Informação mútua para canais em cascata

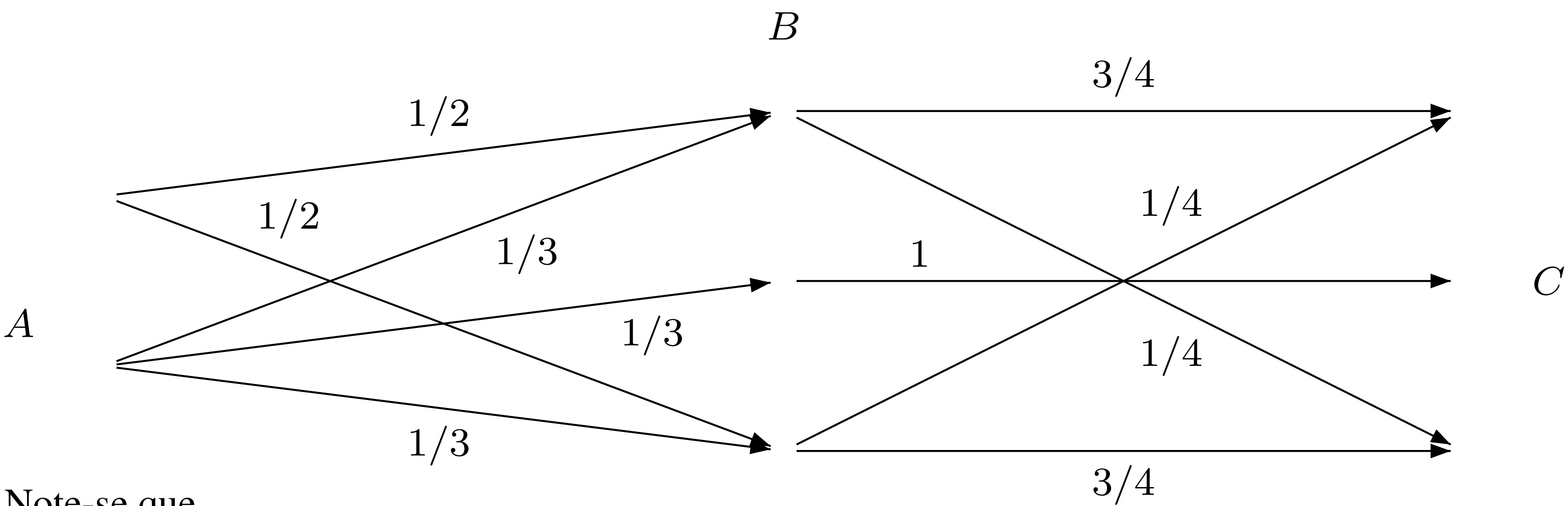
Resultado 2.6 *Para a cascata dos canais AB e BC, tem-se $I(A;B) \geqslant I(A;C)$, verificando-se a igualdade se e só se $P(a|c) = P(a|b)$ sempre que $P(b,c) \neq 0$.*
Por outras palavras, a informação resultante da utilização de uma cascata de canais não pode exceder a informação resultante da utilização do primeiro canal.

Vejamos em seguida que, se o canal BC for isento de ruído, então $I(A;B) = I(A;C)$. Para isso, mostramos que, sendo $P(b,c) \neq 0$, tem-se $P(a|c) = P(a|b)$.
Pelo Teorema de Bayes e pela condição de cascata, temos

$$P(a|c) = \frac{P(a,c)}{P(c)} = \sum_{b' \in B} \frac{P(a,b',c)}{P(b',c)} \frac{P(b',c)}{P(c)} = \sum_{b' \in B} P(a|b',c)P(b'|c) = \sum_{b' \in B} P(a|b')P(b'|c).$$

