An Introduction to Information Theory

Vahid Meghdadi
reference : Elements of Information Theory by Cover and Thomas

September 2007

Contents

1 Entropy

Entropy is a measure of uncertainty of a random variable. The uncertainty or the amount of information containing in a message (or in a particular realization of a random variable) is defined as the inverse of the logarithm of its probability: $\log(1/P_X(x))$. So, less likely outcome carries more information. Let X be a discrete random variable with alphabet $\mathcal{X}$ and probability mass function $P_X(x) = \mathtt{Pr}\{X = x\}$, $x \in \mathcal{X}$. For convenience $P_X(x)$ will be denoted by $p(x)$. The *entropy* of X is defined as follows:

Definition 1. The entropy $H(X)$ of a discrete random variable is defined by

$$\begin{aligned} H(X) &= \mathrm{E} \log \frac{1}{p(x)} \\ &= \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)} \end{aligned} \tag{1}$$

Entropy indicates the average information contained in X. When the base of the logarithm function is 2, the entropy is measured in bits. For example the entropy of a fair coin toss is 1 bit.
Note: The entropy is a function of the distribution of X. It does not depend on the actual values taken by the random variable, but only on the probabilities.
note: $H(X) \geq 0$.

Example 1. : Let

$$X = \begin{cases} 1 & \mathtt{with\ probability}\ p \\ 0 & \mathtt{with\ probability}\ 1-p \end{cases} \tag{2}$$

Show that the entropy of X is

$$H(X) = -p \log p - (1-p) \log(1-p) \tag{3}$$

Some times this entropy is denoted by $H(p, 1-p)$. Note that the entropy is maximized for $p = 0.5$ and it is zero for $p = 1$ or $p = 0$. This makes sense because when $p = 0$ or $p = 1$ there is no uncertainty over the random variable X and hence no information in revealing its outcome. The maximum uncertainty is when the two events are equi-probable.

Example 2. : Suppose X can take on K values. Show that that the entropy is maximized when X is uniformly distributed on these K Values and in this case, $H(X) = \log K$.

Solution: Calculating $H(X)$ results in:

$$H(X) = \sum_{x=1}^{K} p(x) \log \frac{1}{p(x)}$$

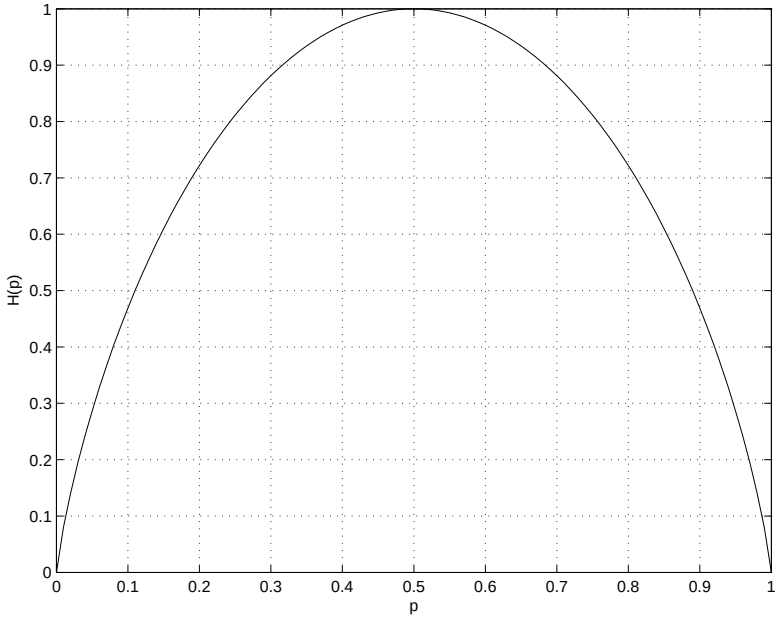

Figure 1: $H(p)$ versus p

Since $\log(x)$ is a concave function, using Jenson inequality for concave function $f(x)$ stated below:

$$\sum \lambda_i f(x_i) \leq f\left(\sum \lambda_i x_i\right)$$

it is clear that

$$H(X) = \sum_x p(x) \log \frac{1}{p(x)} \leq \log\left(\sum_x p(x) \frac{1}{p(x)}\right)$$

So $H(x) \leq \log K$. It means that the maximum value for $H(x)$ can be $\log K$. Choosing $p(X = i) = 1/K$, we can obtain $H(X) = \log 1/K$. So uniformly distributed X maximize the entropy and this entropy is $\log K$.

