Information Theory and Channel Coding

Prof. Rodrigo C. de Lamare
CETUC, PUC-Rio, Brazil
delamare@cetuc.puc-rio.br

III. Channel capacity

- In this chapter, we study channel capacity and examine several implications of the capacity theorem of Shannon.

- In particular, we examine the fundamental limit of how much information can be transmitted over a channel given some key parameters.

- We present mathematical models of discrete and continuous channels and explore how these models can describe realistic channels.

- We introduce the concept of mutual information and its relation to entropy and the channel capacity of both discrete and continuous channels.

A. Discrete memoryless channels

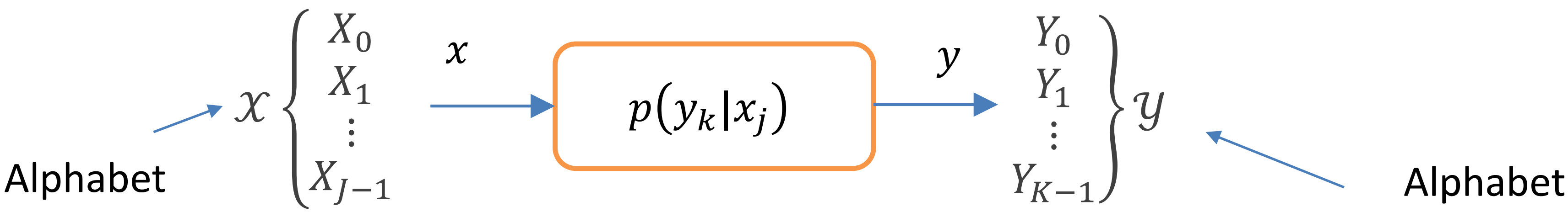
- Communication channels represent the medium over which signals are transmitted.

- In particular, communication channels introduce amplitude and phase distortions in the transmitted signals.

- Modelling communication channels is key because they can be simulated and their capacities can be computed.

- In this section, we will focus our attention on discrete memoryless channels using the concepts of random variables, probability and discrete memoryless sources.

- Let us consider a discrete memoryless channel (DMC) model as



- The model can be written as

$$y = \mathrm{x} + \mathrm{n},$$

where n represents the noise.

- The model is discrete because y and x take on discrete values.

The mathematical description of discrete memoryless channels (DMCs) include:

- The input and output alphabets described by

$$\mathcal{X} = \{X_0, X_1, \dots, X_{J-1}\} \text{ and } \mathcal{Y} = \{Y_0, Y_1, \dots, Y_{K-1}\}$$

- The set of transition probabilities given by

$$p(y_k|x_j) = \mathrm{P}(y_k = Y_k|x_j = X_j), \qquad \text{for all } j \text{ and } k$$

where $0 \leq p(y_k|x_j) \leq 1$ for all j and k.

- The channel can be completely characterized by the set of all transition probabilities as compactly described by

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

- A key property that applies to the set of transition probabilities is

$$\sum_{k=0}^{K-1} p(y_k|x_j) = 1, \qquad \text{for all } j$$

- The input x of the DMC is modelled by the probability

$$p(x_j) = P(x_j = X_j), \qquad j = 0,1,\dots,J-1$$

where $P(x_j = X_j)$ is the probability of an event.

- The joint probability mass function (pmf) of the input x and the ouput y of the DMC is described by

$$\begin{aligned} p(x_j, y_k) &= P(x_j = X_j, y_k = Y_k) \\ &= P(y_k = Y_k | x_j = X_j) P(x_j = X_j) \\ &= p(y_k | x_j) p(x_j) \end{aligned}$$

- The joint pmf is key as it contains the transition and input probabilities.

- The channel output is described by the pmf given by

$$p(y_k) = P(y_k = Y_k)$$
$$= \sum_{j=0}^{J-1} P\big(y_k = Y_k | x_j = X_j\big)\, P\big(x_j = X_j\big)$$
$$= \sum_{j=0}^{J-1} p\big(y_k | x_j\big) p\big(x_j\big),\ \ k = 0,1, \dots , K-1$$

- With the mathematical quantities that constitute the structure of DMCs it possible to fully characterize them.

Example 1

Consider a binary symmetric channel with $J = K = 2$.

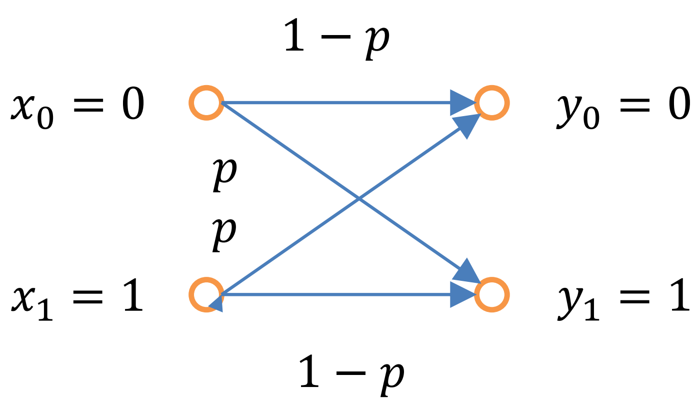
Since the channel is symmetric the probability of receiving a 1 if a 0 was sent is the same as the probability of receiving a 0 if a 1 was sent. This is known as the conditional probability of error and given by p.

a) Describe in a diagram the binary symmetric channel and all its probabilities.

b) Compute the input, transition and output probabilities.

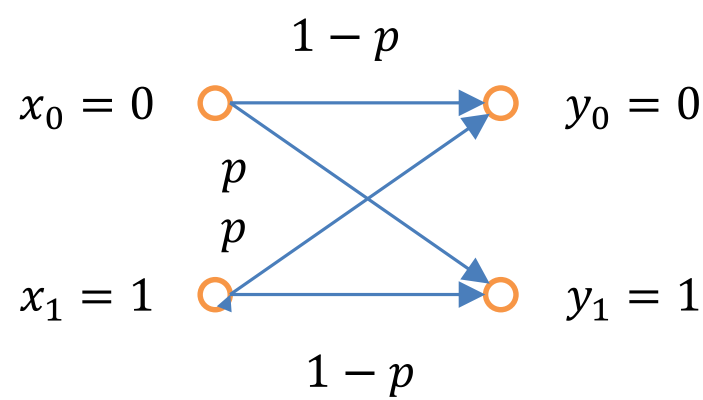
a) The binary symmetric channel (BSC) of this problem deals with $J = 2$ inputs, namely, $x_0 = 0$ and $x_1 = 1$.