Ora, sendo o canal BC isento de ruído, e $P(b,c) \neq 0$, temos $P(b',c) = 0 = P(b'|c)$ para todo o $b' \in B \setminus \{b\}$ e $P(b|c) = 1$. Logo $P(a|c) = P(a|b)$, donde, recordando que $I(A;X) = \sum_{a \in A} \sum_{x \in X} P(a|x)P(a) \log \frac{P(a|x)}{P(a)}$, segue que $I(A;B) = I(A;C)$. $\square$

Exemplo



Note-se que

$$\mathbf{P}_{AC} = \mathbf{P}_{AB}\mathbf{P}_{BC} = \begin{bmatrix} 1/2 & 0 & 1/2 \\ 1/3 & 1/3 & 1/3 \end{bmatrix} \begin{bmatrix} 3/4 & 0 & 1/4 \\ 0 & 1 & 0 \\ 1/4 & 0 & 3/4 \end{bmatrix} = \begin{bmatrix} 1/2 & 0 & 1/2 \\ 1/3 & 1/3 & 1/3 \end{bmatrix},$$

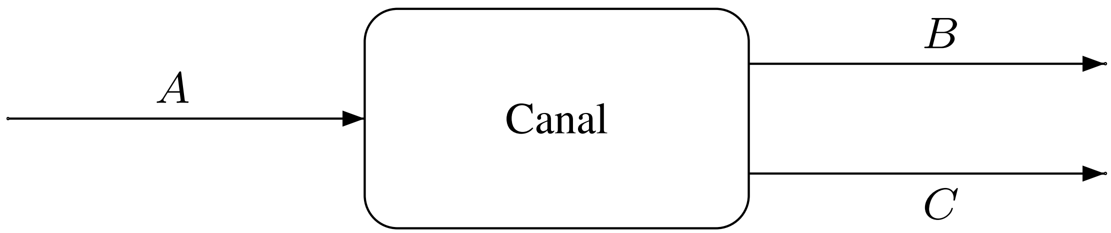
donde resulta que $I(A;B) = I(A;C)$ apesar do canal BC não ser isento de ruído.

Aditividade da informação mútua

- Para compensar a perda de informação resultante do ruído no canal de comunicação, estudaremos várias estratégias para introduzir redundância à entrada que permita recuperar a informação original da informação perturbada obtida à saída.
- Um método simples de introduzir redundância é obtido pela repetição da mensagem original.
- Uma forma de modelar uma tal repetição que tem a vantagem de ter aplicações mais gerais é considerar vários canais através dos quais a mesma informação é transmitida para um mesmo recetor.

Canais com saídas múltiplas

- Por exemplo, para dois canais:



  digamos com $A = \{a_1, \ldots, a_r\}$, $B = \{b_1, \ldots, b_s\}$, $C = \{c_1, \ldots, c_t\}$.
- Supõe-se que, fornecido um símbolo a_i à entrada, o canal produz um símbolo b_j e um símbolo c_k à saída.

- As correspondentes probabilidades serão então:

  $P(a_i)$: probabilidade a priori do símbolo de entrada a_i, desconhecidos os símbolos de saída;

  $P(a_i|b_j)$: probabilidade a posteriori do símbolo de entrada a_i, conhecido o símbolo de saída b_j;

  $P(a_i|b_j, c_k)$: probabilidade a posteriori do símbolo de entrada a_i, conhecidos os símbolos de saída b_j e c_k.

- Definimos a *entropia a posteriori de A dados b_j, c_k*:

$$H(A|b_j, c_k) = \sum_{a \in A} P(a|b_j, c_k) \log \frac{1}{P(a|b_j, c_k)},$$

  e o correspondente *equívoco de A em relação a B e C*:

$$H(A|B, C) = \sum_{b \in B} \sum_{c \in C} P(b, c) H(A|b, c).$$

- Podemos agora definir a *informação mútua entre A e BC* dada pelo canal:

$$I(A;B,C) = H(A) - H(A|B,C)$$

ou seja, a quantidade de informação dada por A menos aquela que se perde na utilização dos canais AB e AC.