There is another way to solve this problem using Lagrange multipliers. We are going to maximize

$$H(X) = -\sum_{k=1}^{K} P(X = k) \log P(X = k)$$

The constraint of this maximization is

$$g(p_1 + p_2 + ... + p_K) = \sum_{k=1}^{K} p_k = 1$$

So by changing p_i we try to find the maximum point of H. For all k from 1 to K we should maximize

$$H + \lambda(g - 1)$$

Hence we require:

$$\frac{\partial}{\partial p_k}\left(H + \lambda(g-1)\right) = 0$$

It means that:

$$\frac{\partial}{\partial p_k}\left(-\sum_{k=1}^{K} p(k)\log p(k) + \lambda(\sum_{k=1}^{K} p_k - 1)\right) = 0$$

After calculating the differentiation we obtain a set of K independent equations as:

$$-\left(\frac{1}{\ln 2} + \log_2 p_k\right) + \lambda = 0$$

Solving this equation, we remark that all the p_k have the same value. So the

$$p_k = \frac{1}{K}$$

As a conclusion, the uniform distribution yields the greatest entropy.

2 Joint and conditional entropy

We saw the entropy of a single random variable (RV) and we now extend it to a pair of RV.

Definition 2. The joint entropy $H(X,Y)$ of a pair of discrete random variables (X,Y) with a joint distribution $p(x,y)$ is defined as:

$$H(X,Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y) \tag{4}$$

which can also be expressed as

$$H(X,Y) = -\,\mathrm{E}\log p(X,Y) \tag{5}$$

The conditional entropy is defined at the same way. It is the expectation of the entropies of the conditional distributions:

Definition 3. If $(X,Y) \sim p(x,y)$, then the conditional entropy $H(Y|X)$ is defined as:

$$\begin{aligned} H(Y|X) &= \sum_{x\in\mathcal{X}} p(x)H(Y|X=x) && (6)\\ &= -\sum_{x\in\mathcal{X}} p(x)\sum_{y\in\mathcal{Y}} p(y|x)\log p(y|x) && (7)\\ &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y|x) && (8)\\ &= \mathrm{E}_{p(x,y)}\log p(Y|X) && (9) \end{aligned}$$

Theorem 1. (chain rule)

$$H(X,Y) = H(X) + H(Y|X) \tag{10}$$

The proof comes from the fact that

$$\log p(X,Y) = \log p(X) + \log p(Y|X)$$

and then take the expectation.

3 Mutual information

The mutual information is a measure of the amount of information that one random variable contains about another random variable. It is the reduction of uncertainty of one random variable due to the knowledge of the other. It is:

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) && (11)\\ &= \mathrm{E}_{p(x,y)}\log\frac{p(X,Y)}{p(X)p(Y)} && (12)\\ &= H(Y) - H(Y|X) && (13)\\ &= I(Y;X) && (14) \end{aligned}$$

The diagram of figure 2 represents the relation between the conditional entropies and mutual information.

The chain rule can be stated here for the mutual information. First we define the conditional mutual information as:

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z) \tag{15}$$

Using this definition the chain rule can be written.

$$\begin{aligned} I(X_1,X_2;Y) &= H(X_1,X_2) - H(X_1,X_2|Y) && (16)\\ &= H(X_1) + H(X_2|X_1) - H(X_1|Y) - H(X_2|X_1,Y) && (17)\\ &= I(X_1;Y) + I(X_2;Y|X1) && (18) \end{aligned}$$

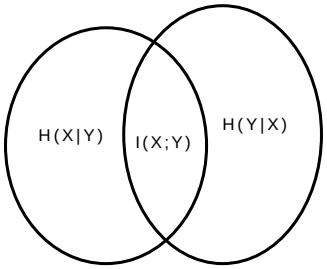

Figure 2: Relation between entropy and mutual information

Some properties

- $$I(X;Y) \geq 0$$
- $$I(X;Y|Z) \geq 0$$
- $H(X) \leq \log |\mathcal{X}|$ where $|\mathcal{X}|$ denotes the number of elements in the range of X, with the equality if and only if X has a uniform distribution over $\mathcal{X}$.
- Condition reduces entropy:

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

  with equality if and only if X and Y are independent.
- $H(X_1, X_2, \ldots, X_n) \leq \sum_{i=1}^{n} H(X_i)$ with equality if and only if the X_i are independent.
- Let $(X, Y) \sim p(x, y) = p(x)p(y|x)$. The mutual information $I(X;Y)$ is a concave function of $p(x)$ for fixed $p(y|x)$ and a convex function of $p(y|x)$ for fixed $p(x)$.

4 Data Compression or Source Coding

Two classes of coding are used: lossless source coding and lossy source coding. In this section, only lossless coding will be considered. The data compression can be achieved by assigning short descriptions to the most frequent outcomes of data source and then longer to the less probable. For example in Morse code, the letter "e" is coded by just a point.

In this chapter some principles of compression will be given.

Definition 4. A source code C for a random variable X is a mapping from $\mathcal{X}$, the range of X, to $\mathcal{D}$, the set of finite length strings of symbols from a D-ary alphabet. Let $C(x)$ denote the codeword corresponding to x and let $l(x)$ denote

the length of $C(x)$.

Example 3. : If you toss a coin, $\mathcal{X} = \{\texttt{tail}, \texttt{head}\}$, $C(\texttt{head}) = 0, C(\texttt{tail}) = 11, l(\texttt{head}) = 1, l(\texttt{tail}) = 2$.