There are also $K = 2$ outputs, namely, $y_0 = 0$ and $y_1 = 1$. The BSC can then be illustrated by



b) The input probabilities are described by

$p(x_0) = P(x_0 = 0)$
$p(x_1) = P(x_1 = 1)$



The transition probabilities are given by

$p(y_0|x_0) = 1 - p$
$p(y_1|x_1) = 1 - p$
$p(y_1|x_0) = p$
$p(y_0|x_1) = p$

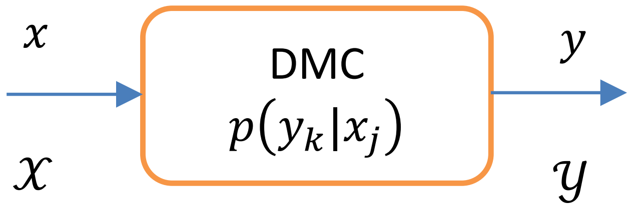
The output probabilities are described by

$$p(y_0) = \sum_{j=0}^{J-1} p(y_0|x_j)p(x_j) = p(y_0|x_0)p(x_0) + p(y_0|x_1)p(x_1) = (1-p)p(x_0) + pp(x_1)$$

$$p(y_1) = \sum_{j=0}^{J-1} p(y_1|x_j)p(x_j) = p(y_1|x_0)p(x_0) + p(y_1|x_1)p(x_1) = pp(x_0) + (1-p)p(x_1)$$

B. Mutual information

- Let us consider a DMC and the entropy associated with the input alphabet $H(\mathcal{X})$, which measures the uncertainty about the input x.



- An important question for DMCs is how to measure $H(\mathcal{X})$ when observing y ?

- We can investigate this by looking into the concept of conditional entropy.

- The conditional entropy for a given output Y_k is described by

$$H(\mathcal{X}|y_k = Y_k) = \sum_{j=0}^{J-1} p(x_j|y_k) \log_2 \left[\frac{1}{p(x_j|y_k)}\right]$$

- If we compute the mean value of $H(\mathcal{X}|y_k = Y_k)$ then we obtain the conditional entropy

$$H(\mathcal{X}|\mathcal{Y}) = \sum_{k=0}^{K-1} H(\mathcal{X}|y_k = Y_k) p(y_k)$$

$$= \sum_{k=0}^{K-1} \sum_{j=0}^{J-1} p(x_j|y_k) p(y_k) \log_2 \left[\frac{1}{p(x_j|y_k)}\right]$$

$$= \sum_{k=0}^{K-1} \sum_{j=0}^{J-1} p(x_j, y_k) \log_2 \left[\frac{1}{p(x_j|y_k)}\right]$$

- The conditional entropy $H(\mathcal{X}|\mathcal{Y})$ measures the uncertainty of the channel after observing the ouput y.

- The mutual information measures the uncertainty about the input x of the DMC while observing the output y of the DMC.

- The mutual information is described by

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

where $H(\mathcal{X})$ measures the uncertainy of the input x and $H(\mathcal{X}|\mathcal{Y})$ measures the uncertainty of the DMC after observing the ouput y of the DMC.

- There is an equivalence of the mutual information if we swap the input and the ouput of the DMC, which yields

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

Properties

i) The mutual information $I(\mathcal{X}, \mathcal{Y})$ is symmetric, i.e.,

$$I(\mathcal{X}, \mathcal{Y}) = I(\mathcal{Y}, \mathcal{X})$$

ii) The mutual information is always nonnegative, i.e.,

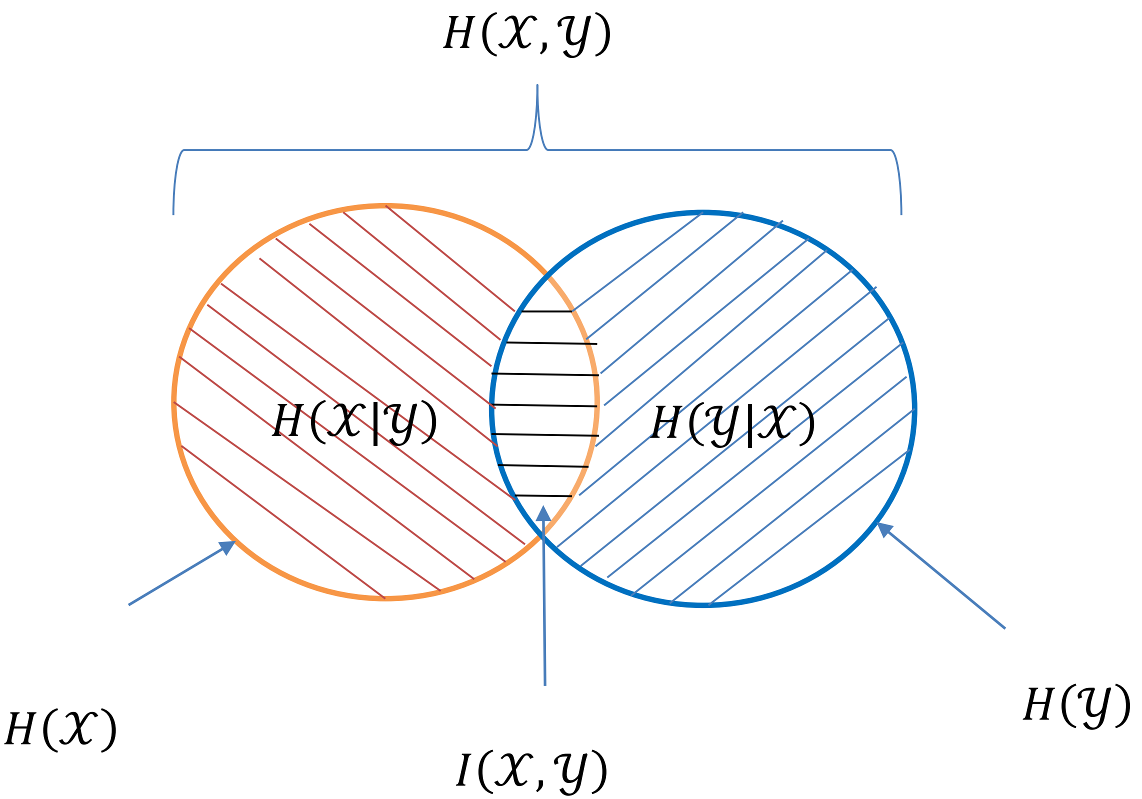
$$I(\mathcal{X}, \mathcal{Y}) \geq 0$$

iii) The mutual information $I(\mathcal{X},\mathcal{Y})$ is related to the joint entropy of the input and the output of the channel by