- Note-se que

$$\begin{aligned} I(A;B,C) &= H(A) - H(A|B) + H(A|B) - H(A|B,C) \\ &= I(A;B) + I(A;C|B) \end{aligned}$$

sendo $I(A;C|B)$ a informação adicional sobre A fornecida pelo canal AC depois de conhecida a saída do canal AB.

Resultado 2.7 *Para um canal com entrada A e saídas B e C, tem-se*

$$I(A;B,C) \geqslant I(A;B),$$

verificando-se a igualdade se e só se $H(A|B) = H(A|B,C)$.

- Em particular, a utilização dupla de um canal aumenta potencialmente a informação sobre a fonte em relação à sua utilização simples.

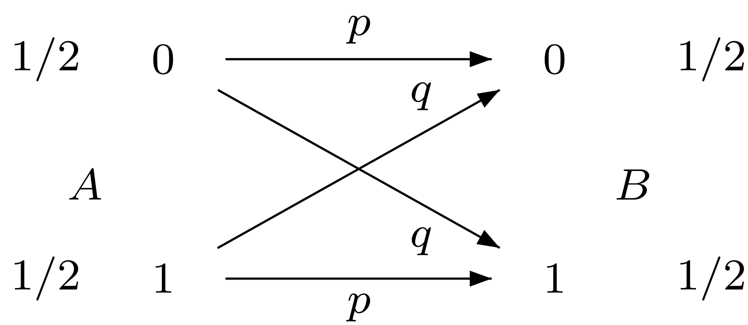
Prova. Basta aplicar o Lema 1.1 em face da seguinte expressão para $I(A;C|B)$:

$$
\begin{aligned}
H(A|B) - H(A|B,C) &= \sum_b P(b) H(A|b) - \sum_{b,c} P(b,c) H(A|b,c) \\
&= \sum_b P(b) \sum_a P(a|b) \log \frac{1}{P(a|b)} - \sum_{b,c} P(b,c) \sum_a P(a|b,c) \log \frac{1}{P(a|b,c)} \\
&= \sum_{a,b} P(a,b) \log \frac{1}{P(a|b)} - \sum_{a,b,c} P(a,b,c) \log \frac{1}{P(a|b,c)} \\
&= \sum_{a,b,c} P(a,b,c) \log \frac{1}{P(a|b)} - \sum_{a,b,c} P(a,b,c) \log \frac{1}{P(a|b,c)} \\
&= \sum_{a,b,c} P(a,b,c) \log \frac{P(a|b,c)}{P(a|b)} \\
&= \sum_{a,b,c} P(a,b,c) \log \frac{P(a,b,c)}{P(a,b)P(c|b)}. \ \square
\end{aligned}
$$

Nota: analogamente, mostra-se que $I(A;B,C) = \sum_{a,b,c} P(a,b,c) \log \dfrac{P(a,b,c)}{P(a)P(b,c)}$.

Exemplo

Consideremos um BSC



no qual cada símbolo a_i é transmitido duas vezes, e representemos por b_j e c_k o resultado da primeira e da segunda transmissão do símbolo, respetivamente, pelo que temos

$$\mathbf{P}_{AB} = \mathbf{P}_{AC} = \begin{bmatrix} p & q \\ q & p \end{bmatrix}.$$

