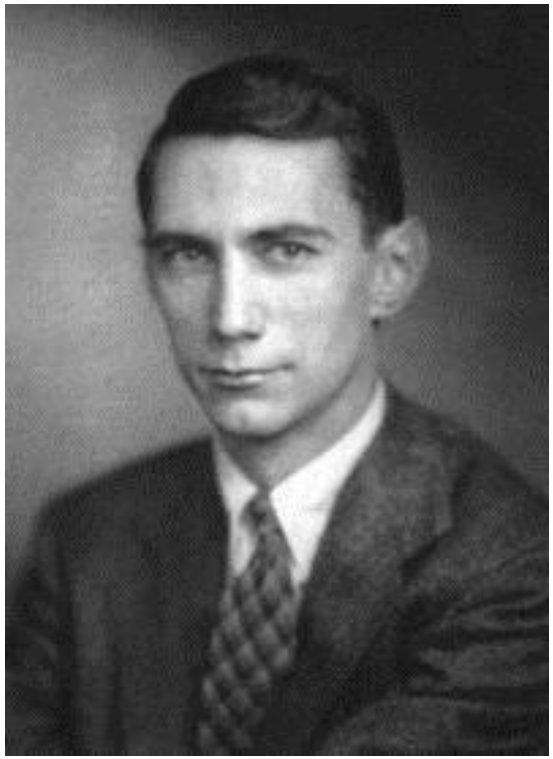
TEORIA DE INFORMAÇÃO – UM NANOCURSO

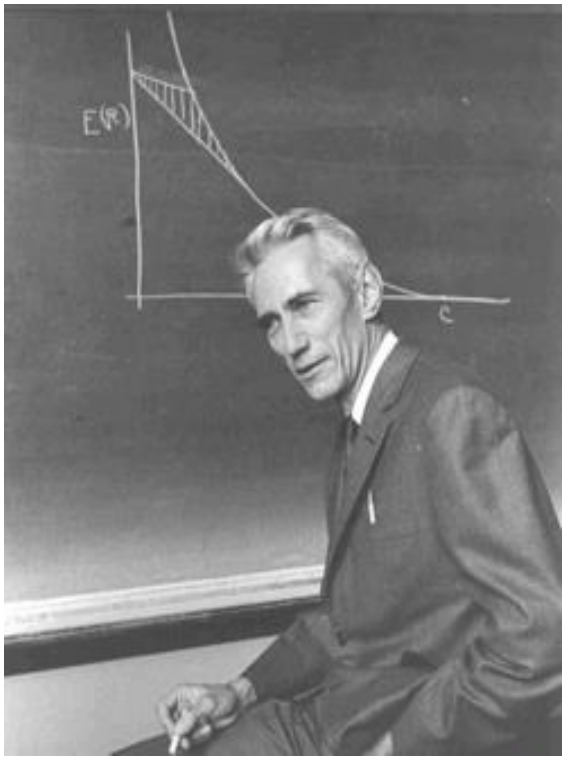
Max H. M. Costa
Unicamp

Set. 2016

Centenário de Shannon - SBrT - Santarém

Dedicação: a memória de Claude Shannon





Claude Shannon – 1916-2001 – matemático,
engenheiro eletrônico, inventor da Teoria de Informação

Sumário

Introdução

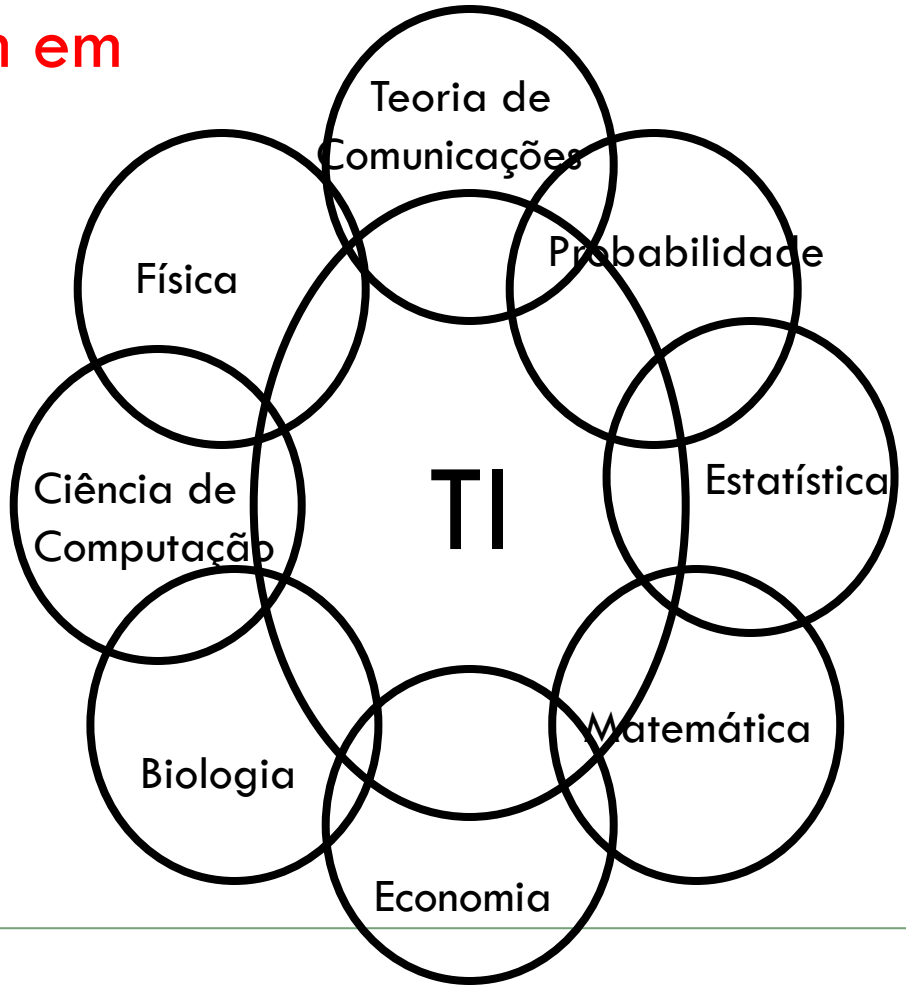
- Teoria de Informação - Interdisciplinaridade
- Entropia, Divergência de K-L, Informação Mútua
- Compressão de dados
- Transmissão por Canais Ruidosos (Codificação de Canal)
- Entropia Diferencial, Canais Gaussianos
- Aplicações de Múltiplos Usuários
- Fechamento

Algumas referências:

- [1] T. Cover and J. Thomas, Elements of Information Theory, Wiley, 2nd ed., 2006 (1991).
- [2] R. Ash, Information Theory, Dover, 1990.
- [3] R. Gallager, Information Theory and Reliable Communication, Wiley, 1968.
- [4] A. El Gamal and Y-H. Kim, Network Information Theory, Cambridge, 2011.

Teoria de Informação e Áreas Afins

- A paisagem em
- torno de TI:



Entropia

- Definição: H(X) = A Entropia de X
- Seja X uma variável aleatória discreta com valores
- em $\{x_1, x_2, ..., x_M\}$ com probabilidades
- $\mathbf{p} = \{p_1, p_2, ..., p_M\}$.

- $H(X) = H(\mathbf{p}) = \sum_{k=1}^{M} p(x_k)\, log_2 \frac{1}{p(x_k)}$ (bits) =

- $= E\left(log_2 \frac{1}{p(X)} \right)$ bits

H(X) é uma medida da incerteza de X.

Mudança de base

- $H_b(X) = E\left(log_b \frac{1}{p(X)} \right) = log_b a \quad H_a(X)$

- Unidades de Entropia:
- Base 2 → bits
- Base 10 → dits ou Hartleys
- Base e → nats
- Base 3 → trits (porque não?)

Exemplos

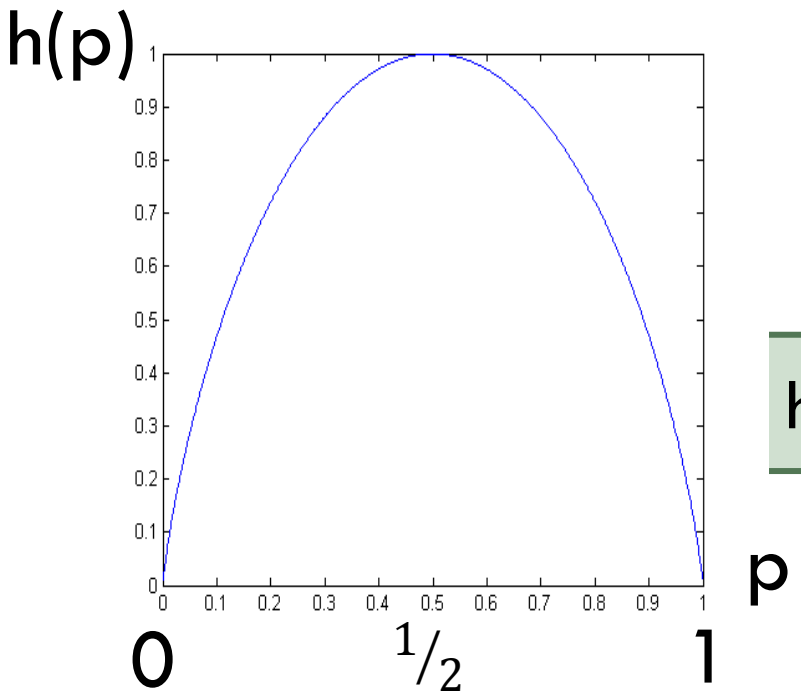
- Ex. 1) $X \in \{0,1\}$, $p(X=0)=0$, $p(X=1)=1$
- $H(X) = - 0 \log 0 - 1 \log 1 = 0$
- Note: $\lim_{P \to 0} p \log p = 0$ por l'Hôpital

Nenhuma incerteza !

X é determinística

Exemplos (continuação)

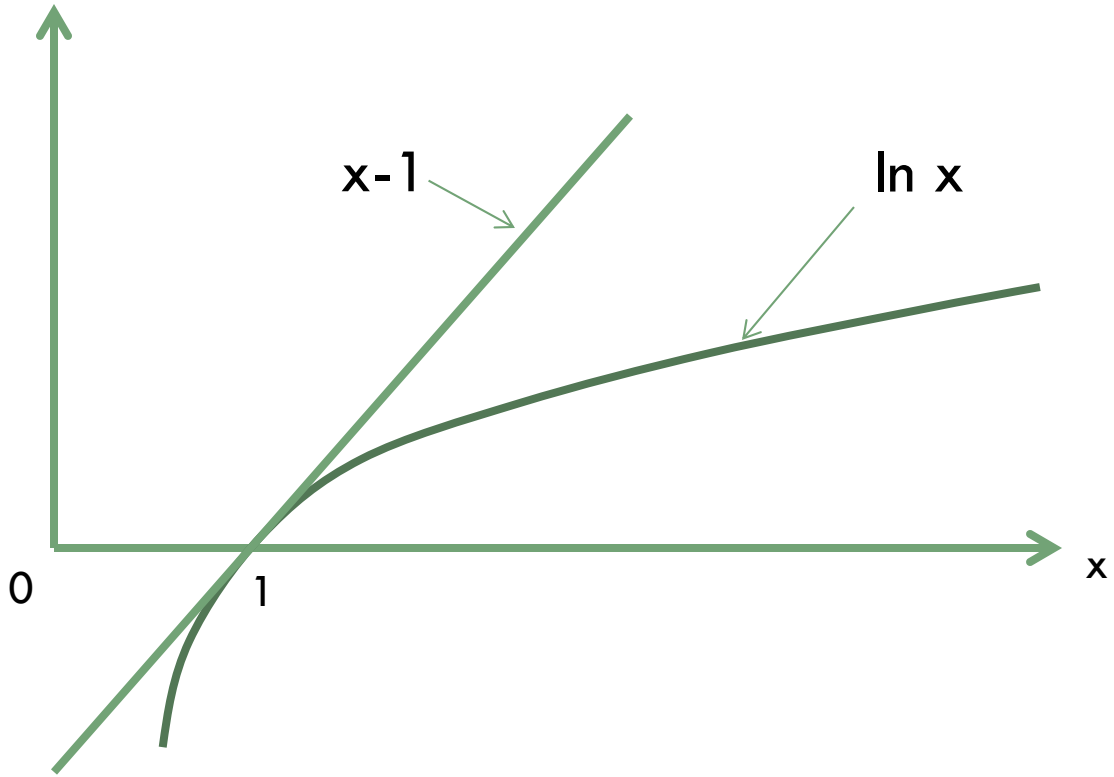
- Ex. 2) $X \in \{0,1\}$, $p(X=0)=p$, $p(X=1)=1-p$,
- $H(X) = -p \log p - (1-p) \log (1-p)$
- $= h(p)$



h(p) é a função binária de entropia

Lema

- $\ln x \leq x-1, \qquad x>0$
- Prova: Série de Taylor com resto



Divergência de Kulbach-Leibler

- Sejam p(x) e q(x) duas funções de massa de probabilidade definidas no alfabeto $\mathcal{X}$.

- *A divergência de Kullbach-Leibler*
- *de p em relação a q é dada por*

$$D(p \parallel q) = \sum_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$$

Proposição: Desigualdade da Informação

$D(p \| q) \geq 0$ com igualdade se e somente se (sse) $p \equiv q$

- Prova: Seja $A = \{x : p(x) > 0\}$
- Use $\ln x \leq x-1$ (Lema)

Observação

- A divergência de K-L é muito útil em T I,
- mas não é uma métrica.