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

$$H(\mathcal{X},\mathcal{Y}) = \sum_{k=0}^{K-1}\sum_{j=0}^{J-1} p(x_j, y_k)\log_2\left[\frac{1}{p(x_j, y_k)}\right]$$

Illustration



Proof of property i)

We first use the formula for entropy and further manipulate it as follows:

$$
\begin{aligned}
H(\mathcal{X}) &= \sum_{j=0}^{J-1} p(x_j)\log_2\left[\frac{1}{p(x_j)}\right] \\
&= \sum_{j=0}^{J-1} p(x_j)\log_2\left[\frac{1}{p(x_j)}\right]\sum_{k=0}^{K-1} p(y_k|x_j) \\
&= \sum_{j=0}^{J-1}\sum_{k=0}^{K-1} p(y_k|x_j)\,p(x_j)\log_2\left[\frac{1}{p(x_j)}\right] \\
&= \sum_{j=0}^{J-1}\sum_{k=0}^{K-1} p(y_k,x_j)\log_2\left[\frac{1}{p(x_j)}\right]
\end{aligned}
$$

Substituting $H(\mathcal{X})$ and $H(\mathcal{X}|\mathcal{Y})$ into $I(\mathcal{X},\mathcal{Y})$, we obtain

$$
\begin{aligned}
I(\mathcal{X},\mathcal{Y}) &= H(\mathcal{X}) - H(\mathcal{X}|\mathcal{Y}) \\
&= \sum_{j=0}^{J-1}\sum_{k=0}^{K-1} p(y_k,x_j)\log_2\left[\frac{p(x_j|y_k)}{p(x_j)}\right]
\end{aligned}
$$

Using Bayes' rule for conditional probabilities, we have

$$\frac{p(x_j|y_k)}{p(x_j)} = \frac{p(y_k|x_j)}{p(y_k)}$$

Substituting the above relation into $I(\mathcal{X},\mathcal{Y})$, we obtain

$$I(\mathcal{X},\mathcal{Y}) = \sum_{j=0}^{J-1}\sum_{k=0}^{K-1} p(y_k,x_j)\log_2\left[\frac{p(x_j|y_k)}{p(x_j)}\right]$$

$$= \sum_{j=0}^{J-1}\sum_{k=0}^{K-1} p(y_k,x_j)\log_2\left[\frac{p(y_k|x_j)}{p(y_k)}\right]$$

$$= I(\mathcal{Y},\mathcal{X})$$

Proof of property ii)

Since $p(x_j|y_k) = \frac{p(y_k, x_j)}{p(y_k)}$, we may express the mutual information of the channel as

$$I(\mathcal{X}, \mathcal{Y}) = \sum_{j=0}^{J-1} \sum_{k=0}^{K-1} p(y_k, x_j) \log_2 \left[\frac{p(x_j|y_k)}{p(x_j)} \right]$$
$$= \sum_{j=0}^{J-1} \sum_{k=0}^{K-1} p(y_k, x_j) \log_2 \left[\frac{p(y_k, x_j)}{p(y_k) p(x_j)} \right]$$

Using the fundamental inequality arising from Jensen's inequality $\sum_{k=0}^{K-1} p_k \log_2 \left[\frac{q_k}{p_k} \right] \leq 0$, we obtain

$$I(\mathcal{X}, \mathcal{Y}) \geq 0$$

The equality only holds if

$$p(y_k, x_j) = p(y_k)p(x_j)$$

and then we have

$$I(\mathcal{X}, \mathcal{Y}) = 0$$

This property shows that we cannot lose information on average by observing the output of a channel.

Moreover, the mutual information is zero only if the random variables x and y are statistically independent.

Proof of property iii)

Let us first rewrite the expression of the joint entropy $H(\mathcal{X},\mathcal{Y})$ as

$$H(\mathcal{X},\mathcal{Y}) = \sum_{k=0}^{K-1}\sum_{j=0}^{J-1} p(x_j,y_k)\log_2\left[\frac{1}{p(x_j,y_k)}\right]$$
$$= \sum_{k=0}^{K-1}\sum_{j=0}^{J-1} p(x_j,y_k)\log_2\left[\frac{p(x_j)p(y_k)}{p(x_j,y_k)}\right]$$
$$+ \sum_{k=0}^{K-1}\sum_{j=0}^{J-1} p(x_j,y_k)\log_2\left[\frac{1}{p(x_j)p(y_k)}\right]$$

The first double summation on the right-hand side of the above expression is the negative of the mutual information, i.e., -$I(\mathcal{X},\mathcal{Y})$.