Definition 5. The expected length of a code $C(x)$ for a random variable X with probability mass function $p(x)$ is given by

$$L(C) = \sum_{x \in \mathcal{X}} p(x) l(x) \tag{19}$$

For example for the above example the expected length of the code is

$$L(C) = \frac{1}{2} * 2 + \frac{1}{2} * 1 = 1.5$$

Definition 6. The code is singular if $C(x_1) = C(x_2)$ and $x_1 \neq x_2$.

Non singular codes are uniquely decodable.

Definition 7. The extension of a code C is a code obtained as:

$$C(x_1 x_2 \dots x_n) = C(x_1) C(x_2) \dots C(x_n) \tag{20}$$

It means that a long message can be coded by concatenating the shorter message code words. For example if $C(x_1) = 11$ and $C(x_2) = 00$, then $C(x_1 x_2) = 1100$.

Definition 8. A code is uniquely decodable if its extension is non-singular.

In other words, any encoded string has only one possible source string and there is no ambiguity.

Definition 9. A code is called a *prefix code* or an *instantaneous code* if no code word is a prefix of any other codeword.

Example 4. The following code is a prefix code: $C(x_1) = 1, C(x_2) = 01, C(x_3) = 001, C(X_4) = 000$. Any encoded sequence is uniquely decodable and its corresponding source word can be obtained as soon as the code word is received. In other word, an instantaneous code can be decoded without reference to the future codewords since the end of a codeword is immediately recognizable. For example the sequence 001100000001 is decoded as $x_3 x_1 x_4 x_4 x_2$.

Example 5. The following code is not instantaneous code but uniquely decodable: $C(x_1) = 1, C(x_2) = 10, C(x_3) = 100, C(X_4) = 000$. Why? Here you should wait to receive a 1 to be able to decode.
Note that if we look at the encoded sequence from right to left, it becomes instantaneous.

Figure 3 illustrates the different nesting of codes.

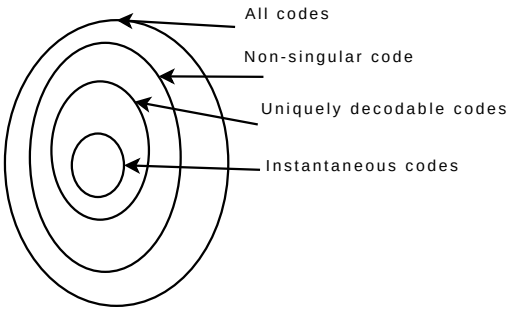

Figure 3: Classes of codes

5 Channel capacity

Assume that we have a channel whose input is the random variable X with input alphabet $\{0, 1, 2, 3\}$ and output alphabet $\{A, B, C, D\}$, as presented in Figure 4. The goal is to send the information without error. If we send 2 bits per input symbol, there is no way to determine precisely which symbol was sent. However, if we use less rate, for example just one bit per channel use it is possible to send information without error. In this case we use only the symbols 0 and 2 (or 1 and 3) and at the channel output we can precisely determine the symbol sent. It means by reducing the rate, reliable transmission is possible. What we did is to modify the $P_X(x)$ to maximize the rate and at the same time to obtain a reliable communication. In this example $P(X = 0) = P(X = 2) = 0.5$ and $P(X = 1) = P(X = 3) = 0$.
In this way we proposed a scheme that achieves 1 bit per channel use. Is it the maximum that we can obtain? The answer is yes because $H(Y)$ is the entropy of Y which is at most equal to 2 (there are 4 possibilities), so $H(Y) \leq 2$. The conditional entropy $H(Y|X)$ explains the uncertainty over Y given X. But if X is known, there is two equi-probable possibilities for Y giving this entropy equals to 1. So

$$[H(Y) - H(Y|X)] \leq 2 - 1 = 1$$

Therefor the information rate cannot be greater than 1, so the scheme proposed is optimal.
Note: If we reduce the rate below a certain number, reliable communication can be obtained.

Definition 10. The channel capacity of a discrete memoryless channel is defined as:

$$\begin{aligned} C &= \max_{p(x)} I(X;Y) \\ &= \max_{p(x)} [H(Y) - H(Y|X)] \end{aligned} \tag{21}$$

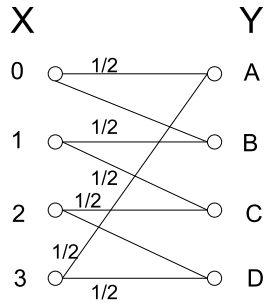

Figure 4: Noisy channel

5.1 examples

5.1.1 Noiseless binary channel

Consider the channel presented in Figure 5. Show that the capacity is 1 bit per symbol (or per channel use).