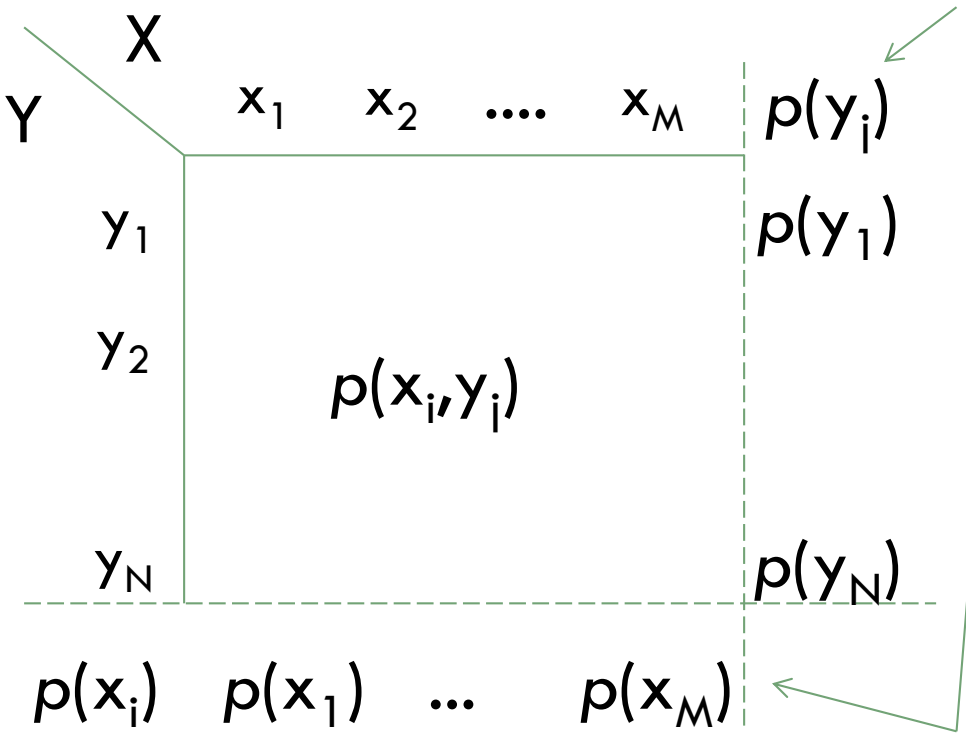
- Ela não é simétrica e não satisfaz a desigualdade do triângulo.

Aplicação

- Seja q a distribuição uniforme
- $q_i = 1/n$ for i=1,...,n
- $p = \{p_1, p_2, ..., p_n\}$
- Então $D(p \| q) \geq 0$
- $$\sum p_i \log \frac{p_i}{q_i} \geq 0$$
- $\sum p_i \log p_i \geq \sum p_i \log q_i = \sum p_i \log 1/n$
- Portanto $H(p) \leq \log n$
- A distribuição uniforme tem máxima entropia.

Distribuições conjunta, marginais e condicionais

- Distribuição conjunta:



Distribuições marginais

$$p(xi) = \sum_{y_j} p(x_i, yj)$$

$$p(y_j) = \sum_{x_i} p(x_i, yj)$$

Distribuições Condicionais:

- $$p(y_j|x_i) = \frac{p(xi,yj)}{p(xi)}$$

- $$p(x_i|y_j) = \frac{p(xi,yj)}{p(y_j)}$$

A distribuição conjunta determina as distribuições marginais e condicionais.

Note: As marginais não determinam a distribuição conjunta.

Entropia Conjunta

- $H(X,Y) = H(\mathbf{p}(\,.\,,.\,)) = \sum_{(x_i, yj)} p(x_i, yj)\ log\ \frac{1}{p(x_i,yj)}$

Entropias Condicionais

- $H(X|Y) = \sum p(x_i, yj)\ log \frac{1}{p(x_i|y_j)} = - E\ (log\ p(X|Y)\ )$

- $H(Y|X) = \sum p(y_j, x_i)\ log \frac{1}{p(y_j|x_i)} = - E\ (log\ p(Y|X)\ )$

Regra da Cadeia (como descascar cebola)

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

- $= H(Y) + H(X|Y)$
- Prova: Sugestão para casa.
- Simples manipulação algébrica.

- Corolário (forma condicional):
- $H(X,Y|Z) = H(X|Z) + H(Y|X,Z)$
- $= H(Y|Z) + H(X|Y,Z)$

Informação Mútua

- A Informação Mútua entre X e Y é
- a divergência de K-L entre a distribuição
- conjunta p(x,y) e o produto das marginais p(x) p(y).

- I(X;Y) = D(p(x,y) || p(x) p(y))

- $$= \sum_{x} \sum_{y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

Propriedades de I(X;Y)

- 1) Não-negatividade: $I(X;Y) \geq 0$, com igualdade
- sse X and Y forem independentes.
- Prova: I(X;Y) é uma divergência de K-L .

- 2) Simetria:
- $I(X;Y) = I(Y;X)$
- Prova: Trivial ($p(x)p(y) = p(y)p(x)$)

Informação Mútua e Entropia

- $I(X;Y) = \sum \sum p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$
- $= H(X) + H(Y) - H(X,Y)$ (simplificando)
- $= H(X) - H(X \mid Y)$ (regra da cadeia)
- $= H(Y) - H(Y \mid X)$ (forma alternativa)

- Note: A Informação mútua entre duas variáveis
- aleatórias é a incerteza residual sobre uma variável quando a outra é revelada.

Informação Mútua

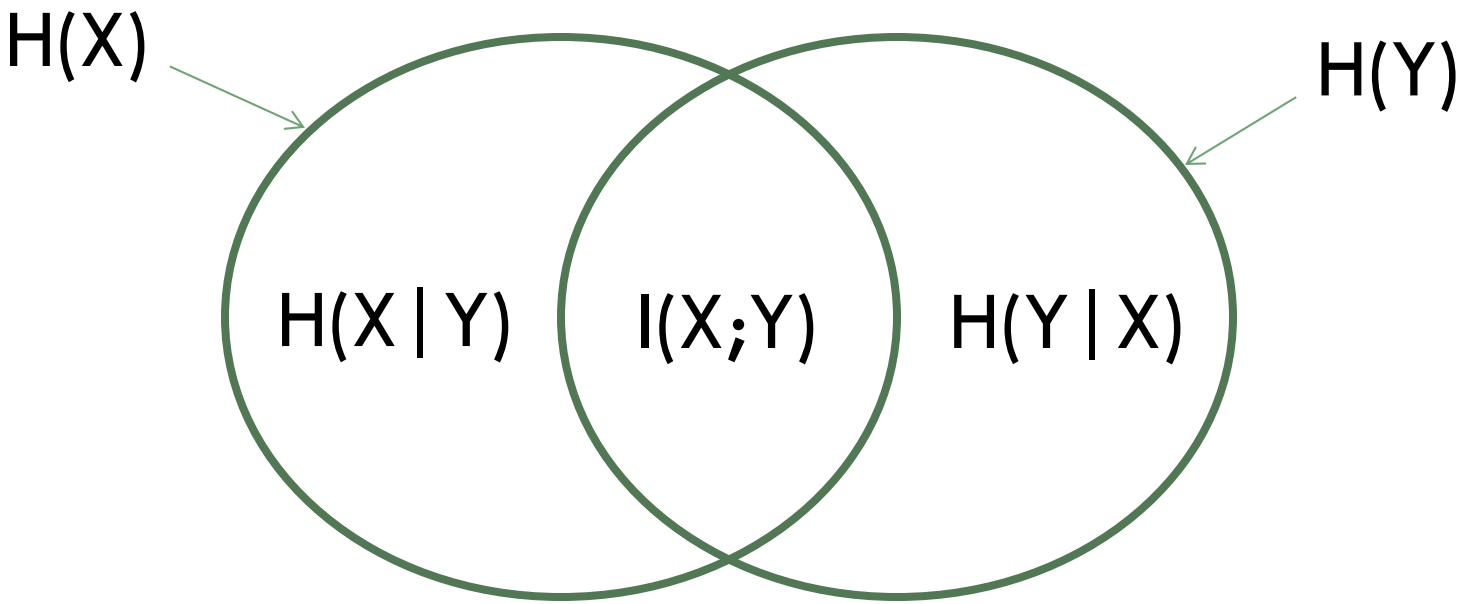
- A informação mútua entre X e Y é a diferença entre as entropias de X antes e depois de conhecer Y:

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

- Também

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

Diagrama de Venn



Informação não atrapalha

- Condicionamento não aumenta entropia:

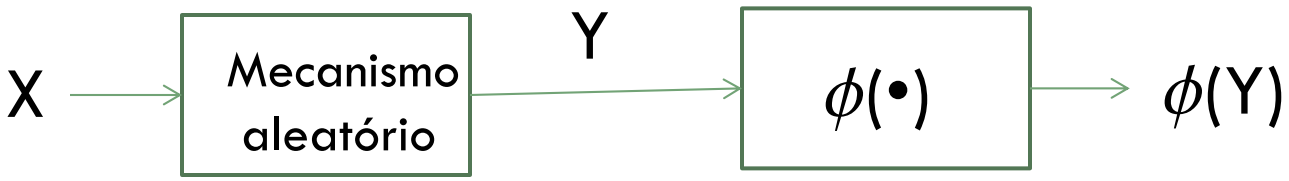
$$H(X|Y) \leq H(X)$$

- Prova: $I(X;Y) = H(X) - H(X|Y) \geq 0$

- Na média o conhecimento de Y não pode aumentar a incerteza sobre X.

Passando informação adiante:

- Sejam X e Y variáveis aleatórias dependentes

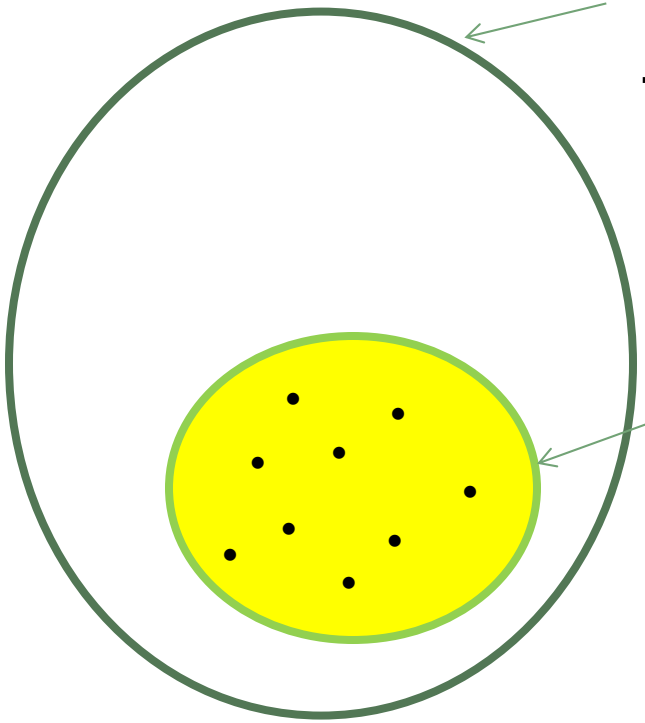


- Proposição: $I(X;Y) \geq I(X; \phi(Y))$
- Prova: $I(X;Y) = H(X) - H(X|Y)$
- $= H(X) - H(X|Y, \phi(Y))$
- $\geq H(X) - H(X|\phi(Y)) = I(X; \phi(Y))$ Condicionamento reduz entropia

- Desigualdade de Processamento de dados.

Propriedade da Equipartição Assintótica

- Sejam $X_1, X_2, \ldots, X_n$ i.i.d. segundo p(x)



Espaço amostra = conjunto de todas sequências $(x_1, x_2, \ldots, x_n)$

A_ϵ = Conjunto de sequências típicas

P.E.A.

1) $\Pr\{A_\epsilon\} \geq 1-\epsilon$
2) $p(x) \cong 2^{-nH(X)}$
3) $\|A_\epsilon\| \cong 2^{nH(X)}$

P.E.A é o DNA de T.I. !

Exemplo de sequências típicas

- Seja X uma moeda viciada com
- $P(Cara)=0.9 \quad e \quad P(Coroa) = 0.1$

- Considere o conjunto de 1000 lançamentos da moeda.
- Sequências típicas são aquelas aproximente 900 Caras e 100 Coroas.

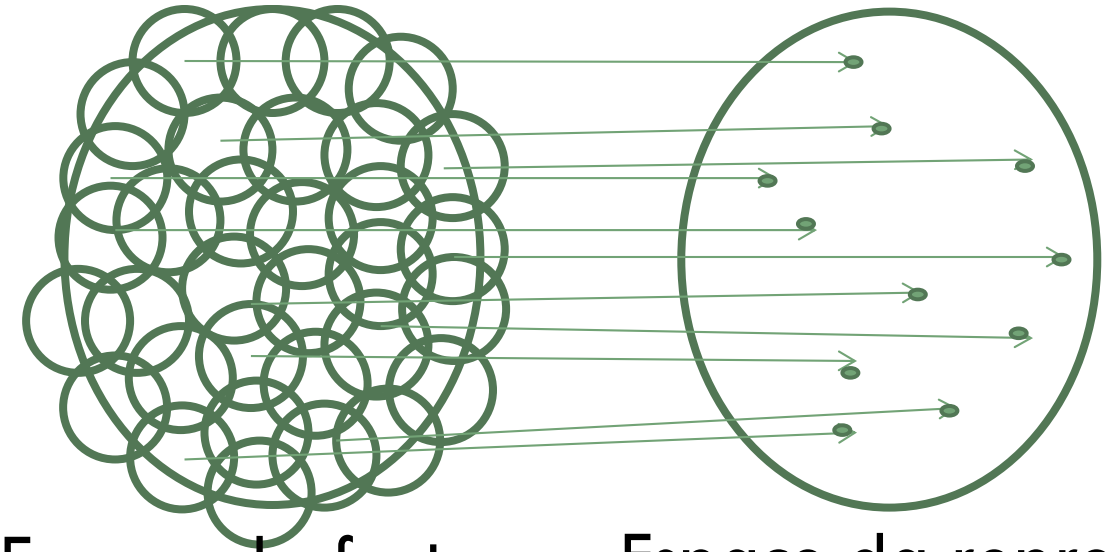
- Note: A sequência mais provável, qual seja a de 1000 Caras, não é típica !

Conclusão da P.E.A. :

- Melhor apostar em A_{ϵ} !

Compressão de dados (Codificação de fonte)

- Objetivo: representar a fonte eficientemente.