The second term can be manipulated as follows:

$$\sum_{k=0}^{K-1}\sum_{j=0}^{J-1} p(x_j, y_k) \log_2\left[\frac{1}{p(x_j)p(y_k)}\right] =$$

$$= \sum_{j=0}^{J-1} \log_2\left[\frac{1}{p(x_j)}\right]\sum_{k=0}^{K-1} p(x_j, y_k) + \sum_{k=0}^{K-1} \log_2\left[\frac{1}{p(y_k)}\right]\sum_{j=0}^{J-1} p(x_j, y_k)$$

$$= \sum_{j=0}^{J-1} p(x_j) \log_2\left[\frac{1}{p(x_j)}\right] + \sum_{k=0}^{K-1} p(y_k) \log_2\left[\frac{1}{p(y_k)}\right]$$

$$= H(\mathcal{X}) + H(\mathcal{Y})$$

Accordingly, we have

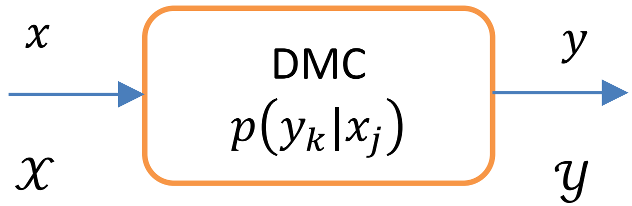
$$H(\mathcal{X},\mathcal{Y}) = -I(\mathcal{X},\mathcal{Y}) + H(\mathcal{X}) + H(\mathcal{Y})$$

and

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

C. Capacity of discrete memoryless channels

- Let us consider a DMC and the entropy associated with the input alphabet $H(\mathcal{X})$, which measures the uncertainty about the input x.



- The mutual information of the input x and the output y of the channel is given by

$$I(\mathcal{X},\mathcal{Y}) = \sum_{j=0}^{J-1}\sum_{k=0}^{K-1} p(y_k, x_j) \log_2\left[\frac{p(x_j|y_k)}{p(x_j)}\right]$$

$$= \sum_{j=0}^{J-1}\sum_{k=0}^{K-1} p(y_k, x_j) \log_2\left[\frac{p(y_k|x_j)}{p(y_k)}\right]$$

- The joint pmf between the input and output variables is given by

$$p(y_k, x_j) = p(y_k | x_j) p(x_j)$$

- The output probabilities can be computed by

$$p(y_k) = \sum_{j=0}^{J-1} p(y_k | x_j) p(x_j), \qquad k = 0,1, \dots, K-1$$

- In order to compute $I(\mathcal{X}, \mathcal{Y})$, we need the input probabilities

$$p(x_j), \qquad j = 0,1, \dots, J-1$$

- The capacity of a DMC can be computed by maximizing the mutual information $I(\mathcal{X},\mathcal{Y})$ subject to appropriate constraints on $p(x_j)$.

- The computation of the capacity can be formulated as the optimization:

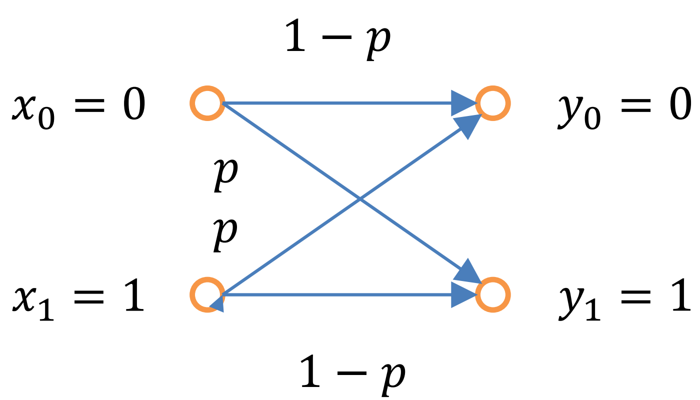
$$C = \max_{p(x_j)} I(\mathcal{X},\mathcal{Y}) \text{ bits/channel use or bits / transmission}$$

$$\text{subject to } p(x_j), \quad \text{for all } j \text{ and } \sum_{j=0}^{J-1} p(x_j) = 1$$

- The optimization involves the maximization of $I(\mathcal{X},\mathcal{Y})$ by adjusting the variables $p(x_1), p(x_2), \ldots, p(x_{J-1})$ subject to appropriate constraints.

Example 2

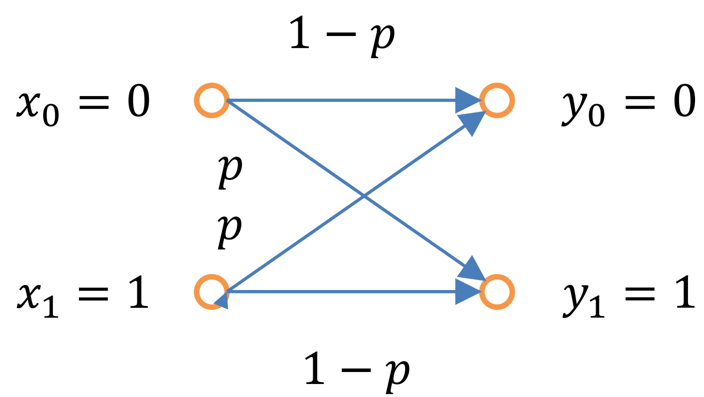
Consider the BSC illustrated by



a) Compute the capacity of the channel

b) Show how the capacity varies with p using a plot.

We consider the BSC.



We know that the entropy $H(\mathcal{X})$ is maximized when $p(x_0) = p(x_1) = \frac{1}{2}$, where x_0 and x_1 are 0 and 1, respectively.

The mutual information $I(\mathcal{X}, \mathcal{Y})$ is similarly maximized as described by

$$C = I(\mathcal{X}, \mathcal{Y}) \text{ when } p(x_0) = p(x_1) = \frac{1}{2},$$

where

$p(y_0|x_0) = 1 - p = p(y_1|x_1)$

$p(y_1|x_0) = p = p(y_0|x_1)$

a) By substituting the transition probabilities in $I(\mathcal{X},\mathcal{Y})$, we obtain

$$I(\mathcal{X},\mathcal{Y}) = \sum_{j=0}^{J-1}\sum_{k=0}^{K-1} p(y_k, x_j)\log_2\left[\frac{p(y_k|x_j)}{p(y_k)}\right]$$