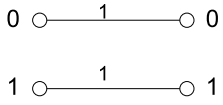

Figure 5: Ideal channel

5.1.2 Binary symmetric channel

For the binary symmetric channel (BSC) of Figure 6 show that the capacity is

$$C_{BSC} = 1 - H(p, 1-p) \tag{22}$$

This channel is equivalent to a channel with $Y = X \oplus Z$ where

$$Z = \begin{cases} 1 & \texttt{prob } p \\ 0 & \texttt{prob } 1-p \end{cases}$$

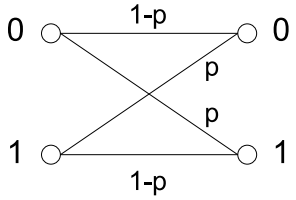

Figure 6: BSC channel

and X and Z are independent. We can say that $H(Y|X) = H(Z)$ because when X is given, knowing Y is the same as knowing Z. At the output, Y has only two possibilities, so we can say $H(Y) \leq 1$. Maximum is achievable by choosing $P_X(0) = P_X(1) = 1/2$. So $I(X;Y) \leq 1 - H(Z)$. The capacity should be: $C_{BSC} = 1 - H(p, 1-p)$.

Exercise

We are using a continuous AWGN channel where the input is a random variable $X \in \{3, -3\}$ and the noise is Gaussian with $N \sim \mathcal{N}(0,1)$. The channel output is: $Y = X + N$. We want to calculate the capacity.
Note: The capacity is not $\log(1 + SNR)$ because the distribution of X is not optimized (the capacity maximizing distribution in this case is Gaussian while here, X can only take two values).
Hint: Put a threshold at 0 and then calculate the probability of error. The channel is now BSC and you can use the results given in this section. Note that the capacity of this channel supposing that the noise variance goes to zero (high SNR) cannot be greater than 1 bit per channel use.

5.1.3 Binary erasure channel

The binary erasure channel is when some bits are lost (rather than corrupted). Here the receiver knows which bit has been erased. Figure 7 shows this channel. We are to calculate the capacity of binary erasure channel.

$$\begin{aligned} C &= \max_{p(x)} I(X;Y) \\ &= \max_{p(x)} [H(Y) - H(Y|X)] \\ &= \max_{p(x)} [H(Y) - H(a, 1-a)] \end{aligned} \tag{23}$$

Because of the symmetry we assume that $P(X = 0) = P(X = 1) = 1/2$. So Y have three possibilities with probabilities $P(Y = 0) = P(Y = 1) = (1 - a)/2$ and $P(Y = e) = a$. So we can write:

$$\begin{aligned} C &= H\left(\frac{1-a}{2}, \frac{1-a}{2}, a\right) - H(a, 1-a) \\ &= 1 - a \texttt{ bit per channel use} \end{aligned} \tag{24}$$

How to attain this capacity? If there is a feedback from the receiver, each time an erasure is detected, the receiver asks the transmitter to resend the erased bit. Because of the erasure probability of a, it happens to repeat a symbol once every $1/a$ symbols. This means that for a large sequence of N symbols sent (or N channel use) only $N - aN$ information bits are passed to the receiver. So with this structure the rate $1 - a$ bit per channel use is achieved.
However, it can be proved that this rate is the maximal rate even if there is no feedback. In fact, *feedback does not increase the capacity of discrete memoryless channels.*

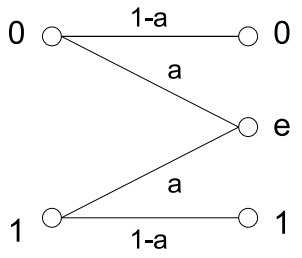

Figure 7: Erasure channel

5.1.4 Two fold channel

Suppose we have two independent memoryless channels drawn in Figure 8. Suppose also that the receiver knows which of the channels is used. The transmitter uses the channel following the symbol to be transmitted. In other word, two different alphabets are used for channels 1 and 2. We are to calculate the capacity of the channel: $\max I(X;Y)$.
Let's define a new variable S as:

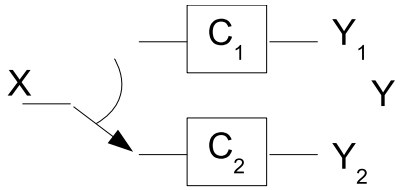

Figure 8: Two channel configuration

$$S = \begin{cases} 1 & \texttt{if channel 1 is used with prob } a \\ 2 & \texttt{if channel 2 is used with prob } 1-a \end{cases}$$

We have $I(X;Y) = I(X;Y,S)$. This is because the received alphabet are distinct so knowing Y implies that we know $\mathbf{S}$. However to prove this, using chain rule we can write: $I(X;Y,S) = I(X;Y) + I(X;S|Y)$. The second term is $H(S|Y) - H(S|Y,X)$ which is equal to zero because if we know Y we know S. So the above equation is proved.
Developing the above relation we have:

$$\begin{aligned} I(X;Y) &= I(X;Y,S) \\ &= I(X;S) + I(X;Y|S) \\ &= H(S) - H(S|X) + I(X;Y|S=1)P(S=1) + I(X;Y|S=2)P(S=2) \\ &= H(a,1-a) + 0 + I(X;Y_1)a + I(X;Y_2)(1-a) \end{aligned}$$

To maximize the mutual information with respect to a, P_{X_1}, P_{X_2}, we calculate the derivative equals to zero with respect to a, which gives:

$$a = \frac{2^{C_1}}{2^{C_1} + 2^{C_2}}$$

Replacing this value in I we obtain:

$$C = \log_2(2^{C_1} + 2^{C_2}) \tag{25}$$

6 Differential entropy

In this section we consider continuous random variables rather than discrete random variables. The entropy as defined before uses the probability mass function and does not work here. Instead, we use the probability density function (PDF) to define the entropy of X.