Código eficiente: Código de Shannon

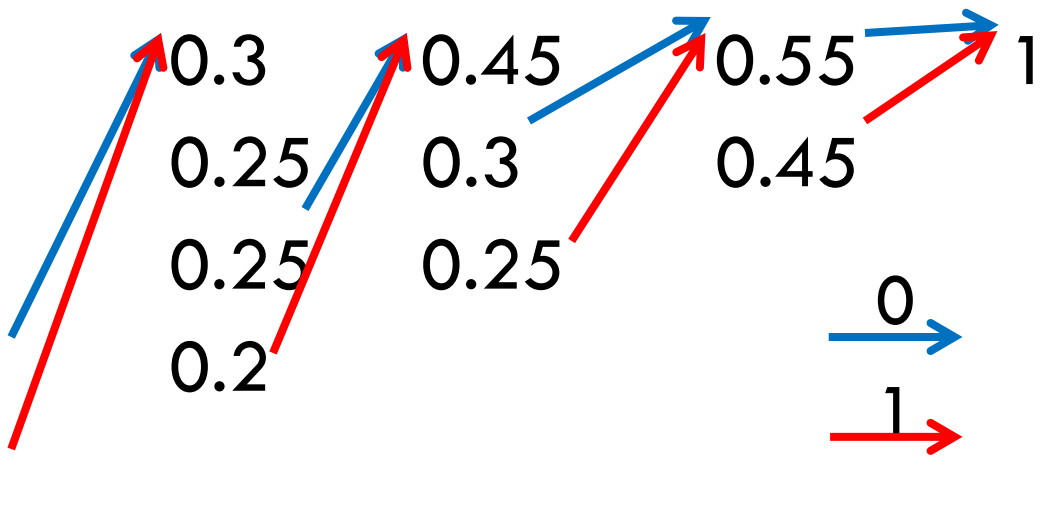
- Seja $\ell_i = \left\lceil log_D \frac{1}{p_i} \right\rceil$,
- Use uma palavra de código com este comprimento.

- Este código satisfaz
- $H_D(X) \leq L \leq H_D(X) + 1$

Código eficiente: Huffmann (1952)

- O código de Huffman é o código ótimo de prefixo (menor comprimento esperado) para uma dada distribuição p(x).
- Exemplo:

X	p(x)	Código
1	0.25	01
2	0.25	10
3	0.2	11
4	0.15	000
5	0.15	001



Este código tem comprimento médio de 2.3 bits/símbolo.

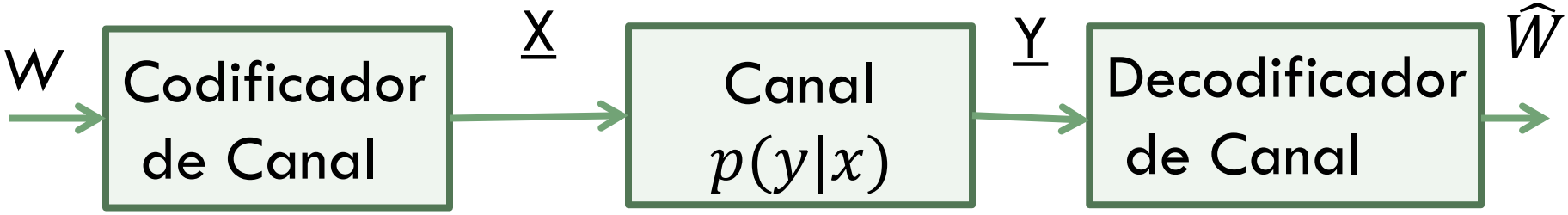
Códigos eficientes:

Outros códigos eficientes:

- Códigos Shannon-Fano-Elias
- Códigos Aritméticos
- Códigos de Lempel-Ziv (Universal – aprende a distribuição da fonte)
- Códigos de corridas + códigos de Golomb (muito simples)

Transmissão por Canais Ruidosos

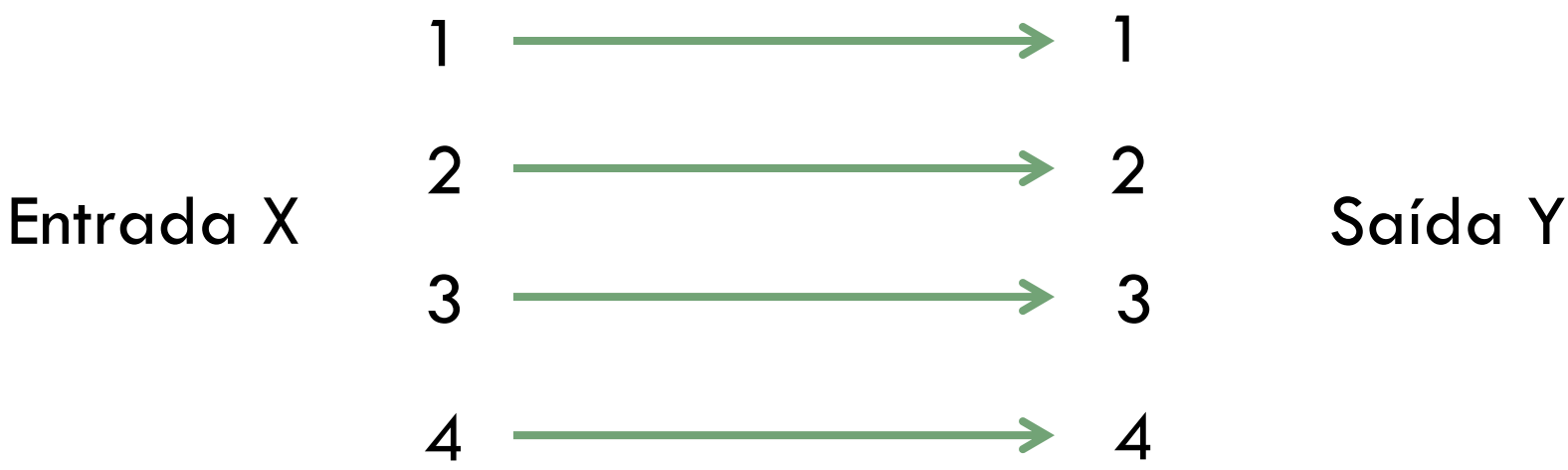
- O Problema da codificação de canal:



- $W \in \{1,2,\ldots,2^{nR}\}$ = conjunto de mensagens (Taxa R)
- $\underline{X} = (x_1\ x_2\ \ldots\ x_n)$ = palavra-código de entrada no canal
- $\underline{Y} = (y_1\ y_2\ \ldots\ y_n)$ = palavra-código de saída do canal
- $\widehat{W}$ = mensagem decodificada P(erro) = $P\{W \neq \widehat{W}\}$

Exemplos

- Máquina de escrever sem ruído:

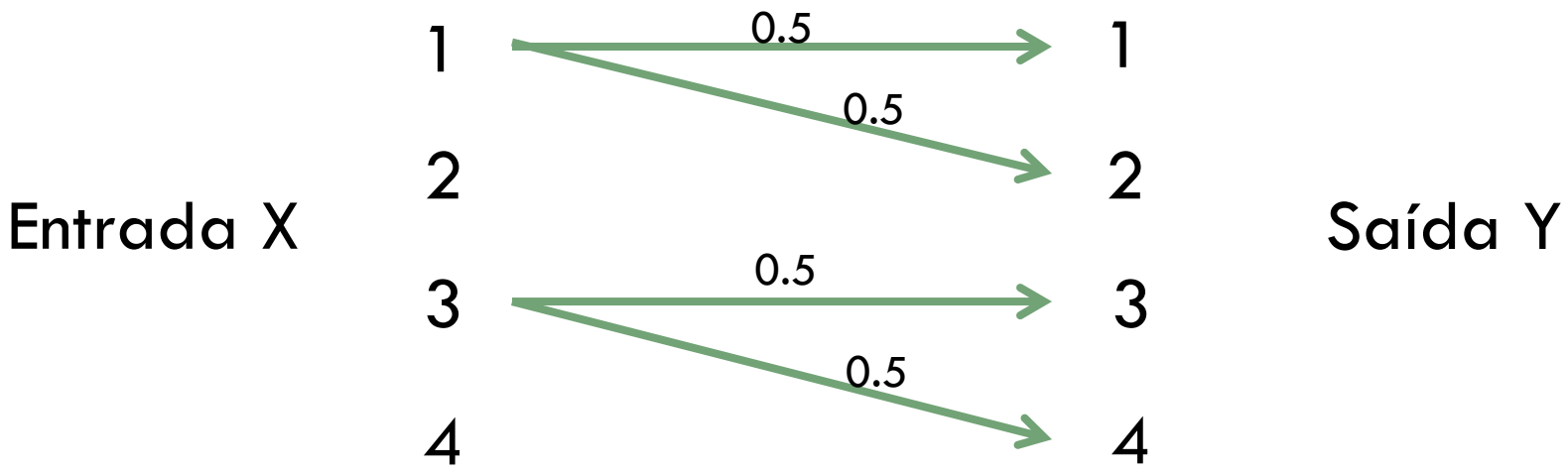


Número de símbolos sem ruído $= 4$

Pode-se transmitir R $= log_2\ 4 = 2$ bits/transmissão

Exemplos simples

- Máquina de escrever ruidosa (tipo 1):

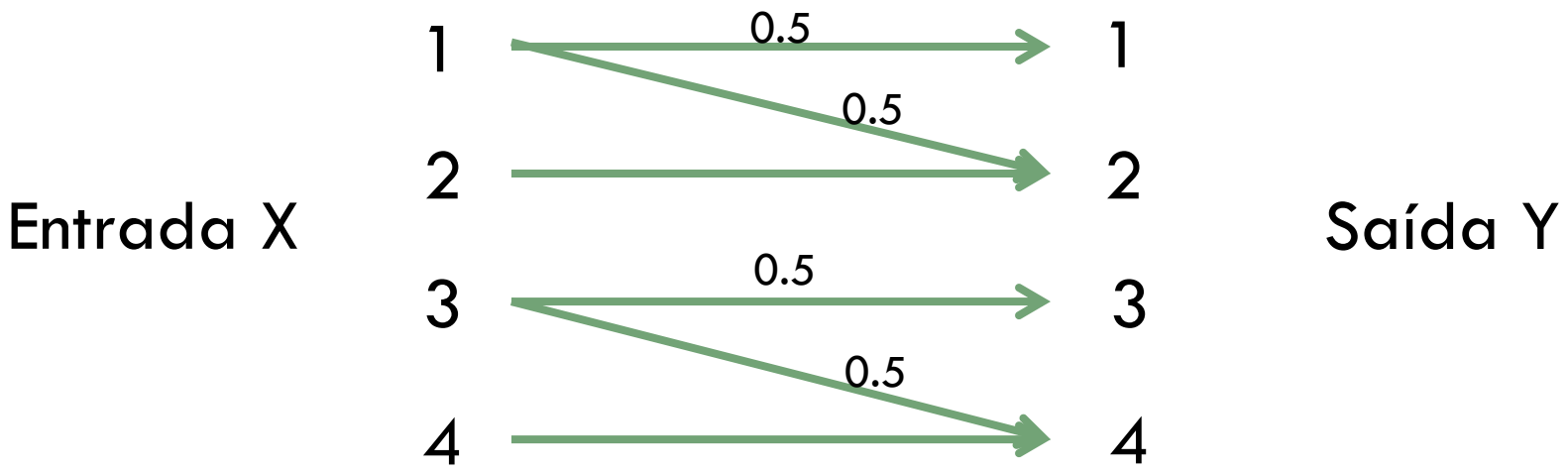


Número de símbolos sem ruído = 2

Pode-se transmitir R = $log_2\ 2 = 1$ bit/transmissão

Exemplos simples

- Máquina de escrever ruidosa (tipo 2):

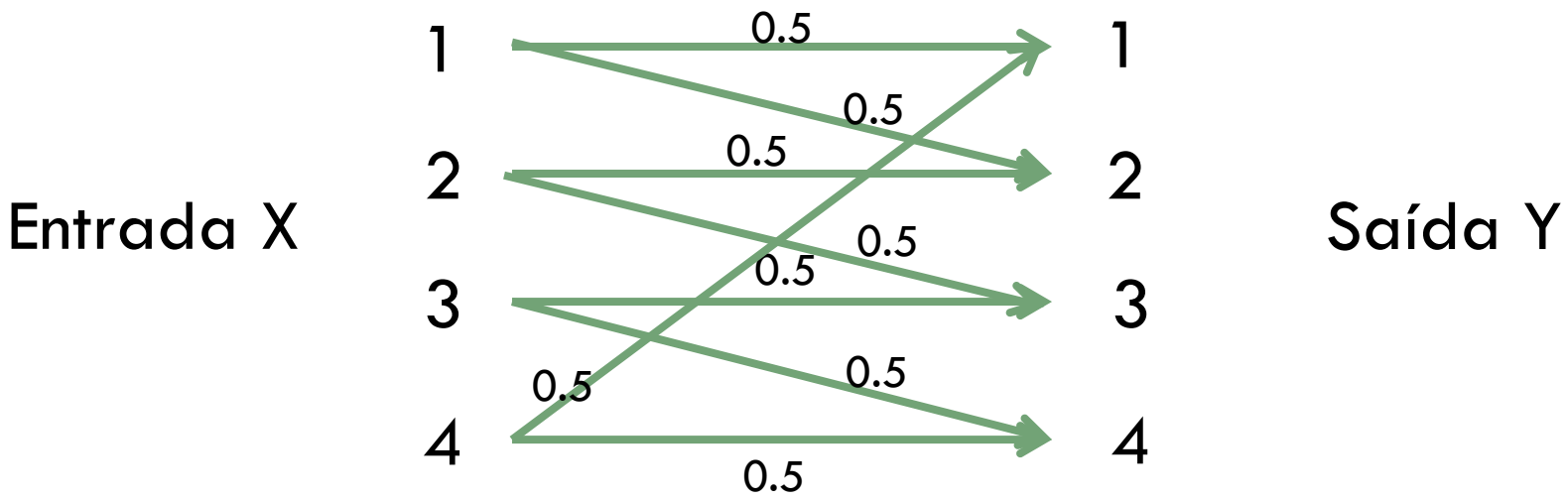


Número de símbolos sem ruído = 2

Pode-se transmitir R = log_2 2 = 1 bit/transmissão

Exemplos simples

- Máquina de escrever ruidosa (tipo 3):