With $J = K = 2$ and then setting $p(x_0) = p(x_1) = \frac{1}{2}$, we have

$$C = \max_{p(x_j)} \sum_{j=0}^{1}\sum_{k=0}^{1} p(y_k, x_j)\log_2\left[\frac{p(y_k|x_j)}{p(y_k)}\right]$$

$$= p(y_0, x_0)\log_2\left[\frac{p(y_0|x_0)}{p(y_0)}\right] + p(y_0, x_1)\log_2\left[\frac{p(y_0|x_1)}{p(y_0)}\right]$$

$$+p(y_1, x_0)\log_2\left[\frac{p(y_1|x_0)}{p(y_1)}\right] + p(y_1, x_1)\log_2\left[\frac{p(y_1|x_1)}{p(y_1)}\right]$$

$$= p(y_0|x_0)\,p(x_0)\log_2\left[\frac{p(y_0|x_0)}{p(y_0)}\right] + p(y_0|x_1)p(x_1)\log_2\left[\frac{p(y_0|x_1)}{p(y_0)}\right]$$

$$+p(y_1|x_0)p(x_0)\log_2\left[\frac{p(y_1|x_0)}{p(y_1)}\right] + p(y_1|x_1)p(x_1)\log_2\left[\frac{p(y_1|x_1)}{p(y_1)}\right]$$

$$= \frac{1-p}{2}\log_2[2(1-p)] + \frac{p}{2}\log_2[2p] + \frac{p}{2}\log_2[2p] + \frac{1-p}{2}\log_2[2(1-p)]$$

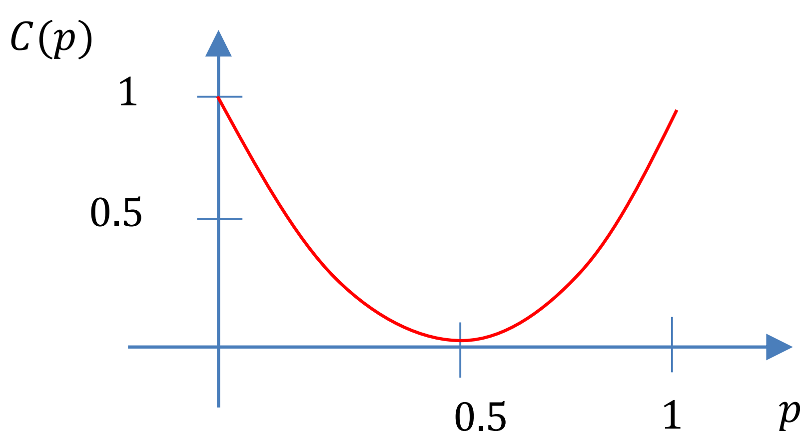
$$= 1 + \mathrm{p}\log_2 p + (1-p)\log_2(1-p)$$

b) Using the definition of entropy and their mathematical relations we have the capacity of the BSC

$$C(p) = 1 - H(p),$$

where $\mathrm{H(p)} = -\mathrm{p}\log_2 p - (1-p)\log_2(1-p)$.

The channel capacity varies with p in a convex manner as shown below.



When $p = 0$, C attains its maximum value of 1 bit/ channel use

When $p = \frac{1}{2}$, C attains its minimum value of 0 bit/ channel use (useless channel)

D. Differential entropy and mutual information for continuous variables

- In this section, we extend the previous concepts to continuous sources and channels, which are modelled as continuous random variables.

- Consider a random variable x with the probability density function $p_x(X)$, the differential entropy of x is described by

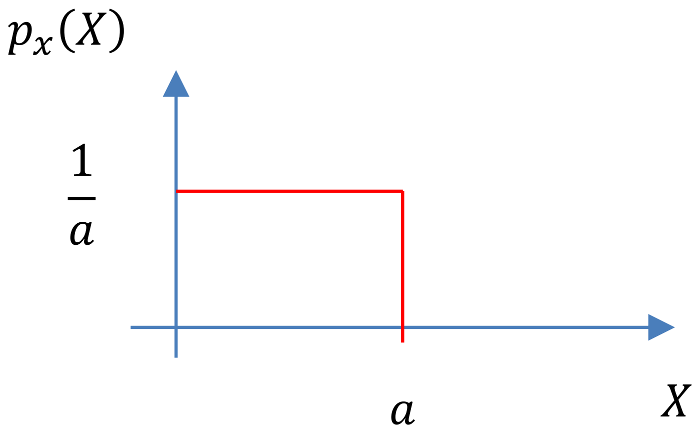
$$h(x) = \int_{-\infty}^{\infty} p_x(X) \log_2 \left[\frac{1}{p_x(X)} \right] dX$$

- As in the discrete case, the differential entropy depends only on the probability density of the random variable x.

Example 3

Compute the differential entropy of a random variable with uniform distribution described by

$$p_x(X) = \begin{cases} \frac{1}{a}, & 0 < X < a \\ 0, & \text{otherwise} \end{cases}$$



Solution:

$$h(x) = \int_{-\infty}^{\infty} p_x(X) \log_2 \left[\frac{1}{p_x(X)}\right] dX$$

$$= \int_0^a \frac{1}{a} \log_2 a \, dX$$

$$= \log_2 a \text{ bits}$$

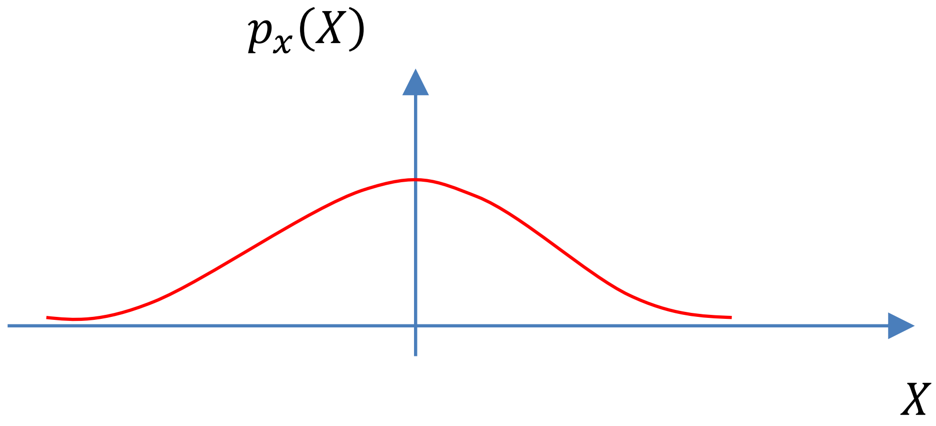
Note that $\log_2 a < 0$ for $a < 1$.

The entropy of a continuous random variable can be negative unlike the case for a discrete random variable.