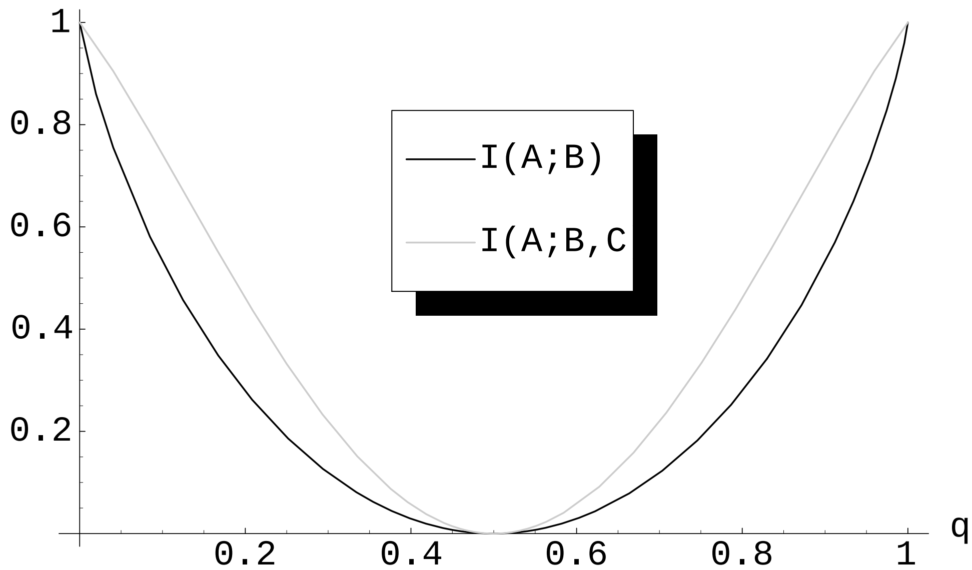
Tínhamos anteriormente calculado $I(A;B) = 1 - (q \log \frac{1}{q} + p \log \frac{1}{p})$.

Para calcular $I(A;B,C) = \sum_{a,b,c} P(a,b,c) \log \dfrac{P(a,b,c)}{P(a)P(b,c)}$, sendo $P(a) = 1/2$, $P(b,c) = \sum_a P(a,b,c)$ e $P(a,b,c) = P(a)P(b|a)P(c|b,a) = \frac{1}{2}P(b|a)P(c|a)$, consideremos a seguinte tabela dos valores relevantes, onde a última coluna representa a contribuição global das parcelas desse tipo para a expressão acima para $I(A;B,C)$:

$a\,bc$	$P(a,b,c)$	$P(b,c)$	Tipo	Contribuição
0 00 1 11	$p^2/2$	$\frac{1}{2}(p^2+q^2)$	X	$2\left(\frac{1}{2}p^2 \log \frac{\frac{1}{2}p^2}{\frac{1}{2}\frac{1}{2}(p^2+q^2)}\right) = p^2 \log \frac{2p^2}{p^2+q^2}$
0 01 1 01 0 10 1 10	$pq/2$	pq	Z	$4\left(\frac{1}{2}pq \log \frac{\frac{1}{2}pq}{\frac{1}{2}pq}\right) = 0$
0 11 1 00	$q^2/2$	$\frac{1}{2}(p^2+q^2)$	Y	$2\left(\frac{1}{2}q^2 \log \frac{\frac{1}{2}q^2}{\frac{1}{2}\frac{1}{2}(p^2+q^2)}\right) = q^2 \log \frac{2q^2}{p^2+q^2}$

donde $I(A;B,C) = p^2 \log \dfrac{2p^2}{p^2+q^2} + q^2 \log \dfrac{2q^2}{p^2+q^2}$.

Comparação de $I(A;B)$ com $I(A;B,C)$:



Casos particulares:

- $q = 0$: sem ruído, $I(A;B) = I(A;B,C) = I(A) = 1$;
- $q = 1$: garantia de inversão do bit, $I(A;B) = I(A;B,C) = I(A) = 1$;
- $q = 1/2$, ambiguidade total do resultado, $I(A;B) = I(A;B,C) = 0$.

Nota sobre notação

Na Aula 2, enunciámos e provámos o seguinte resultado:

Teorema 1.6 $H(P) = H(P_T) + H(P_{S|T}) = H(P_S) + H(P_{T|S})$,

onde $H(PS_{S|T})$ teve de ser intepretado não como a entropia duma distribuição de probabilidades mas como uma "entropia condicional", definida como a média pesada (com pesos dados por P_T) das entropias $H(P_{S|t})$.
Na notação que temos vindo a usar no estudo dos canais de informação, preferimos para $H(P_{S|T})$ a notação $H(S|T)$, onde desaparece a confusão com a entropia duma distribuição de probabilidades.
Com esta notação, não abusiva, o Teorema 1.6 passa a ter o seguinte enunciado, onde uniformizamos também a notação de forma a torná-la mais sugestiva:

Teorema 1.6 $H(S, T) = H(T) + H(S|T) = H(S) + H(T|S)$.