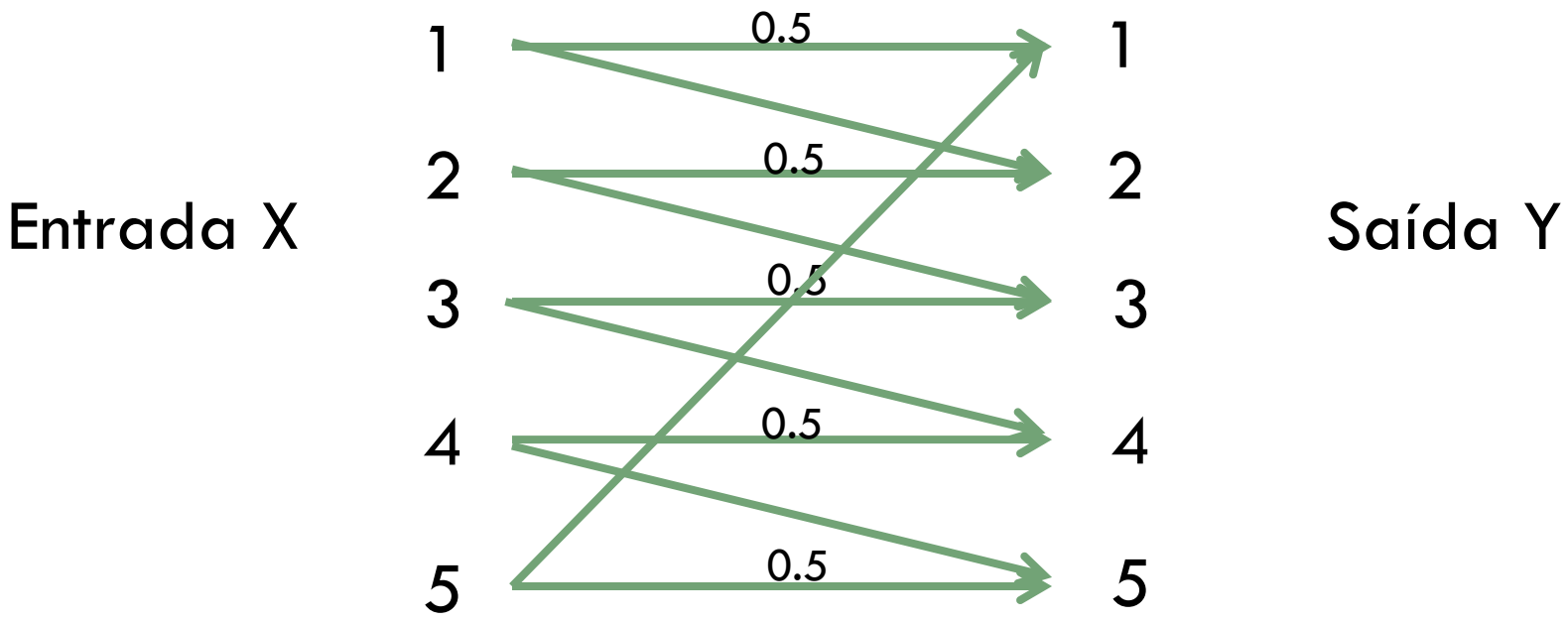


Número de símbolos sem ruído = 2 Use só X=1 e X=3

Pode-se transmitir R = $log_2\ 2 = 1$ bit/transmissão

Exemplos simples

- Máquina de escrever mais complicada:

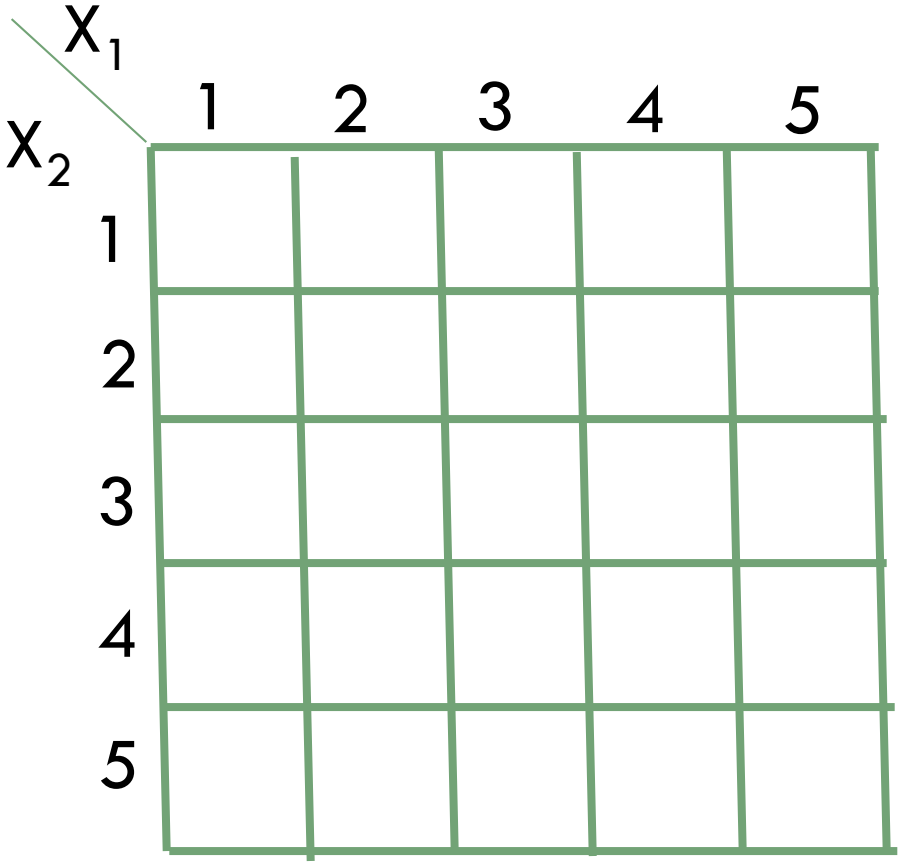


Quantos símbolos livres de ruído?

Claramente pelo menos 2, talvez mais.

Exemplos

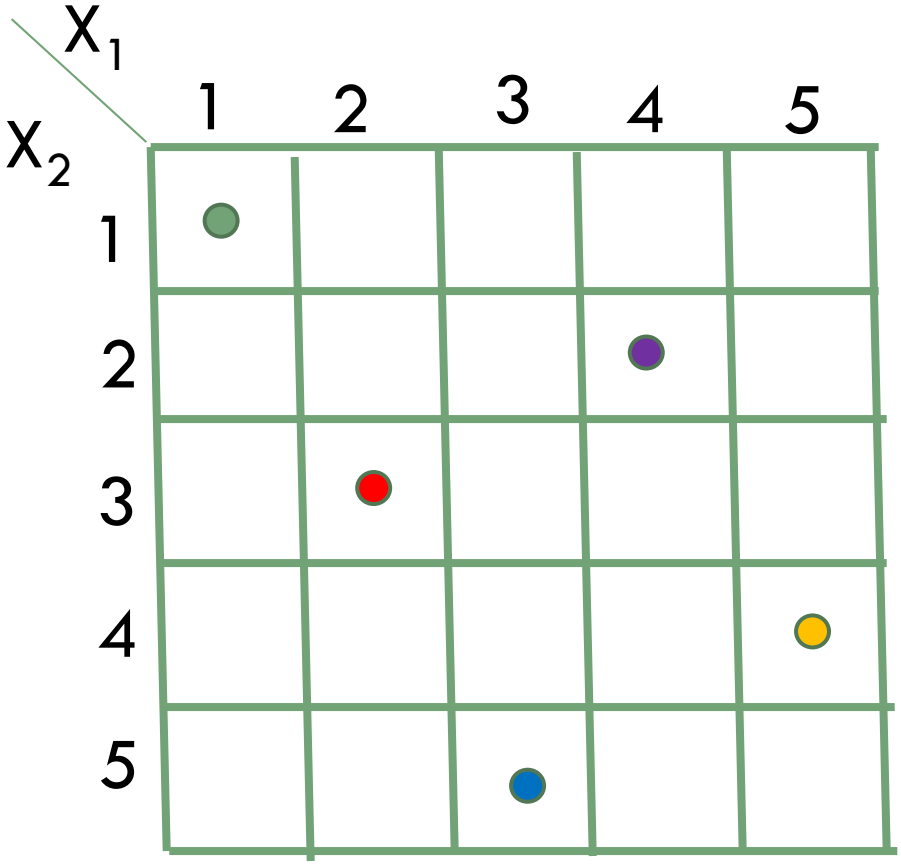
- Considere dois usos consecutivos do canal:



Código:
Que quadrados
Escolher ?

Exemplos simples

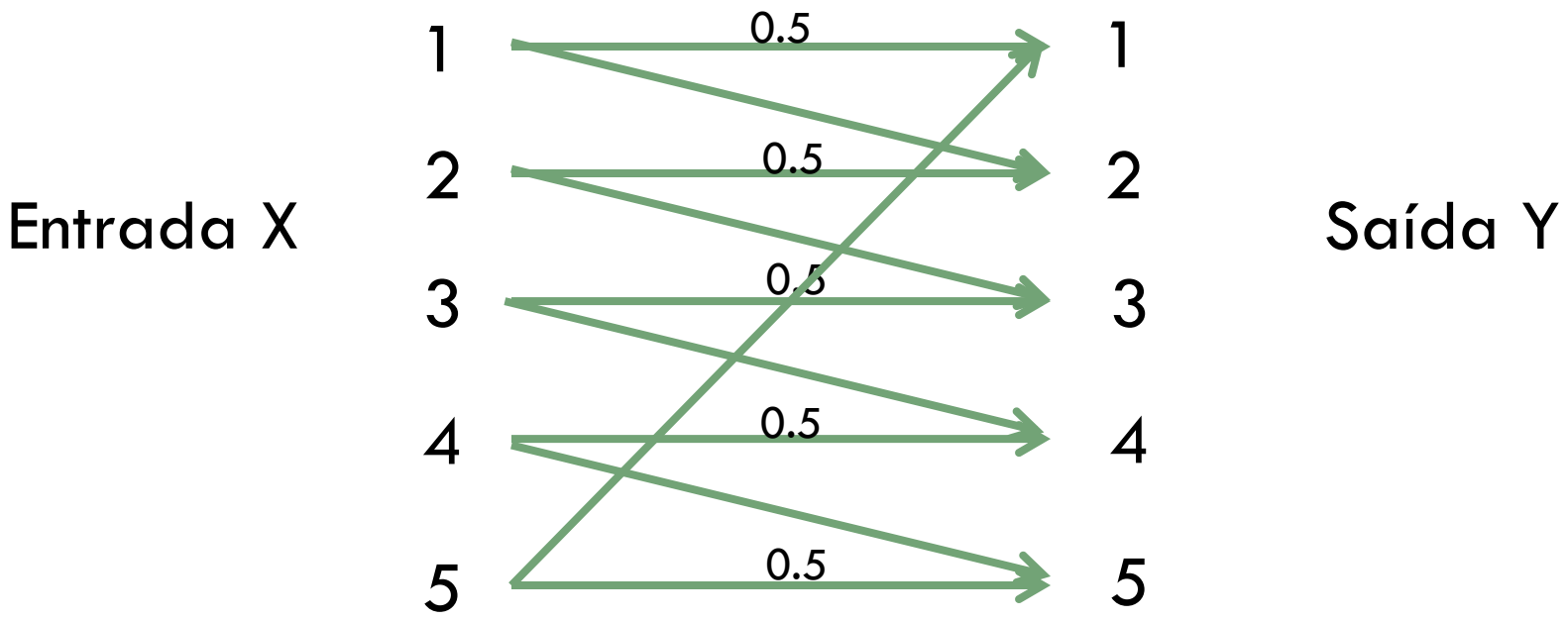
- Dois usos consecutivos do canal:



Sejam os pontos $\{X_1, X_2\} = \{(1,1), (2,3), (3,5), (4,2), (5,4)\}$

Relembrando o canal

- Máquina de escrever complicada:

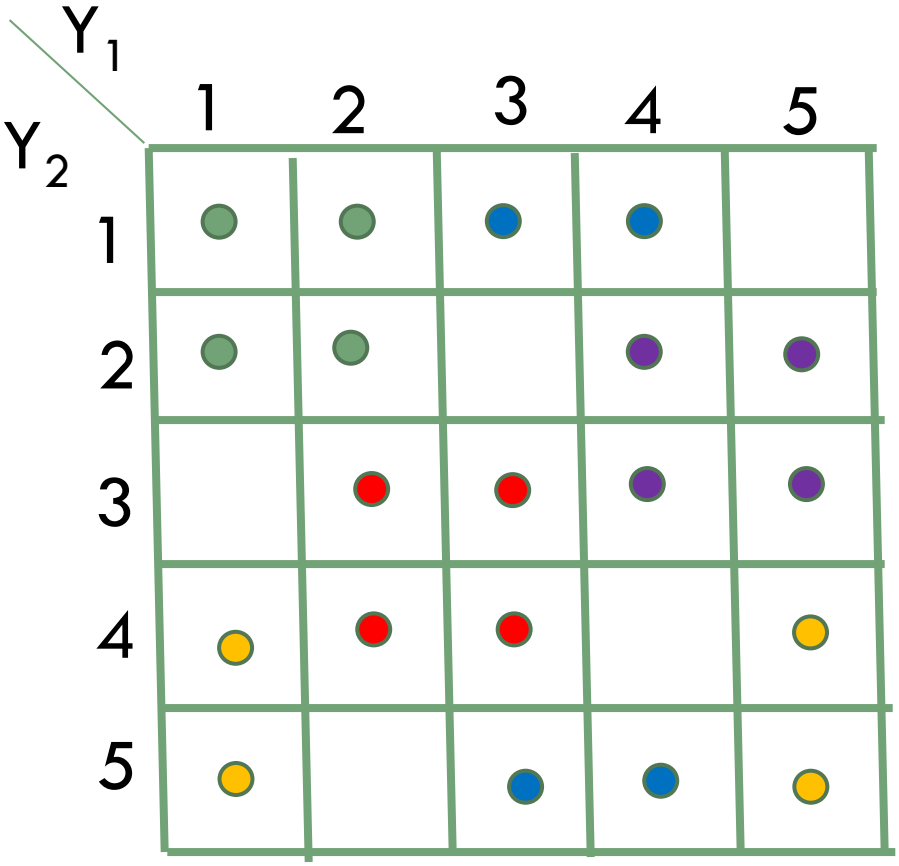


Quantos símbolos livres de ruído?