Example 4

Compute the differential entropy of a random variable with Gaussian distribution described by

$$p_x(X) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{X^2}{2\sigma^2}}$$



Solution:

$$\mathrm{h}(x) = \int_{-\infty}^{\infty} p_x(X) \ln \left[\frac{1}{p_x(X)}\right] dX \quad \text{(nats)}$$

$$= -\int_{-\infty}^{\infty} p_x(X) \ln p_x(X)\, dX$$

$$= -\int_{-\infty}^{\infty} p_x(X) \left[-\frac{X^2}{2\sigma^2} - \ln\sqrt{2\pi\sigma^2}\right] dX$$

$$= \frac{1}{2}\ln 2\pi\sigma^2 + \frac{1}{2}\frac{E[x^2]}{\sigma^2}$$

$$= \frac{1}{2}\ln 2\pi\sigma^2 + \frac{1}{2}\ln \mathrm{e}$$

$$= \frac{1}{2}\ln 2\pi e\sigma^2 \text{ nats}$$

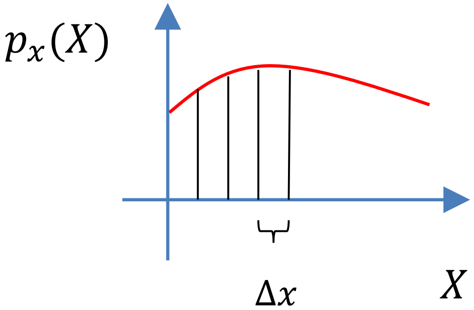
Changing the basis from $\ln$ to $\log_2$, we have

$$\mathrm{h}(x) = \frac{1}{2}\log_2 2\pi e\sigma^2 \quad \text{bits}$$

Relation of differential entropy to entropy of discrete variables

- Let us consider the random variable x as the limiting form of a discrete random variable $x_k = k\Delta x, k = 0, \pm 1, \pm 2, \ldots$, where $\Delta x \to 0$.

- In this case, x takes on a value in the range $[x_k, x_k + \Delta x\ ]$ with probability given by



$$p_x(X_k)\Delta x = \int_{k\Delta x}^{(k+1)\Delta x} p_x(X)dX$$

- Consider the quantized random variable x_q described by

$$x_q = x_k, \qquad k\Delta x \leq X_q < (k+1)\Delta x$$

- Then the probability that $x_q = X_k$ is given by

$$P(x_q = X_k) = p_x(X_k)\Delta x = \int_{k\Delta x}^{(k+1)\Delta x} p_x(X)dX$$

- Let us now compute the entropy of x_k by letting $\Delta x \to 0$ as follows:

$$\begin{aligned}
H(x_k) &= \lim_{\Delta x \to 0} \sum_{k=-\infty}^{\infty} p_x(X_k)\,\Delta x \log_2\left(\frac{1}{p_x(X_k)\Delta x}\right) \\
&= \lim_{\Delta x \to 0}\left[\sum_{k=-\infty}^{\infty} p_x(X_k)\,\Delta x \log_2\left(\frac{1}{p_x(X_k)}\right) - \log_2 \Delta x \sum_{k=-\infty}^{\infty} p_x(X_k)\Delta x\right] \\
&= \int_{-\infty}^{\infty} p_x(X) \log_2\left(\frac{1}{p_x(X)}\right) dX - \lim_{\Delta x \to 0} \log_2 \Delta x \int_{-\infty}^{\infty} p_x(X)\mathrm{d}X \\
&= h(x) - \lim_{\Delta x \to 0} \log_2 \Delta x
\end{aligned}$$

Theorem 1:

The previous development leads to

$$H(x_k) = h(x) - \lim_{\Delta x \to 0} \log_2 \Delta x$$

or

$$h(x) = H(x_k) + \lim_{\Delta x \to 0} \log_2 \Delta x,$$

which for $\Delta x \to 0$ results in

$$h(x) = H(x_k)$$

and for an arbitrary Δx related to n quantization bits yields

$$h(x) = H(x_k) + \log_2 \Delta x = H(x_k) + n$$

Example 5

Compute the entropy for the following cases:

a) If a random variable x *has* uniform distribution on $[0, 1]$ and we let $\Delta x = 2^{-n}$.

b) If a random variable x *has Gaussian* distribution with zero mean, $\sigma^2 = 100$.

Solution:

a) For a random variable x with uniform distribution on $[0, 1]$ and $\Delta x = 2^{-n}$, we have

$$H(x_k) = \sum_{k=-\infty}^{\infty} p_x(X_k)\,\Delta x\,\log_2\left(\frac{1}{p_x(X_k)\Delta x}\right) = n$$

and

$$h(x) = H(x_k) + \log_2 \Delta x = n - n = 0,$$

which means that n bits suffice to describe x to an accuracy of n bits.

b)

For a random variable x *with* Gaussian distribution with zero mean and $\sigma^2 = 100$, we have

$$h(x) = H(x_k) + \log_2 \Delta x = H(x_k) + n$$
$$= \frac{1}{2} \log_2 2\pi e \sigma^2 + \mathrm{n} = 5.37\mathrm{bits} + \mathrm{n}$$

Joint and conditional entropy: extension to vectors

- We can extend the definition of differential entropy to random vectors.