Definition 11. The random variable X is said to be continuous if its cumulative distribution function $F(x) = \texttt{Pr}(X \leq x)$ is continuous.

Definition 12. The differential entropy $h(X)$ of a continuous random variable X with a PDF $P_X(x)$ is defined as

$$\begin{aligned} h(X) &= \int_S P_X(x) \log \frac{1}{P_X(x)} dx \\ &= \mathrm{E}\left[\log \frac{1}{P_X(x)}\right] \end{aligned} \tag{26}$$

where S is the support set of the random variable.

Uniform distribution

Show that for $X \sim U(0, a)$ the differential entropy is $\log a$. Note that unlike discrete entropy, the differential entropy can be negative. However, $2^{h(X)} = 2^{\log a} = a$ is the volume of the support set, which is always non-negative, as expected.

Normal distribution

Show that for $X \sim \mathcal{N}(0, \sigma^2)$ the differential entropy is

$$h(x) = \frac{1}{2} \log(2\pi e \sigma^2) \texttt{ bits} \tag{27}$$

Exponential distribution

Show that for $P_X(x) = \lambda e^{-\lambda x}$ for $X \geq 0$ the differential entropy is

$$h(x) = \log \frac{e}{\lambda} \texttt{ bits} \tag{28}$$

What is the entropy if $P_X(x) = \frac{\lambda}{2} e^{-\lambda |x|}$?

6.1 Relation between differential and discrete entropy

Referring to the Figure 9, the continuous random variable X is quantized to generate a discrete random variable denoted by X^{Δ}. This random variable takes the value x_i if $i\Delta \leq X \leq (i+1)\Delta$. Then the probability that $X^{\Delta} = x_i$ is

$$p_i = \int_{i\Delta}^{(i+1)\Delta} P_X(x)dx \tag{29}$$

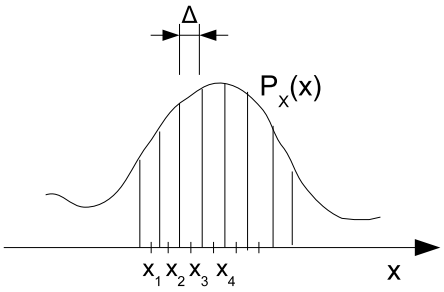

Figure 9: Quantization of a continuous random variable

Now, because X^{Δ} is a discrete random variable,m we can write the discrete entropy as:

$$\begin{aligned} H(X^{\Delta}) &= -\sum_{-\infty}^{\infty} p_i \log p_i \\ &= -\sum_{-\infty}^{\infty} P_X(x_i)\Delta \log(P_X(x_i)\Delta) \\ &= -\sum \Delta P_X(x_i) \log P_X(x_i) - \log \Delta \end{aligned} \tag{30}$$

and as $\Delta \to 0$ we can write:

$$H(X^{\Delta}) + \log \Delta \to h(X) \tag{31}$$

The important result is *the entropy of an n-bit quantization of a continuous random variable X is approximately $h(X) + n$.*

6.2 joint and conditional entropy

The differential entropy can be extended to several random variables. so:

$$h(X_1, X_2, ..., X_n) = -\int p(x_1, x_2, ..., x_n) \log p(x_1, x_2, ..., x_n) dx_1 dx_2 ... dx_n \tag{32}$$

$$h(X|Y) = -\int p(x,y)\log p(x|y)dxdy \quad (33)$$

$$= h(X,Y) - h(Y) \quad (34)$$

Theorem 2. (Entropy of multivariate normal distribution): Let $\mathbf{X}$ be a random normal vector with mean vector μ and covariance matrix $\mathbf{K}$. Then

$$h(X_1, X_2, ..., X_n) = \frac{1}{2}\log\left((2\pi e)^n|\mathbf{K}|\right) \quad \texttt{bits} \quad (35)$$

Note that the mean of the distribution has no effect on entropy. In general:

$$h(Y) = h(Y + cte)$$

6.3 Some properties

- Chain rule $h(X,Y) = h(X) + h(Y|X)$
- Uncorrelated Gaussian random vector $X = [X_1X_2...X_n]^T$ with $X_1, X_2, ..., X_n$ i.i.d. $\sim \mathcal{N}(0,1)$
$$h(\mathbf{X}) = \frac{1}{2}\log(2\pi e)^n$$
- Given a random vector $\mathbf{X}$ with $h(\mathbf{X})$, the differential entropy of the random vector $\mathbf{Y} = \mathbf{AX}$ will be
$$h(\mathbf{Y}) = h(\mathbf{X}) + \log|\mathbf{A}|$$
- The same case but for scalar random variable. If $Y = cX$, we have $h(Y) = h(X) + \log|c|$. Note: For discrete random variables if $Y = cX$, the entropy of X and Y are the same: $H(X) = H(Y)$.