Mais de 2 ?

Exemplos simples

- Olhando as saídas do canal:



Sejam $\{X_1,X_2\}$ = {(1,1), (2,3), (3,5), (4,2), (5,4)}

Exemplo simples - observações

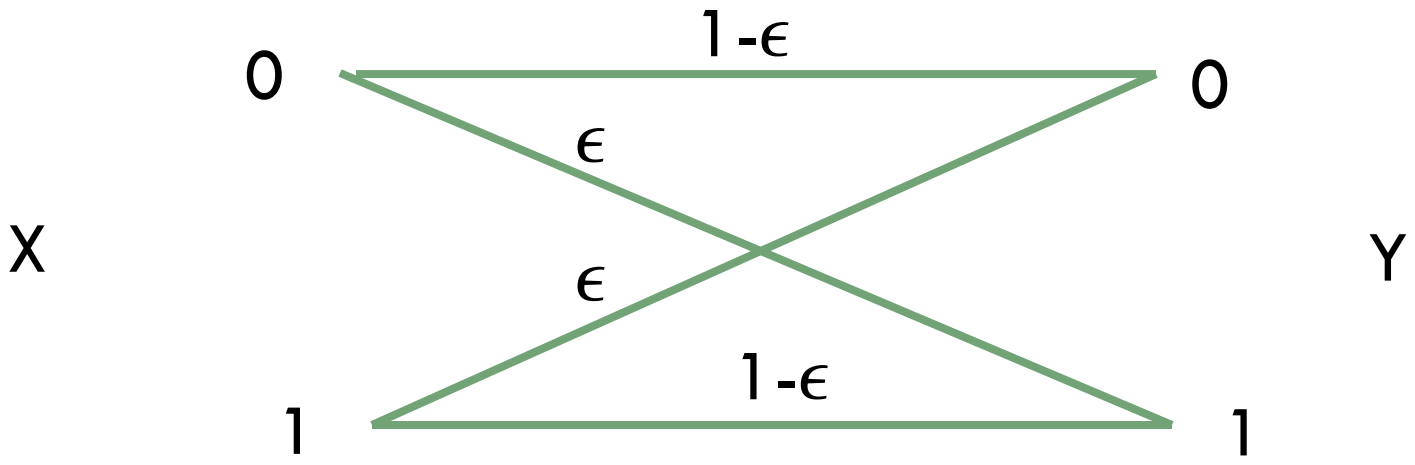
- Pode-se transmitir 5 símbolos livres de ruído em n=2 transmissões.

- Portanto tem-se taxa de $\frac{log_2 5}{2} = 1.16$ bits/transmissão
- com P(erro) = 0.

- Pode-se usar códigos mais longos (n→∞) para
- obter $\log_2 (5/2) = 1.32$ bits/transmissão, a
- capacidade do canal.

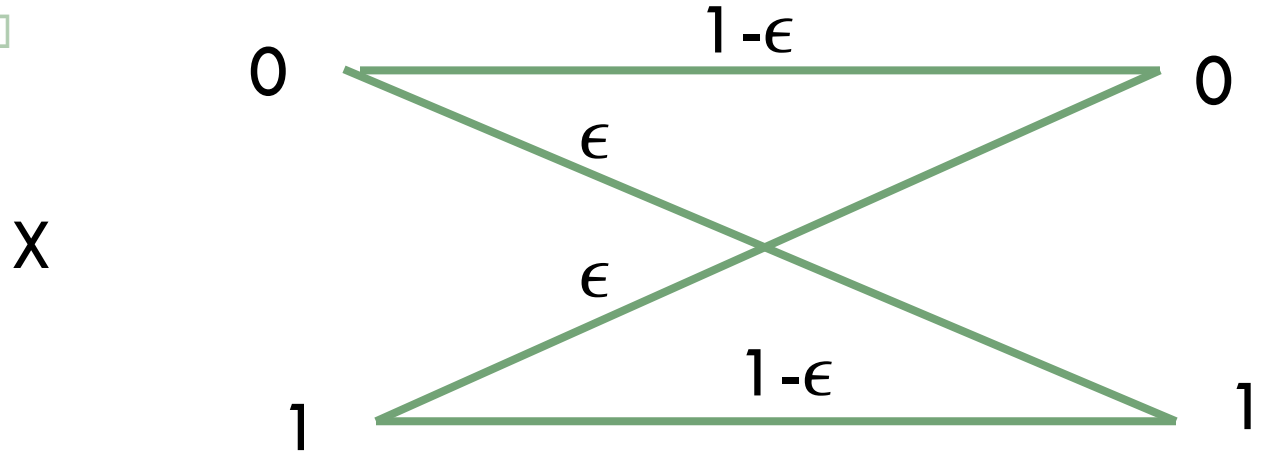
Exemplos: Canal Binário Simétrico (BSC)

- Quantos símbolos livres de ruído ?

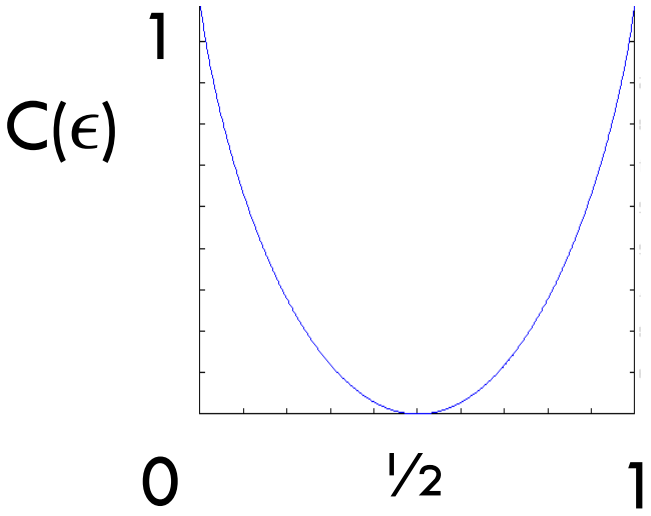


A.: Claramente para n=1 não há nenhum.
Que tal para n grande ?

Exemplo: Canal Binário Simétrico (BSC)



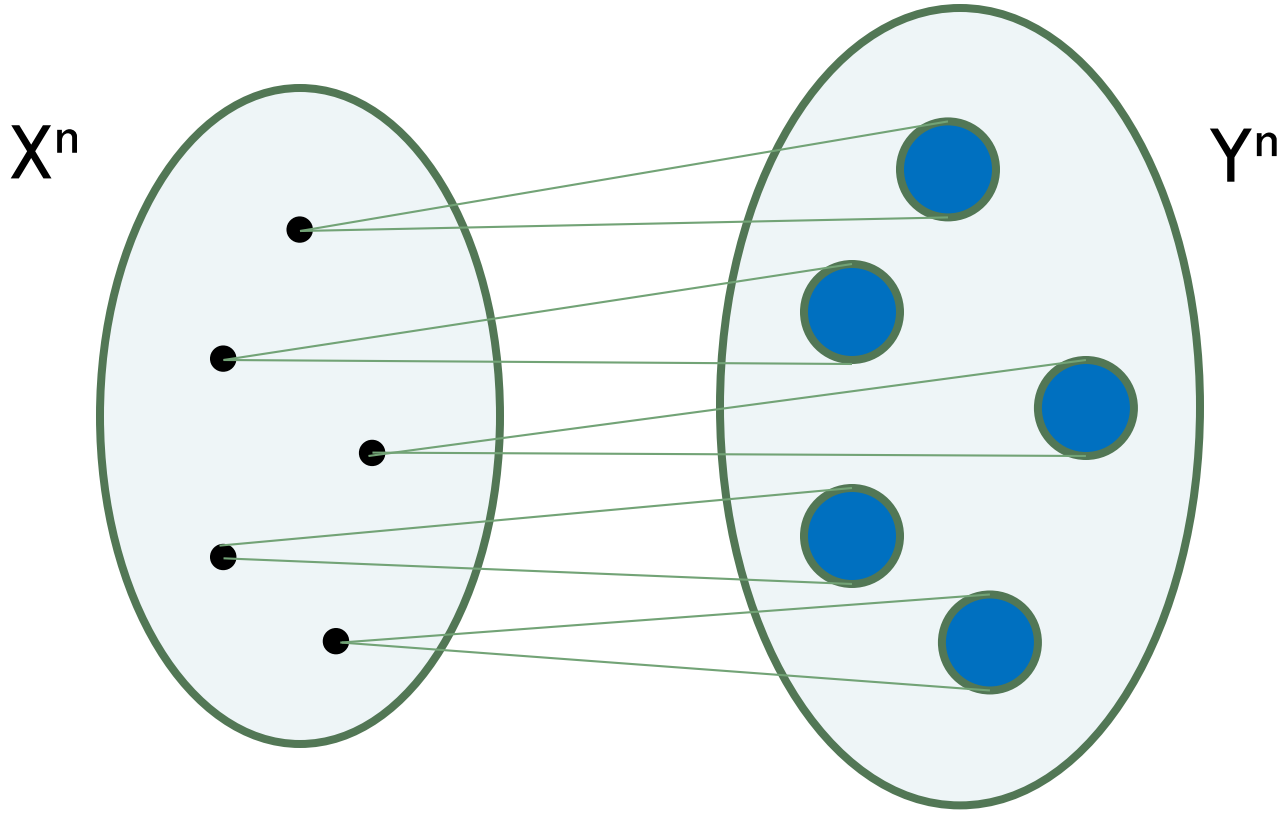
- $C = \max (H(Y) - H(Y|X))$
- $= 1 - h(\epsilon)$ bits/transmissão



- Note: C=0 para $\epsilon = ½$

Segundo Teorema de Shannon

- Usa-se o canal n vezes:



Segundo Teorema de Shannon

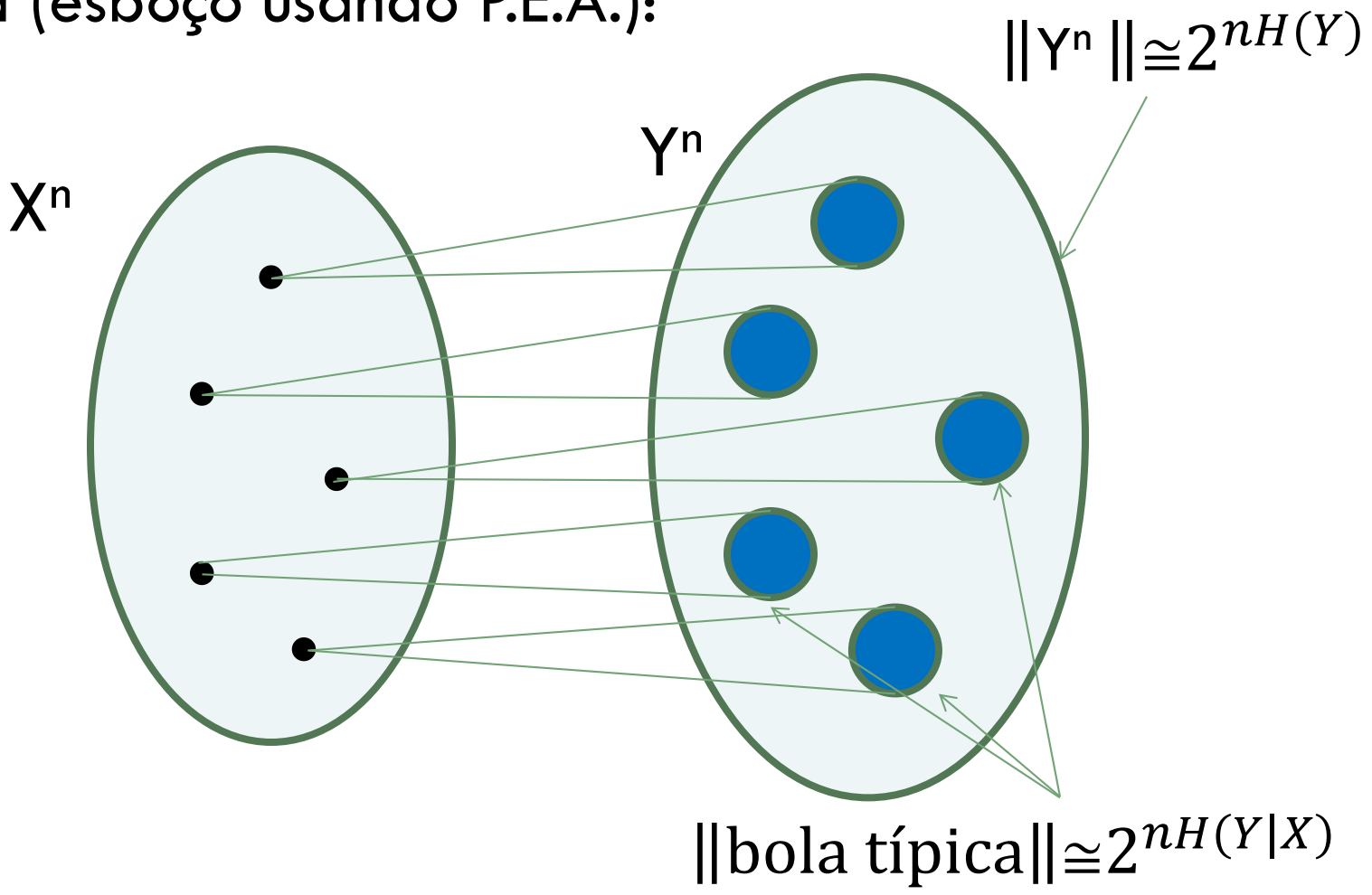
- A Capacidade de um canal discreto sem memória
- é
-

$$C = \max_{p(x)} I(X;Y).$$

- Note:
- $I(X;Y)$ é uma function de $p(x,y) = p(x)p(y|x)$.
- Mas $p(y|x)$ é fixada pelo canal.

Segundo Teorema de Shannon

- Prova (esboço usando P.E.A.):



Entropia Diferencial

- Seja X uma variável aleatória contínua com densidade de probabilidade $f(x)$ e suporte S.
- A entropia diferencial de X é dada por

$$h(X) = -\int_S f(x) \log f(x) dx \qquad \text{(se existir).}$$

Note: Também é denotada por $h(f)$.