Nesta forma, o Teorema 1.6 apareceu também na página 107, no contexto do estudo de canais de informação. (Ver também o diagrama da página 108.)

Capacidade de um canal

- Recorde-se que a informação mútua dada por um canal foi definida como a informação fornecida à entrada menos o equívoco da informação à entrada condicionada ao conhecimento da saída.

$$I(A; B) = H(A) - H(A|B).$$

- Em particular, a informação mútua só faz sentido quando é dada uma fonte de informação a transmitir pelo canal e depende desta, nomeadamente da distribuição de probabilidades $\{P_A(a)\}_{a \in A}$.
- A *capacidade de um canal AB* é

$$C = \max_{\{P_A(a)\}_{a \in A}} I(A; B).$$

- Note-se que:
  - $C \geqslant 0$ pois todo o $I(A;B) \geqslant 0$;
  - $C \leqslant \min\{\log|A|, \log|B|\}$ pois cada $I(A;B) \leqslant H(A) \leqslant \log|A|$ e idem para B.
- O cálculo de C é um problema de maximização da função contínua $\mathbb{R}^{|A|} \to \mathbb{R}$

$$
\begin{aligned}
I(A;B) &= \sum_{a\in A}\sum_{b\in B} P(a,b)\log\frac{P(a,b)}{P(a)P(b)} \\
&= \sum_{a\in A}\sum_{b\in B} P(a)P(b|a)\log\frac{P(a)P(b|a)}{P(a)\sum_{a\in A}P(a)P(b|a)}
\end{aligned}
$$

onde as probabilidades do canal $P(b|a)$ poderão ser conhecidas ou estimadas estatisticamente e a maximização é feita relativamente ao domínio compacto

$$
\{(P(a_i))_i \in \mathbb{R}^{|A|} : P(a_i) \geqslant 0\ (\forall i),\ \sum_i P(a_i) = 1\}.
$$

Problema de maximização

- A existência de um tal máximo segue de resultados bem conhecidos da Análise/Topologia.
- Mas o seu cálculo está longe de ser um problema trivial em geral. Em alguns casos concretos, os seguintes métodos poderão ser eficazes:
  - método dos multiplicadores de Lagrange;
  - algoritmos baseados na procura na direção do gradiente (que é a direção de variação máxima);
  - algoritmos iterativos desenvolvidos por alguns autores (Arimoto (1972), Blahut (1972)).

Capacidade de um BSC

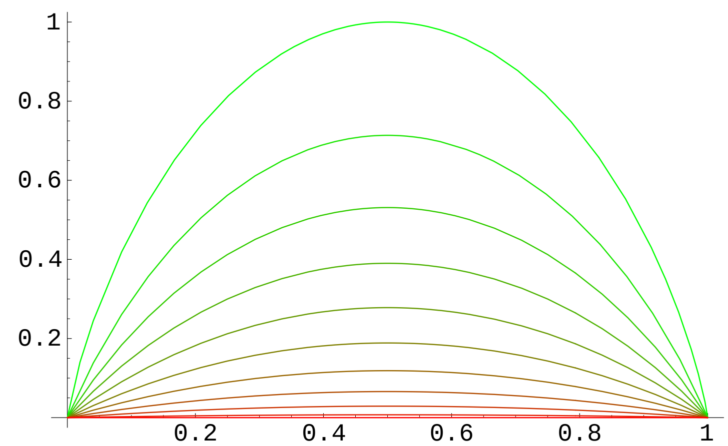
Para um BSC, obtivemos anteriormente a seguinte expressão para a informação mútua

$$I(A;B) = \left((\omega p + \overline{\omega} q) \log \frac{1}{\omega p + \overline{\omega} q} + (\overline{\omega} p + \omega q) \log \frac{1}{\overline{\omega} p + \omega q} \right) - (q \log \frac{1}{q} + p \log \frac{1}{p}),$$

que é aqui considerada função $f_q(\omega)$ de ω, sendo $\overline{\omega} = 1 - \omega$ e sendo p, q parâmetros não negativos com $p + q = 1$.