Theorem 3. Suppose $\mathbf{X}$ is a random vector with $\mathrm{E}(\mathbf{X}) = \mathbf{0}$ and $\mathrm{E}(\mathbf{XX}^T) = \mathbf{K}$, then $h(\mathbf{X}) \leq \frac{1}{2}\log(2\pi e)^n|\mathbf{K}|$. The equality is achieved only if $\mathbf{X}$ is Gaussian $\sim \mathcal{N}(\mathbf{0}, \mathbf{K})$

Application (Hadamard's inequality)

Let $\mathbf{X} \sim \mathcal{N}(0, \mathbf{K})$ be a multi-variant normal random variable, then the Hadamard inequality states that:

$$|\mathbf{K}| = \prod_{i=1}^{n} \mathbf{K}_{ii}$$

Proof. Using chain rule, we can write:

$$\begin{aligned} h(\mathbf{X}) &= h(X_1) + h(X_2|X_1) + h(X_3|X_1, X_2) + \cdots + h(X_n|X_1, \cdots, X_n) \\ \frac{1}{2}\log(2\pi e)^n|\mathbf{K}| &\leq h(X_1) + h(X_2) + \cdots + h(X_n) \\ &= \frac{1}{2}\log(2\pi e)K_{11} + \frac{1}{2}\log(2\pi e)K_{22} + \cdots + \frac{1}{2}\log(2\pi e)K_{nn} \\ &= \frac{1}{2}\log(2\pi e)^n K_{11}K_{22}\cdots K_{nn} \end{aligned}$$

□

7 The Gaussian channel

A Gaussian channel is a time discrete channel presented on Figure 10. The input output relationship at instant i is: $Y_i = X_i + Z_i$. Z_i is an i.i.d. zero mean Gaussian process with power $P_N = \sigma^2$.

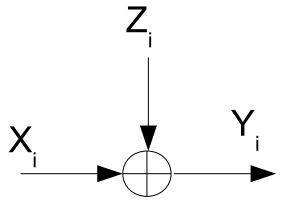

Figure 10: The Gaussian channel

7.1 Capacity of Gaussian channel

The capacity of the channel is defined as the maximum of the mutual information between input and output over all distribution on the input that satisfy the power constraint:

$$C = \max_{p(x):EX^2 \leq P} I(X;Y) \tag{36}$$

In order to calculate this for a Gaussian channel, we expand $I(X;Y)$:

$$\begin{aligned} I(X;Y) &= h(Y) - h(Y|X) \\ &= h(Y) - h(X+Z|X) \\ &= h(Y) - h(Z|X) \\ &= h(Y) - h(Z) \end{aligned} \tag{37}$$

For the Gaussian process Z the entropy is $h(Z) = \frac{1}{2}\log 2\pi eN$ where N is the noise variance (or power). Since X and Z are independent, $EY^2 = P+N$ where P is the power of X. To maximize the mutual information, one should maximize

$h(Y)$ with the power constraint of $P_Y = P + N$. We saw that the distribution maximizing the entropy for a continuous random variable is Gaussian. This can be obtain if X is Gaussian. Applying this to the mutual information formula (37), we obtain

$$\begin{aligned} I(X;Y) &= h(Y) - h(Z) \\ &\leq \frac{1}{2}\log 2\pi e(P+N) - \frac{1}{2}\log 2\pi eN \\ &= \frac{1}{2}\log\left(1+\frac{P}{N}\right) \end{aligned} \tag{38}$$

Hence the information capacity of a Gaussian channel is

$$C = \max_{p(x):EX^2\leq P} I(X;Y) = \frac{1}{2}\log\left(1+\frac{P}{N}\right) \tag{39}$$

and this maximum is attained when $X \sim \mathcal{N}(0, P)$.

Example 6. What is the capacity of the following transmission system:

$$\mathbf{Y} = 3\mathbf{X} + \mathbf{N}$$

where $\mathrm{E}\,\mathbf{X}^2 \leq P_x$ and $N \sim \mathcal{N}(0, P_N)$.
Let $\mathbf{X'} = 3\mathbf{X}$ so $P_{X'} \leq 9P_x$ and the capacity of this channel will be:

$$C = \frac{1}{2}\log(1+\frac{9P_x}{P_N})$$

Therefore the capacity is increased. Here the channel gain is a deterministic value and the channel is flat.