Exemplos: Distribuição Uniforme

- Seja X uniforme no intervalo $[0, a]$. Então
- $f(x) = \frac{1}{a}$ no interval e $f(x) = 0$ fora dele.

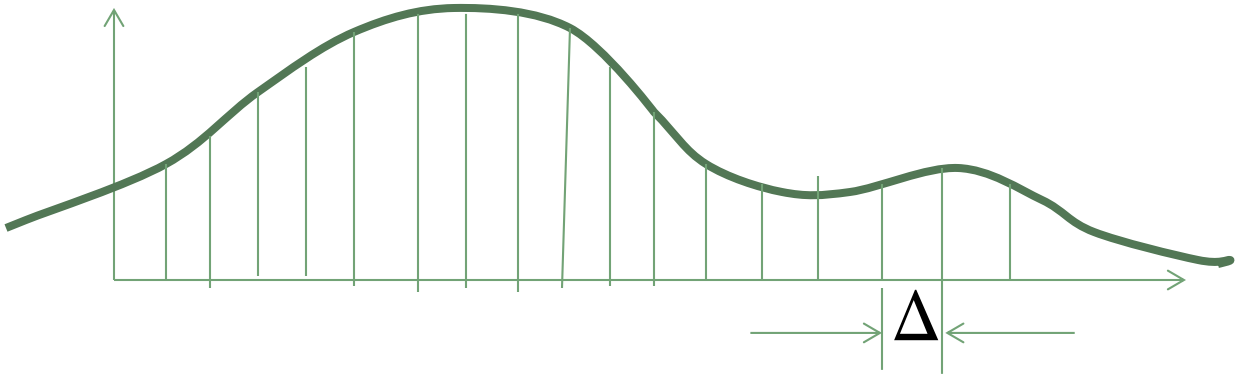
- $h(X) = -\int_0^a \frac{1}{a}\ log\ \frac{1}{a}\, dx = \log a$
- Note que $h(X)$ pode ser negativa (quando $a < 1$).
- No entanto, $2^{h(f)} = 2^{\log a} = a$ é o tamanho do conjunto-suporte, que é não-negativo.

Exemplo: Distribuição Gaussiana

- Seja $X \sim \phi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(\frac{-x^2}{2\sigma^2}\right)$
- Então $h(X) = h(\phi) = -\int \phi(x)[-\frac{x^2}{2\sigma^2} - ln\sqrt{2\pi\sigma^2}]\, dx$
- $= \frac{EX^2}{2\sigma^2} + \frac{1}{2} ln\ 2\pi\sigma^2$
- $= \frac{1}{2} ln\ 2\pi e\sigma^2$ nats
- Mudando a base tem-se $h(X) = \frac{1}{2} log_2\ 2\pi e\sigma^2$ bits

Relação entre Entropias Diferencial e Discreta

- Considere uma quantização de X, denotada por X^{Δ}



- Seja $X^{\Delta} = x_i$ dentro do intervalo i.

Então $H(X^{\Delta})$ $= - \sum_i p_i \log p_i$

$= - \sum_i \Delta f(x_i) \log f(x_i) - \log \Delta$

$\cong h(f) - \log \Delta$

Entropia diferencial de um vetor Gaussiano

- Teorema: Seja $\boldsymbol{X}$ um vetor Gaussiano *n*-dimensional com média μ e matriz covariância K. Então

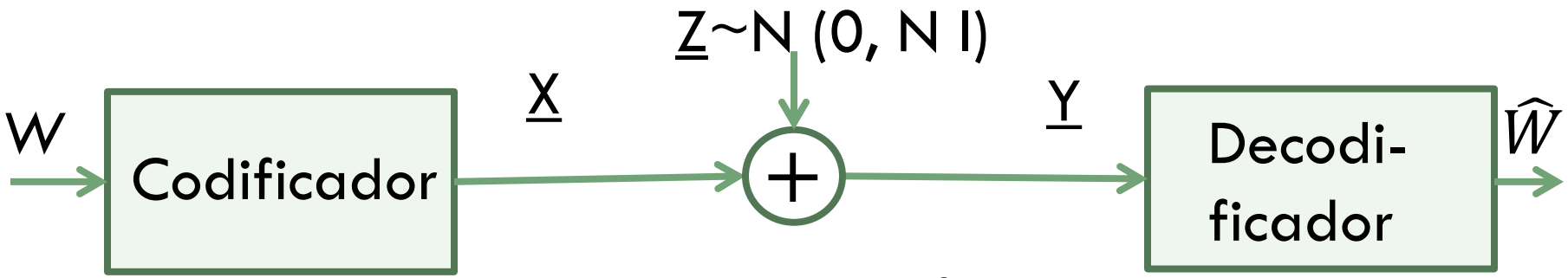
- $$h(\boldsymbol{X}) = \frac{1}{2}\log\left((2\pi e)^n |K|\right)$$

- onde $|K|$ denota o determinante de K.

- Prova: Manipulação algébrica.

O Canal Gaussiano

- O Problema do canal Gaussiano:



$W \in \{1,2,\ldots,2^{nR}\}$ = conjunto de mensagens de taxa R

- $\underline{X} = (x_1\ x_2\ \ldots\ x_n)$ = entrada do canal
- $\underline{Y} = (y_1\ y_2\ \ldots\ y_n)$ = saída do canal
- $\widehat{W}$ = mensagem decodificada P(erro) = $P\{W \neq \widehat{W}\}$

O canal Gaussiano

-
- $\mathrm{C}apacidade \qquad C = \max_{f(x):\, EX^2 \leq P} I(X;Y)$

- $I(X;Y) = h(Y) - h(Y|X) = h(Y) - h(X+Z|X)$
- $\quad = h(Y) - h(Z) \leq \frac{1}{2}\log 2\pi e(P+N) - \frac{1}{2}\log 2\pi eN$
- $\quad = \frac{1}{2}\log\left(1+\frac{P}{N}\right)$ bits/transmissão

O Canal Gaussiano

- $\mathrm{C}apacidade$: $C = \max_{f(x):\, EX^2 \leq P} I(X;Y)$

- $$C = \frac{1}{2}\log\left(1 + \frac{P}{N}\right) \text{ bits/transmissão}$$

O Canal Gaussiano de Banda Limitada

- $$C = W \log(1 + \frac{P}{N_0 W}) \text{ bits/segundo}$$

- Note: Se W→ ∞

- tem-se $C = \frac{P}{N_0} log_2 e$ bits/segundo.

Canal Gaussiano de Banda W

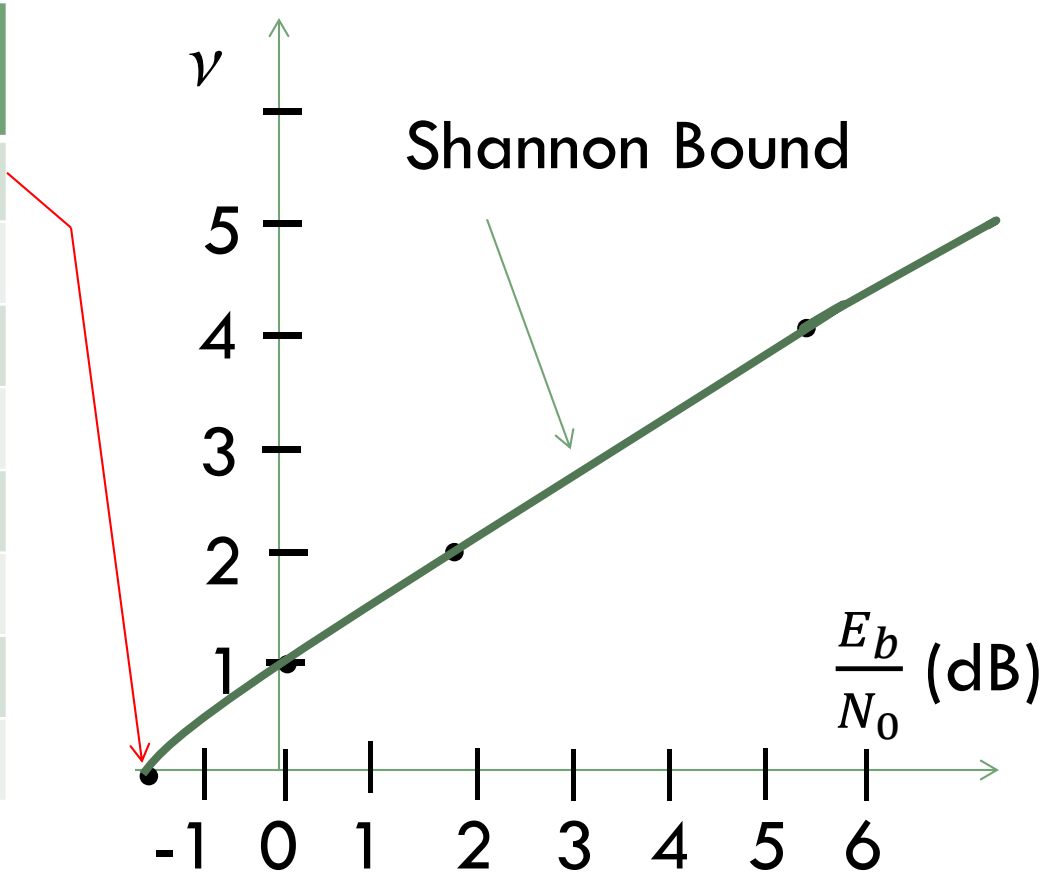
- Seja $\frac{R}{W}$ a eficiência espectral ν em bits por segundo por Hertz. Também seja $P = E_b R$ onde E_b é a energia disponível por bit de informação.
- Obtem-se
- $$\frac{R}{W} \leq \frac{C}{W} = \log(1 + \frac{E_b R}{N_0 W}) \text{ bits/segundo.}$$
- Portanto
- $$\frac{E_b}{N_0} \geq \frac{2^{\nu} - 1}{\nu}$$

Esta relação define o chamado Limitante de Shannon.

Limitante de Shannon

$$\frac{E_b}{N_0} \geq \frac{2^{\nu}-1}{\nu}$$

ν	$\frac{E_b}{N_0}$	$\frac{E_b}{N_0}$ (dB)
→0	0.69	-1.59
0.1	0.718	-1.44
0.25	0.757	-1.21
0.5	0.828	-0.82
1	1	0
2	1.5	1.76
4	3.75	5.74
8	31.87	15.03



Solução de “Water-Filling” (Shannon)

Canais Gaussianos Paralelos

Exemplo de “Water Filling”

- Canais com níveis de ruído 2, 1 and 3.
- Potência disponível = 2
- Capacidade= $\frac{1}{2}\log\left(1+\frac{0.5}{2}\right)+\frac{1}{2}\log\left(1+\frac{1.5}{1}\right)+\frac{1}{2}\log\left(1+\frac{0}{3}\right)$
- Nível de ruído mais potência de sinal = 2.5
- Nenhuma potência alocada para o terceiro canal.

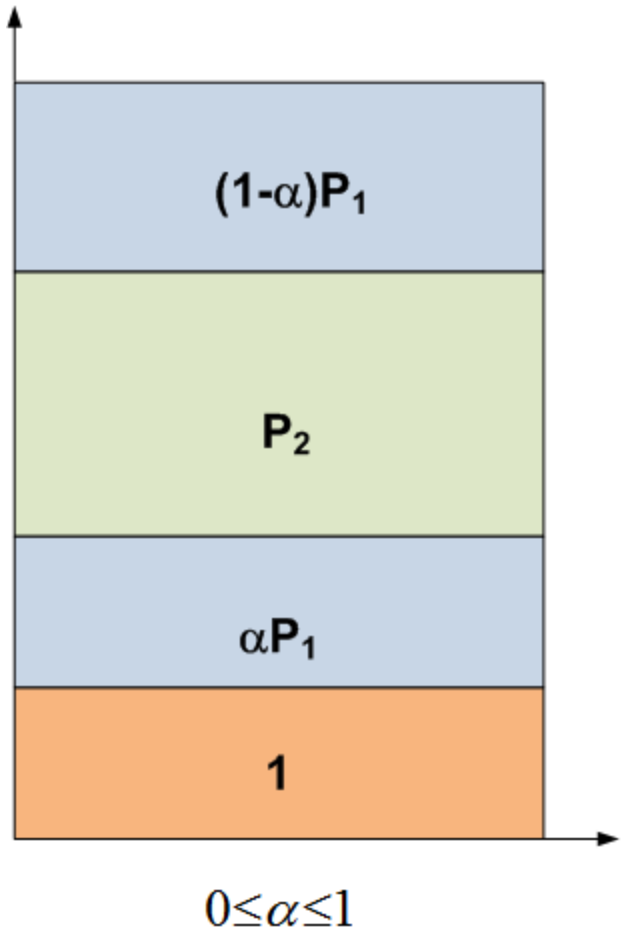
Teoria de informação de múltiplos usuários

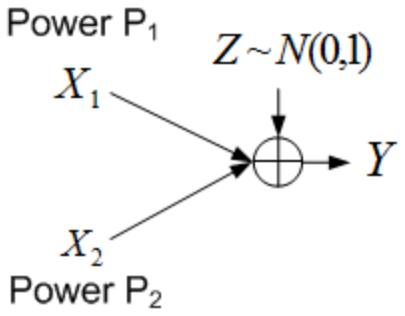
- Blocos constituintes:

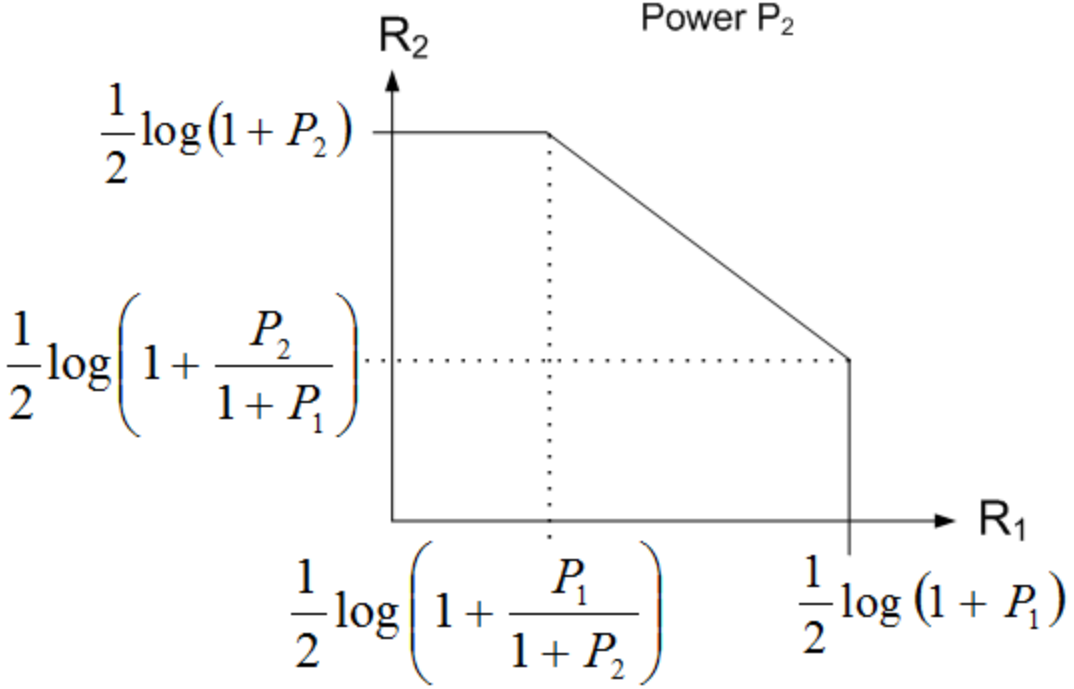
- Canal de Múltiplo Acesso (MACs)
- Canais de Broadcast (BCs)
- Canais de Interferência (IFCs)
- Canais de Relay (RCs)

- Note: Estes canais têm versões discretas sem memória e versões Gaussianas. Por simplicidade vamos ver apenas as versões Gaussianas.