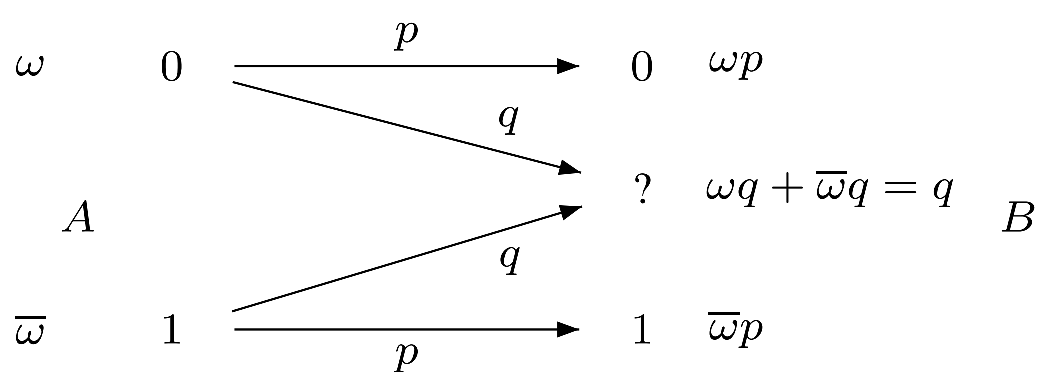
Note-se que:

- $f_q(\omega) = f_q(\overline{\omega})$, sendo $f_q(0) = f_q(1) = 0$;
- $f_{1/2} \equiv 0$;
- gráficos de $f_{n/20}$ para $n = 0, \ldots, 10$ (com cores evoluindo gradualmente de verde a vermelho):



Capacidade de um BEC

Consideremos um BEC dado por



Calculámos anteriormente

$$
\begin{aligned}
I(A;B) &= \omega p \log \frac{1}{\omega p} + \overline{\omega} p \log \frac{1}{\overline{\omega} p} - p \log \frac{1}{p} \\
&= p \left(\omega \log \frac{1}{\omega} + \overline{\omega} \log \frac{1}{\overline{\omega}} \right) = (1-q) H(A)
\end{aligned}
$$

pelo que concluímos que $C_{BEC} = 1 - q$ pois $\max_{\{P_A(a)\}_{a \in A}} H(A) = \log 2 = 1$, sendo o máximo atingido para $\omega = P(a=0) = 1/2$.
Para $q = 0$, o canal BEC atinge a sua capacidade de 1 bit, para a distribuição de probabilidade uniforme à entrada. Para $q = 1$, o canal atinge a capacidade, de 0 bit, independentemente da distribuição de probabilidade de entrada.

Capacidade de canal fracamente simétrico

- Um canal diz-se *simétrico* se as linhas (respetivamente as colunas) da matriz do canal forem obtidas umas das outras por permutações.
- Um canal diz-se *fracamente simétrico* se as linhas da matriz do canal forem obtidas umas das outras por permutações e as somas das colunas $c_j = \sum_{a_i \in A} P(b_j|a_i)$ forem todas iguais.
- Note-se que todo o canal simétrico também é fracamente simétrico.
- Recorde-se que, num canal AB, temos

$$\begin{aligned} I(A;B) &= H(B) - H(B|A) \\ &= H(B) - \sum_{a \in A} P(a) \sum_{b \in B} P(b|a) \log \frac{1}{P(b|a)} \\ &\leqslant \log|B| - \sum_{a \in A} P(a) H(B|a). \end{aligned}$$