7.2 Band limited channel

Suppose we have a continuous channel with bandwidth B and the power spectral density of noise is $N_0/2$. So the analog noise power is N_0B. On the other hand, supposing that the channel is used over the time interval $[0, T]$. So the power of analog signal times T gives the total energy of the signal in this period. Using Shannon sampling theorem, there are $2B$ samples per second. So the power of discrete signal per sample will be $PT/2BT = P/2B$. The same argument can be used for the noise, so the power of samples of noise is $\frac{N_0}{2}2B\frac{T}{2BT} = N_0/2$. So the capacity of the Gaussian channel per sample is:

$$C = \frac{1}{2}\log\left(1+\frac{P}{N_0B}\right) \texttt{ bits per sample} \tag{40}$$

Since there are maximum 2B independent samples per second the capacity can be written as:

$$C = B\log\left(1+\frac{P}{N_0B}\right) \texttt{bits per second} \tag{41}$$

Example 7. For the channel presented in the figure 11 , what is the capacity? We can imagine that there are two parallel independent channels as presented in figure 12 with just one power constraint. Let P_1 and P_2 be the power transmitted through the first and the second channel, respectively. In this case the channel capacity can be written as:

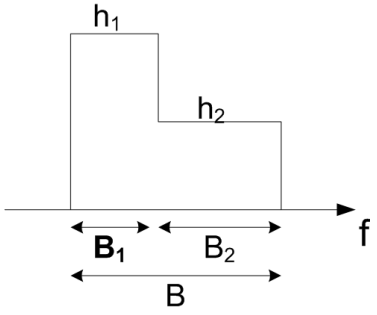

Figure 11: Channel used in the example

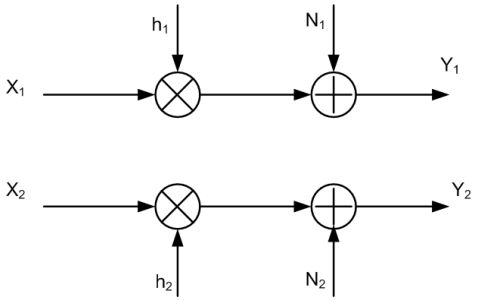

Figure 12: Channel model of figure 11

$$C = \max_{P_1+P_2 \leq P_x} \left[B_1 \log(1 + \frac{P_1 h_1^2}{N_0 B_1}) + B_2 \log(1 + \frac{P_2 h_2^2}{N_0 B_2}) \right] \tag{42}$$

So we should maximize C subjected to $P_1 + P_2 \leq P_x$. Using Lagrangian, one can define:

$$L(P_1, P_2, \lambda) = B_1 \log(1 + \frac{P_1 h_1^2}{N_0 B_1}) + B_2 \log(1 + \frac{P_2 h_2^2}{N_0 B_2}) - \lambda(P_1 + P_2 - P_x)$$

Let $d(.)/dp_1 = 0$ and $d(.)/dp_2 = 0$ and using ln instead of $\log_2$:

$$\frac{B_1}{1 + \frac{P_1 h_1^2}{N_0 B_1}} \frac{h_1^2}{N_0 B_1} = \lambda$$

$$\frac{P_1}{B_1 N_0} = \frac{1}{\lambda N_0} - \frac{1}{h_1^2}$$

With the same operations we obtain:

$$\frac{P_1}{B_1 N_0} = Cst - \frac{1}{h_1^2} \quad (43)$$

$$\frac{P_2}{B_2 N_0} = Cst - \frac{1}{h_2^2} \quad (44)$$

Where the Cte can be found by setting $P_1 + P_2 = P_x$. Since the two powers are found, the capacity of the channel is calculated using equation 42. The only constraint that to be considered is that P_1 and P_2 cannot be negative. If one of these is negative, the corresponding power is zero and all the power are assigned to the other one. This principle is called *water filling*.

Exercise 1. Use the same principle (water filling) and give the power allocation for a channel with three frequency bands defined as follows: $h_1 = 1/2$, $h_2 = 1/3$ and $h_3 = 1$; $B_1 = B$, $B_2 = 2B$ and $B_3 = B$; $P_x = P_1 + P_2 + P_3 = 10$.
solution: $P_1 = 3.5$, $P_2 = 0$ and $P_3 = 6.5$.

7.3 Parallel Gaussian channel

Here we consider k independent Gaussian channels in parallel with a common power constraint as depicted in Figure 13. The objective is to maximize the capacity by optimal distribution of the power among the channels:

$$C = \max_{p_{X_1,\ldots,X_k}(x_1,\ldots,x_k):\sum EX_i^2 \leq P} I(X_1, ..., X_k; Y_1, ..., Y_k) \quad (45)$$

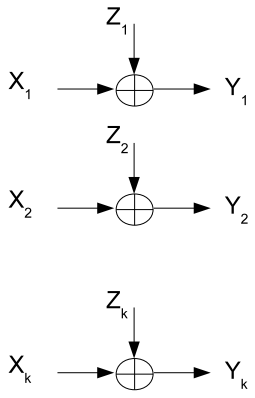

Figure 13: Parallel Gaussian channels

Using the independence of $Z_1, ..., Z_k$:

$$\begin{aligned} C &= I(X_1, ..., X_k; Y_1, ..., Y_k) \\ &= h(Y_1, ..., Y_k) - h(Y_1, ..., Y_k | X_1, ..., X_k) \\ &= h(Y_1, ..., Y_k) - h(Z_1, ..., Z_k) \\ &\leq \sum_i h(Y_i) - h(Z_i) \\ &\leq \sum_i \frac{1}{2} \log\left(1 + \frac{P_i}{N_i}\right) \end{aligned}$$

8 Capacity of SIMO channel

Consider the channel presented in figure 14. **X** is a binary random variable with $P_x(1) = P_x(0) = 1/2$. Z_1 and Z_2 are two binary random variables (i.i.d) representing the effect of channel noise with $P(Z_1 = 1) = P(Z_2 = 1) = p$. The capacity of the channel is $C_1 = I(X; (Y_1, Y_2))$. We can add a data processing unit at the output as presented in figure 15. Now the whole channel is called C_2. We can write for Z the following equation.