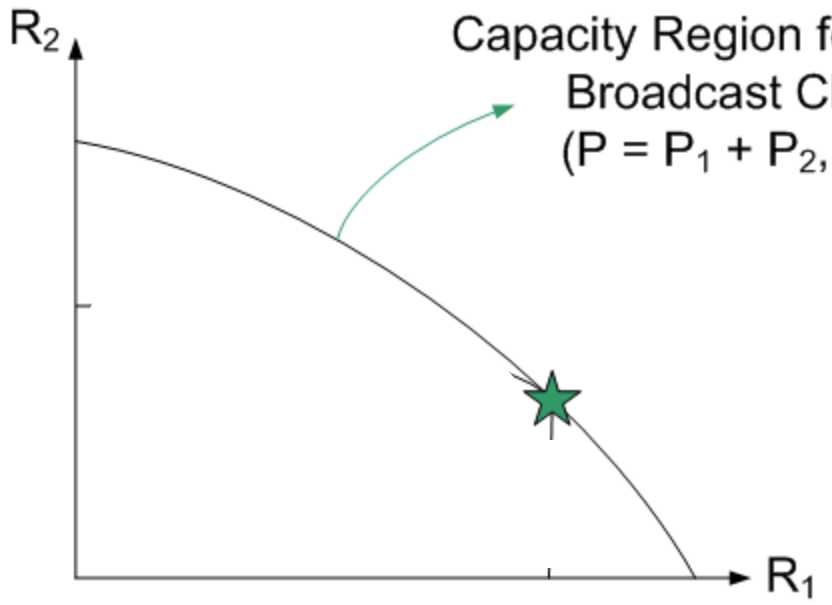
Canal de Múltiplo Acesso (MAC)

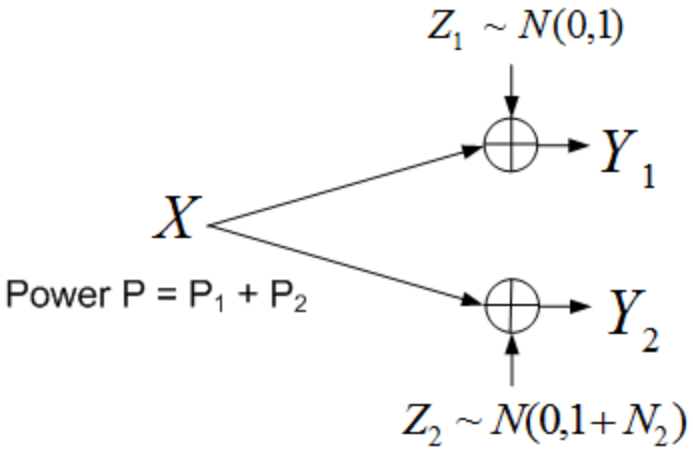






Canal de Broadcast Gaussiano





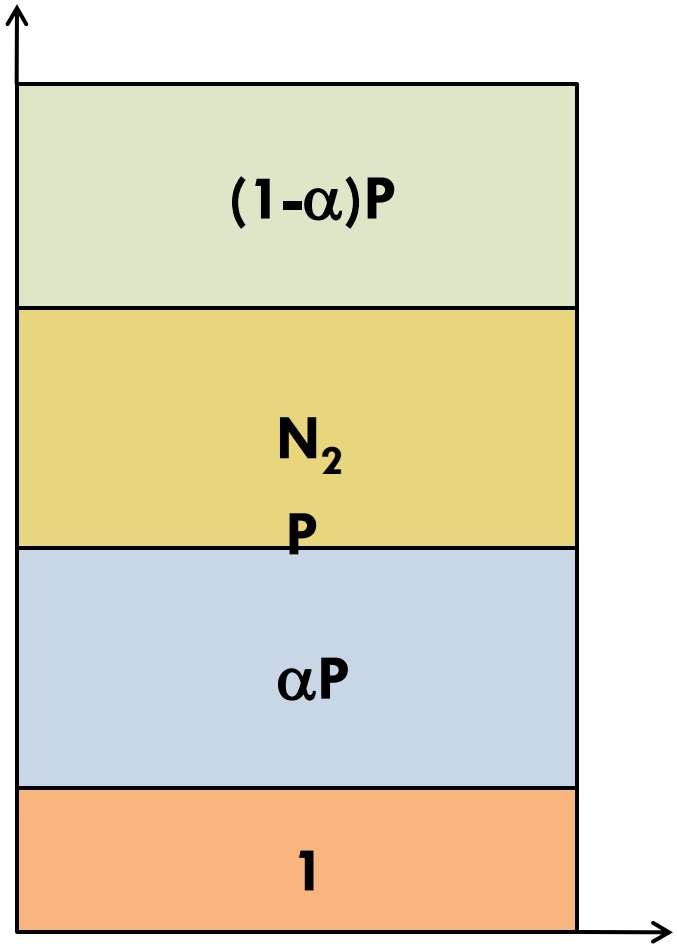
$$C_{BC}\{R_1, R_2\}:$$

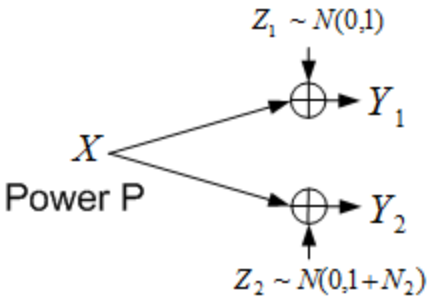
$$0 \le \alpha \le 1$$

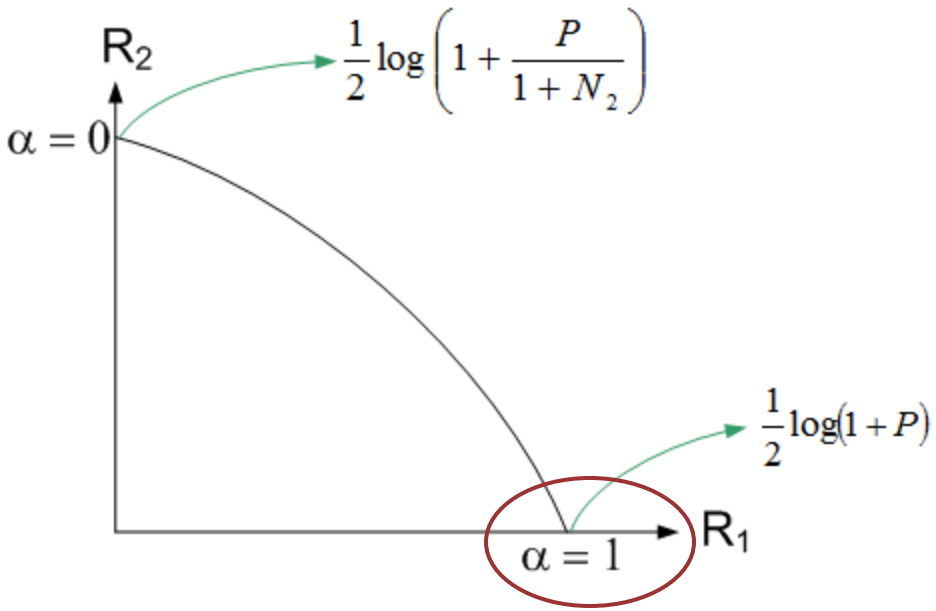
$$0 \le R_1 \le \frac{1}{2}\log(1 + \alpha P)$$

$$0 \le R_2 \le \frac{1}{2}\log\left(1 + \frac{(1-\alpha)P}{1 + N_2 + \alpha P}\right)$$