- The joint differential entropy for a random vector $\boldsymbol{x} = [x_1 \quad \cdots \quad x_n]^T$ is defined by

$$h(\boldsymbol{x}) = \int_{-\infty}^{\infty} p_{\boldsymbol{x}}(\boldsymbol{X}) \log_2 \left[\frac{1}{p_{\boldsymbol{x}}(\boldsymbol{X})} \right] d\boldsymbol{X}$$

- The conditional differential entropy of two variables x and y is described by

$$h(x|y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p_{x,y}(X,Y) \log_2 \left[\frac{1}{p_x(X|Y)} \right] dXdY$$

- Since in general $p_x(X|Y) = p_{x,y}(X,Y)/p_y(Y)$, we can write

$$h(x|y) = h(x,y) - h(y)$$

Example 6

Compute the differential entropy of the random vector $\boldsymbol{x} = [x_1 \quad \cdots \quad x_n]^T$ whose joint probability density function is

$$p_{\boldsymbol{x}}(\boldsymbol{X}) = \frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{\det(\boldsymbol{K})}} e^{-\frac{1}{2}(\boldsymbol{X}-\boldsymbol{m}_x)^T \boldsymbol{K}^{-1}(\boldsymbol{X}-\boldsymbol{m}_x)}$$

Solution:

$$h(\boldsymbol{x}) = \int_{-\infty}^{\infty} p_{\boldsymbol{x}}(\boldsymbol{X})\, ln\left[\frac{1}{p_{\boldsymbol{x}}(\boldsymbol{X})}\right] d\boldsymbol{X} \quad \text{(nats)}$$

$$= -\int_{-\infty}^{\infty} p_{\boldsymbol{x}}(\boldsymbol{X}) \left(-\frac{1}{2}(\boldsymbol{X}-\boldsymbol{m}_x)^T \boldsymbol{K}^{-1}(\boldsymbol{X}-\boldsymbol{m}_x) - \ln(2\pi)^{\frac{n}{2}} \det(\boldsymbol{K})^{\frac{1}{2}}\right) d\boldsymbol{X}$$

$$= \frac{1}{2}\mathrm{E}[(\boldsymbol{x}-\boldsymbol{m}_x)^T \boldsymbol{K}^{-1}(\boldsymbol{x}-\boldsymbol{m}_x)] + \frac{1}{2}\ln(2\pi)^n \det(\boldsymbol{K})$$

$$= \frac{1}{2} tr[\boldsymbol{K}\boldsymbol{K}^{-1}] + \frac{1}{2}\ln(2\pi)^n \det(\boldsymbol{K})$$

$$= \frac{1}{2} n \ln e \; + \frac{1}{2}\ln(2\pi)^n \det(\boldsymbol{K})$$

$$= \frac{1}{2} \ln e^n \; + \frac{1}{2}\ln(2\pi)^n \det(\boldsymbol{K})$$

$$= \frac{1}{2}\ln(2\pi e)^n \det(\boldsymbol{K})$$

By changing the basis of the logarithm, we have

$$h(\boldsymbol{x}) = \frac{1}{2}\log_2(2\pi e)^n \det(\boldsymbol{K}) \quad \text{bits}$$

E. Mutual information

- Consider a pair of random variables x and y that can represent the input and the output of a communication channel.

- The mutual information between x and y is defined by

$$I(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p_{x,y}(X,Y)\log_2\left[\frac{p_x(X|Y)}{p_x(X)}\right] dXdY,$$

where $p_{x,y}(X,Y)$ is the joint pdf of x and y , and $p_x(X|Y)$ is the conditional pdf of x subject to $y = Y$.

- The conditional differential entropy of two variables x and y is described by

$$h(x|y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p_{x,y}(X,Y) \log_2 \left[\frac{1}{p_x(X|Y)} \right] dX dY$$

- Since in general $p_x(X|Y) = p_{x,y}(X,Y)/p_y(Y)$, we can write

$$h(x|y) = h(x,y) - h(y)$$

- These relations are useful to compute the mutual information in practical situations.

Properties of mutual information

i) $I(x, y) = I(y, x)$ (symmetry)

ii) $I(x, y) \geq 0$ (non negativity)

iii) $I(x, y) = h(x) - h(x|y)$
$= h(y) - h(y|x)$

- The proofs are similar to those of mutual information with discrete variables.

Example 7

Compute the mutual information between the input x and the output y of the channel

when both x and y are drawn from Gaussian random variables with zero mean and variance σ^2 and the covariance matrix of $\boldsymbol{u} = [x\ y]^{\boldsymbol{T}}$

$$\boldsymbol{K} = E[(\boldsymbol{u} - \boldsymbol{m}_u)(\boldsymbol{u} - \boldsymbol{m}_u)^T] = \begin{bmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{bmatrix},$$

where $\boldsymbol{m}_u$ is the mean vector of $\boldsymbol{u}$.

Solution:

The differential entropies of the input x and the output y of the channel are

$$h(x) = \frac{1}{2}\log_2(2\pi e)\sigma^2 = h(y)$$

The joint differential entropy is given by

$$\begin{aligned} h(x,y) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p_{x,y}(X,Y)\log_2\left[\frac{1}{p_x(X|Y)}\right] dX dY \\ &= \frac{1}{2}\log_2(2\pi e)^2 \det(\boldsymbol{K}) \\ &= \frac{1}{2}\log_2(2\pi e)^2 \sigma^4(1-\rho^2) \end{aligned}$$

Therefore, the mutual information is described by

$$
\begin{aligned}
I(x,y) &= h(x) - h(x|y) \\
&= h(x) + h(y) - h(x,y) \\
&= \frac{1}{2}\log_2(2\pi e)\sigma^2 + \frac{1}{2}\log_2(2\pi e)\sigma^2 - \frac{1}{2}\log_2(2\pi e)^2\sigma^4(1-\rho^2) \\
&= \frac{1}{2}\log_2(1-\rho^2),
\end{aligned}
$$

where $h(x|y) = h(x,y) - h(y)$

F. Capacity of Gaussian channels

- The information capacity of Gaussian channels is the maximum of the mutual information between the input and the output of the channel.

- To this end, we need to consider all distributions on the input that satisfy a power constraint P.

- Mathematically, the information capacity of Gaussian channels with power constraint P is given by

$$C = \max_{p_x(X)} I(x, y)$$

$$\text{subject to } E[x^2] \leq P$$

Channel capacity theorem (Shannon, 1948)

The information capacity of a continuous channel bandlimited to B Hz perturbed by additive white Gaussian noise (AWGN) with power spectral density $\frac{N_0}{2}$ is given by

$$C = B\log_2\left(1+\frac{P}{N_0 B}\right), \quad \text{bits/ s}$$

where P is average transmit power.

This theorem shows that given P and B we can transmit information at a rate of C bits per second.

Computation of the information capacity

- In order to solve the optimization problem given by

$$C = \max_{p_x(X)} I(x, y)$$

$$\text{subject to } E[x^2] \leq P$$

- We first consider the channel model described by

$$y = x + n,$$

where n is AWGN with zero mean and variance σ^2.