Suponhamos agora que temos um canal fracamente simétrico.
Ora $H(B|a_i) = \sum_{b \in B} P(b|a_i) \log \frac{1}{P(b|a_i)}$ depende das entradas na linha i da matriz do canal, pelo que se conclui que $H(B|a_i) = H(B|a)$ é constante, i.e. não varia com a_i.
Logo $I(A;B) \leqslant \log |B| - H(B|a)$ e este é, portanto, um majorante da capacidade do canal.
Considerando a distribuição de probabilidade uniforme à entrada $P(a) = \frac{1}{r}$, onde $r = |A|$, temos

$$P(b_j) = \sum_{a_i \in A} P(b_j|a_i)P(a_i) = \frac{1}{r} \sum_{a_i \in A} P(b_j|a_i) = \frac{c_j}{r},$$

sendo este valor constante, $\frac{c}{r}$, pois as somas das colunas da matriz do canal são constantes.
Logo a distribuição à saída resultante é também uma distribuição uniforme, pelo que $H(B) = \log |B|$.

Teorema 2.8 *Para um canal fracamente simétrico* AB, *a capacidade é dada por*

$$C = \log |B| - H(B|a)$$

onde $H(B|a) = \sum_{b \in B} P(b|a) \log \frac{1}{P(b|a)}$ *pode ser calculada a partir de qualquer linha da matriz do canal. A capacidade do canal é atingida para a distribuição uniforme de probabilidades à entrada,* $P(a) = \frac{1}{|A|}$.$\square$

Exemplo: BSC de novo

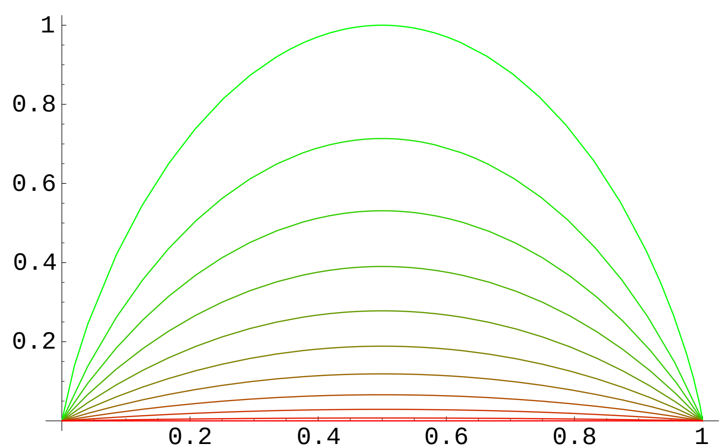
Note-se que, num BSC, tem-se

$$\mathbf{P} = \begin{bmatrix} p & q \\ q & p \end{bmatrix}.$$

Usando o Teorema 2.8 e as igualdades $\log |B| = \log 2 = 1$ e $H(B|a) = p \log \frac{1}{p} + q \log \frac{1}{q}$ concluímos que

$$C = 1 - \left(p \log \frac{1}{p} + q \log \frac{1}{q} \right)$$

capacidade que é atingida quando $P(a) = 1/2$, o que corrobora a expetativa criada com os gráficos



Outro exemplo

Consideremos o canal com matriz

$$\mathbf{P} = \begin{bmatrix} \frac{1}{6} & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{6} & \frac{1}{3} \end{bmatrix}.$$

Temos aqui $\log|B| = \log 3 \simeq 1.58496$ e
$H(B|a) = \frac{1}{6}\log 6 + \frac{1}{2}\log 2 + \frac{1}{3}\log 3 \simeq 1.45915$ pelo que

$$C \in [1.5849 - 1.4592, 1.5850 - 1.4591] = [0.1257, 0.1259],$$

o que mostra que $C \simeq 0.126$, capacidade esta que é atingida tomando $P(a) = 1/2$.