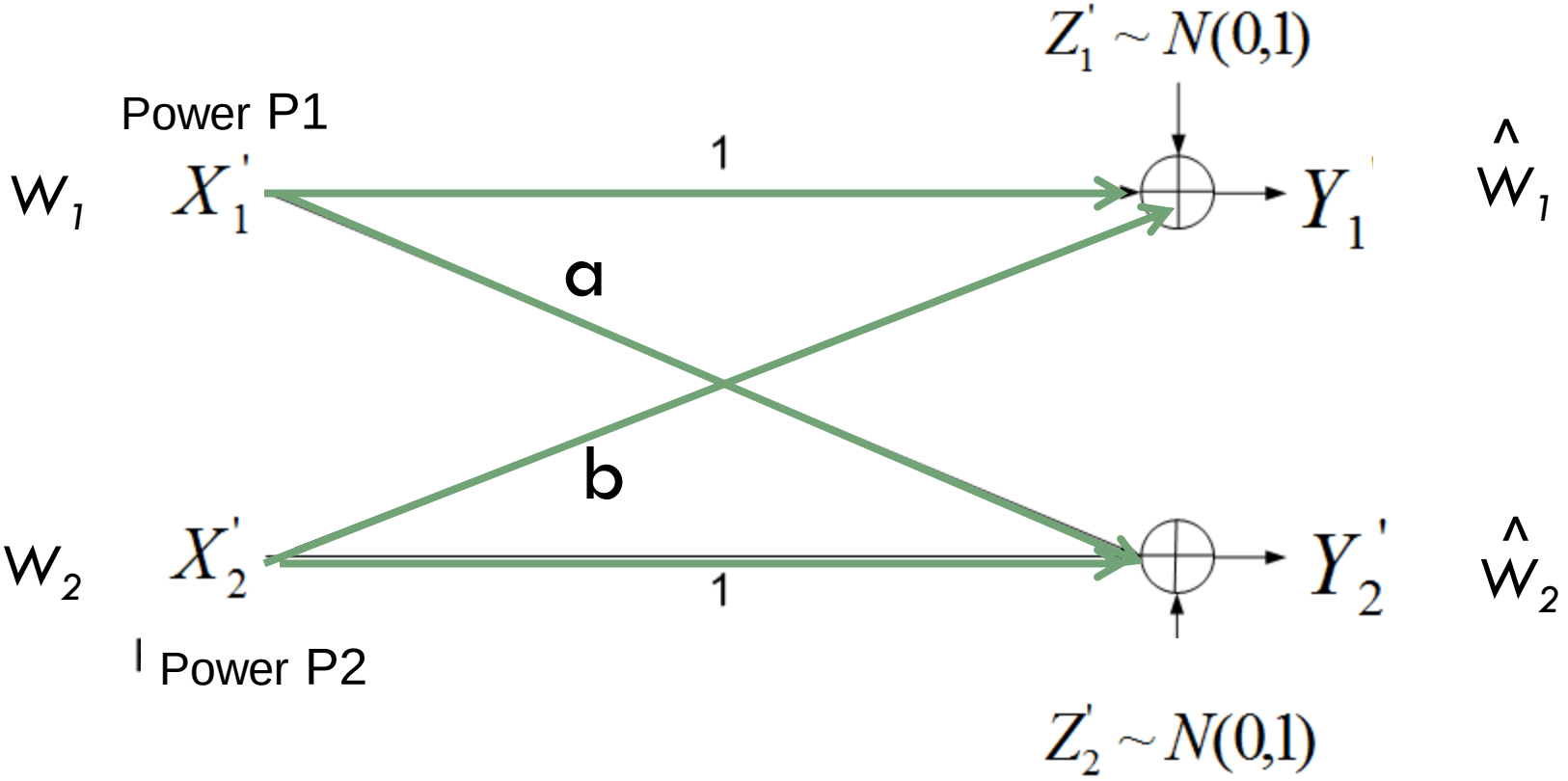
Codificação por Superposição



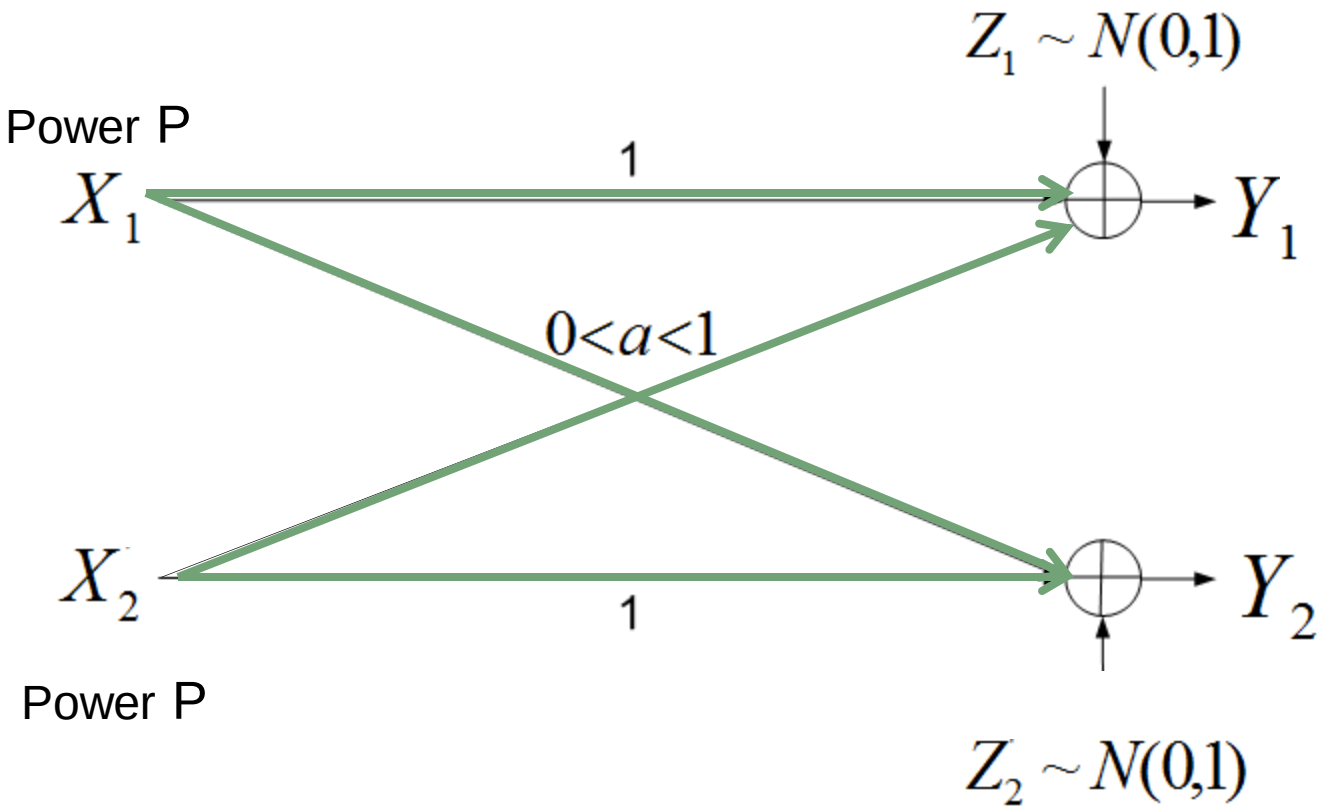




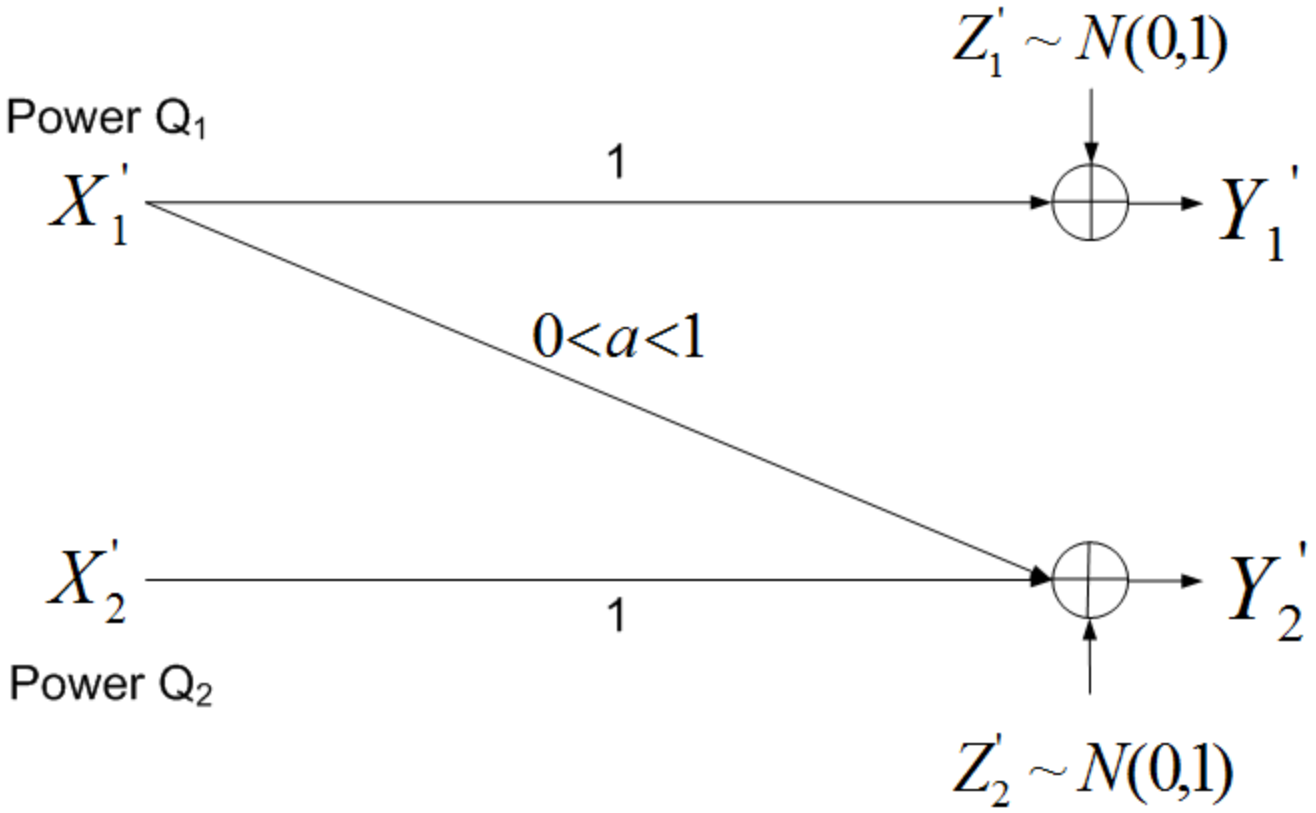
Canal de interferência Gaussiano padrão



Canal de interferência Gaussiano simétrico



Canal de Interferência em Z

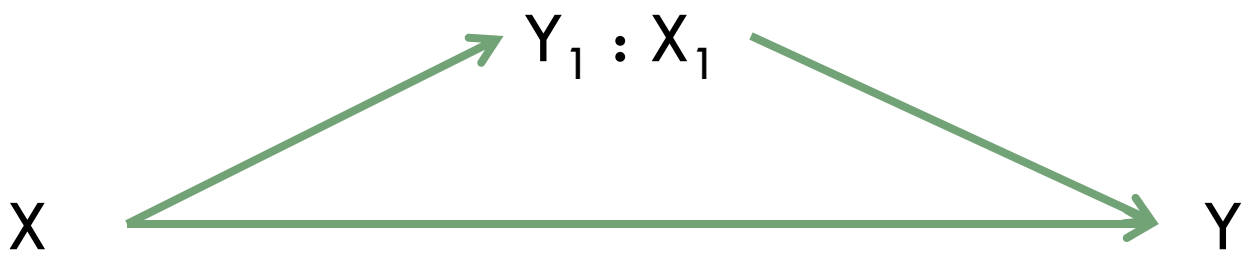


Canal de Interferência: estratégias

O que se pode fazer com interferência:

1. Ignorar (tomar a interferência como ruído,
2. Evitar (dividir o espaço de sinal (TDM/FDM)),
3. Parcialmente decodificar os dois sinais de interferência,
4. Parcialmente decodificar um e totalmente o outro,
5. Decodificar os dois sinais de interferência (a melhor opção para interferência forte, $a \geq 1$).

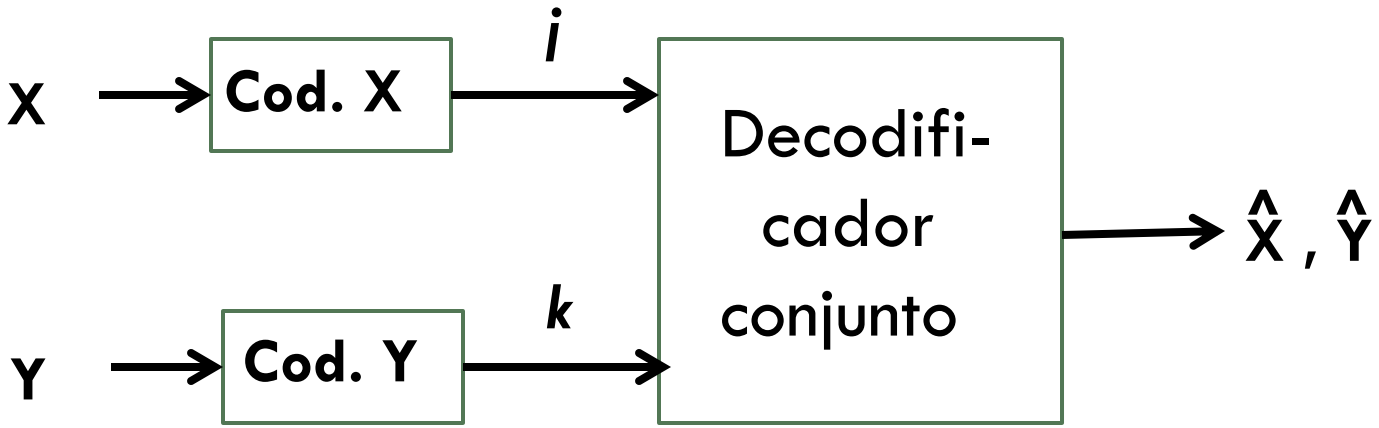
Canal de Relay



- O canal de relay é dito fisicamente degradado se $p(y,y_1 \mid x,x_1)=p(y_1 \mid x,x_1)\ p(y \mid y_1,x_1)$.
- Portanto Y é uma versão degradada do sinal de relay Y_1 .
- Teorema: $C = \sup_{p(x,x_1)} \min \{ I(X,X_1;Y_1), I(X;Y_1 \mid X_1)\}$

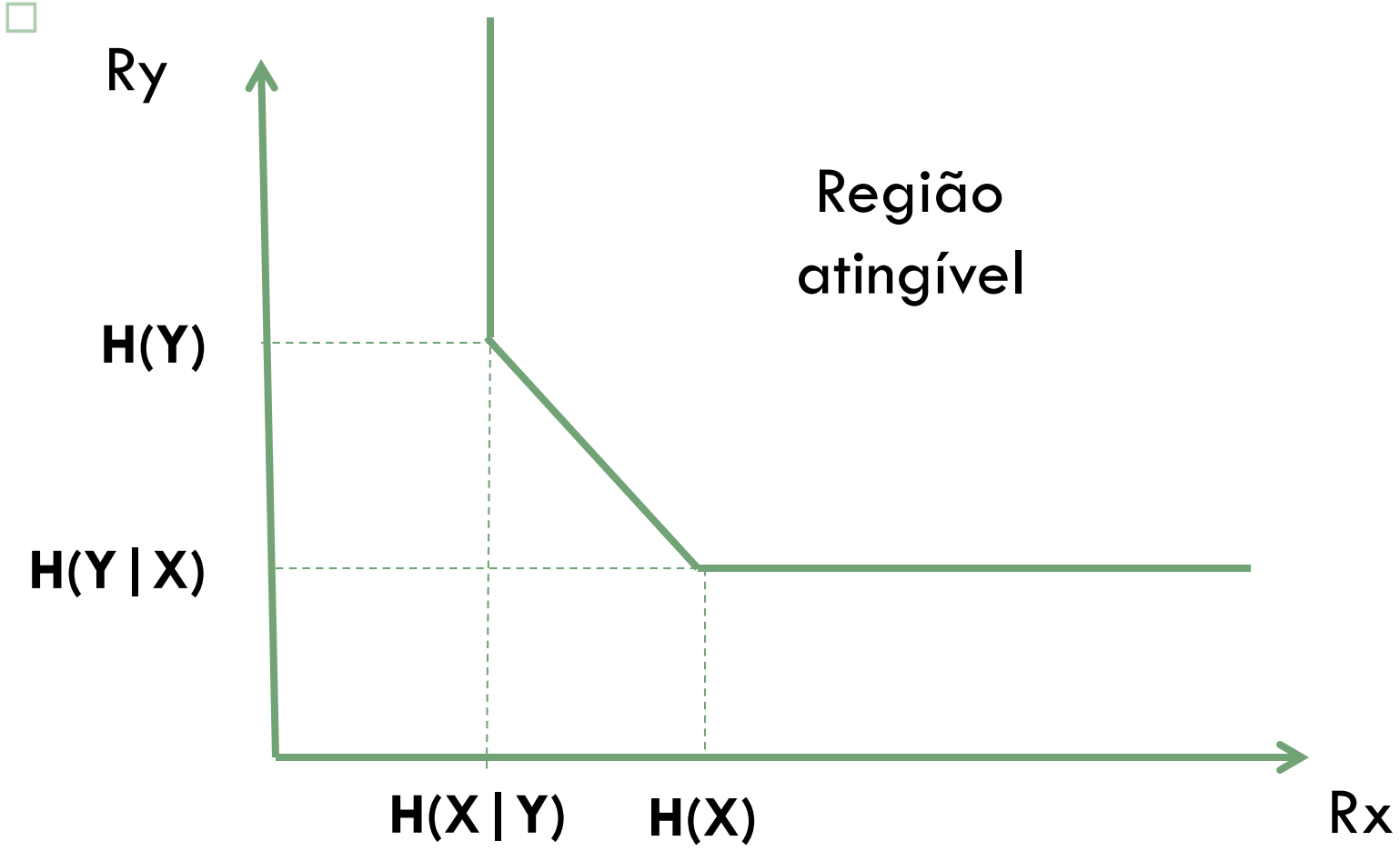
Codificação de Fonte de Slepian Wolf

- X,Y variáveis aleatórias correlatadas ~ p(x,y)



Índices *j* e *k* com taxas R_x and R_y, resp.

Slepian Wolf (continuação)



Fechamento

Muitas frentes de pesquisa:

Codificação conjunta de fonte e canal

Codificação para canais com informação lateral

Codificação distribuída de fonte

Estratégias de codificação para redes

“Casamento” de “Network Coding” e T I de múltiplos usuários.

- Obrigado !

max@fee.unicamp.br