- We then work out the mutual information expression as follows:

$$I(x, y) = h(y) - h(y|x)$$

- The mutual information expression can be simplified as

$$
\begin{aligned}
I(x, y) &= h(y) - h(y|x) \\
&= h(y) - h(x + n|x) \\
&= h(y) - h(n|x) \\
&= h(y) - h(n),
\end{aligned}
$$

which takes into account that x and n are statistically independent.

- Next, we need to compute the differential entropies $h(y)$ and $h(n)$.

- The differential entropy of the AWGN noise is given by

$$h(n) = \frac{1}{2}\log_2(2\pi e\sigma^2)$$

- Now, we need to compute the variance of y, which is given by

$$\begin{aligned}\sigma_y^2 &= E[y^2] \\ &= E[(x+n)^2] = E[x^2] + E[n^2] = P + \sigma^2\end{aligned}$$

- The differential entropy of y is expressed by

$$\begin{aligned}h(y) &= \frac{1}{2}\log_2\left(2\pi e\sigma_y^2\right) \\ &= \frac{1}{2}\log_2\left(2\pi e(P+\sigma^2)\right)\end{aligned}$$

- The capacity is the maximum of the mutual information subject to the power constraint, which is taken into account in $h(y)$, and yields

$$
\begin{aligned}
C_t &= \max I(x,y) = h(y) - h(n) \\
&= \frac{1}{2}\log_2\left(2\pi e(P+\sigma^2)\right) - \frac{1}{2}\log_2(2\pi e\sigma^2) \\
&= \frac{1}{2}\log_2\left(\frac{P+\sigma^2}{\sigma^2}\right) \\
&= \frac{1}{2}\log_2\left(1+\frac{P}{\sigma^2}\right) \text{ bits / transmission}
\end{aligned}
$$

- We note that the maximization of $h(y)$ requires that y be Gaussian as Gaussian random variables have the largest differential entropy.

- The capacity can also be expressed per unit of time by considering that K samples have been transmitted over T seconds, which results in

$$\begin{aligned} \mathrm{C} &= \frac{\mathrm{K}}{\mathrm{T}} \mathrm{C_t} = \frac{\mathrm{K}}{\mathrm{T}} \frac{1}{2} \log_2 \left(1 + \frac{P}{\sigma^2}\right) \\ &= \frac{2\mathrm{BT}}{\mathrm{T}} \frac{1}{2} \log_2 \left(1 + \frac{P}{\sigma^2}\right) \\ &= B \log_2 \left(1 + \frac{P}{N_0 B}\right) \text{ bits / second} \end{aligned}$$

- In the above expression, which has been derived by Shannon, we make use of $\mathrm{K} = 2\mathrm{BT}$ samples, where B is the bandwidth.

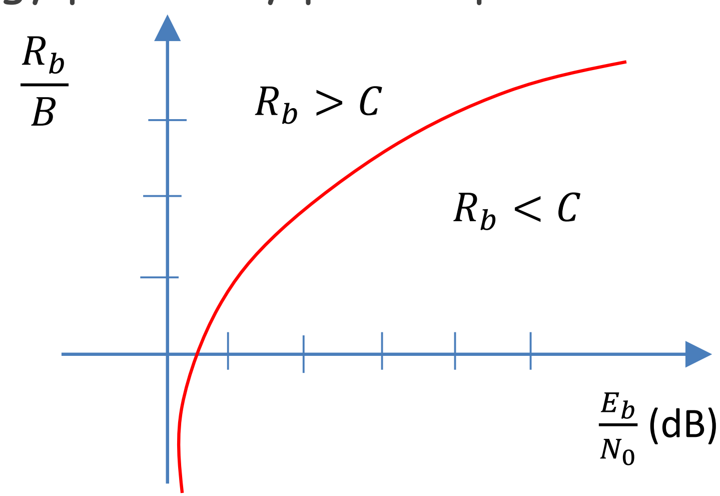
G. Implications of the channel capacity theorem

- In an ideal system, we transmit at a rate equal to $R_b = C$ bits /s.
- If we take into account $P = E_b C$, where E_b is the transmit energy per bit, we have

$$\frac{C}{B} = \log_2\left(1 + \frac{P}{N_0 B}\right) = \log_2\left(1 + \frac{E_b C}{N_0 B}\right)$$

- The spectral efficiency is the ratio of energy per bit by power spectral density is given by

$$\frac{E_b}{N_0} = \frac{2^{\frac{C}{B}} - 1}{\frac{C}{B}}$$



i) When $B \to \infty$ $\frac{E_b}{N_0}$ approaches

$$\left(\frac{E_b}{N_0}\right)_{\infty} = \lim_{B \to \infty}\left(\frac{E_b}{N_0}\right)$$

$$= \frac{1}{\log_2 e} = -0.693 \text{ or } -1.6 \text{ dB}$$

Shannon limit

The capacity limit is then given by

$$C_{\infty} = \lim_{B \to \infty} \mathrm{C} = \frac{\mathrm{P}}{\mathrm{N_0}} \log_2 \mathrm{e}$$

Proof

Since $\log_2(1+x) = x\, log_2\left((1+x)^{\frac{1}{x}}\right)$ and $\lim_{x\to\infty}(1+x)^{\frac{1}{x}} = e$, we have

$$\frac{C}{B} = \log_2\left(1+\frac{P}{N_0B}\right)$$

$$= \frac{C}{B}\frac{E_b}{N_0}\log_2\left(1+\frac{C}{B}\frac{E_b}{N_0}\right)^{\frac{N_0B}{CE_b}}$$

We can then simplify the above as

$$\frac{E_b}{N_0}\log_2\left(1+\frac{C}{B}\frac{E_b}{N_0}\right)^{\frac{N_0B}{CE_b}} = 1$$

If $\frac{C}{B}\to\ \infty$ or $B\to\ \infty$ then we obtain

$$\frac{E_b}{N_0} = \frac{1}{\log_2 \mathrm{e}} = 0.693$$

ii) Capacity bound $R_b = C$

- When $R_b \leq C \rightarrow$ error-free transmission is possible
- When $R_b > C \rightarrow$ error-free transmission is not possible