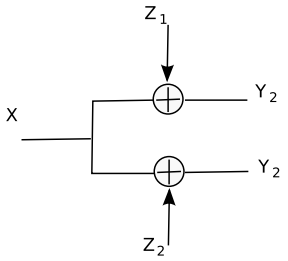

Figure 14: SIMO channel

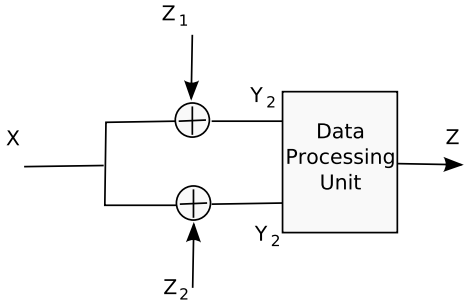

Figure 15: SIMO channel with data processing

$$Z = \begin{cases} 1 & if \quad Y_1 = Y_2 = 1 \\ 0 & if \quad Y_1 = Y_2 = 0 \\ \epsilon_1 & if \quad Y_1 = 1, Y_2 = 0 \\ \epsilon_2 & if \quad Y_1 = 0, Y_2 = 1 \end{cases}$$

The capacity of C_2 can be calculated using the probabilities $P(Z|X)$. Also with the same principle the capacity of C_1 can be calculated. But which one is greater C_1 or C_2?
First C_2 cannot be greater than C_1, it means that $C_1 \geq C_2$; that is because *signal processing cannot increase the capacity*. Second, here, there is no loss of information because of signal processing. It means that the processing is completely invertible: $(Y_1, Y_2) \Leftrightarrow Z$. So $C_1 = C_2$.
But if we had as data processing the following relation:

$$Z = \begin{cases} 1 & if \quad Y_1 = Y_2 = 1 \\ 0 & if \quad Y_1 = Y_2 = 0 \\ \epsilon & if \quad Y_1 \neq Y_2 \end{cases}$$

Here there is some information loss, what we can say in this case? First as before $C_1 \geq C_2$. Here we are not interested in Y but in X. So the two channels are equivalent. This can be shown mathematically.

$$I(X; Y_1, Y_2, Z) = I(X; Y_1, Y_2) + I(X; Z|Y_1, Y_2)$$

This is because $I(A; B, C) = I(A; B) + I(A; C|B)$. The first term in the above equation is the capacity of the first channel and the second term is equal to zero because if you know $Y1$ and Y_2 you know perfectly Z. We can also write:

$$I(X; Y_1, Y_2, Z) = I(X; Z) + I(X; Y_1, Y_2|Z)$$

The first term is the capacity of the second channel. If we show that the second term is zero, we have shown that $C_1 = C_2$. We can say:

$$\begin{aligned} I(X; Y_1, Y_2|Z) &= P(Z = 0) I(X; Y_1, Y_2|Z = 0) \\ &+ P(Z = 1) I(X; Y_1, Y_2|Z = 1) \\ &+ P(Z = \epsilon) I(X; Y_1, Y_2|Z = \epsilon) \end{aligned}$$

Since there is no information on Y_1 and Y_2 when $Z = 0$ or $Z = 2$, we can write:

$$\begin{aligned} I(X; Y_1, Y_2|Z) &= P(Z = \epsilon) I(X; Y_1, Y_2|Z = \epsilon) \\ &= P(Z = \epsilon)[H(Y_1, Y_2|Z = \epsilon) - H(Y_1, Y_2|Z = \epsilon, X)] \end{aligned}$$

It can be shown that the two terms are equal to one which gives zero as the result. It means that we have shown $C1 = C2$. So Z has all information of X; it is a sufficient statistics for X.
We are now interested in Gaussian variables resulting from Gaussian channels.

Suppose $X \sim \mathcal{N}(0, P)$, $N_1 \sim \mathcal{N}(0, 1)$ and $N_2 \sim \mathcal{N}(0, 1)$. N_1 and N_2 are jointly i.i.d Gaussian random variable.

$$Y_1 = X + Z_1$$

$$Y_2 = X + Z_2$$

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} X + Z_1 \\ X + Z_2 \end{bmatrix} = \mathbf{X} + \mathbf{Z}$$

The random variable Z is defined as $Z = Y_1 + Y_2$. It means that the signal processing unit is an addition block (here it is called maximum ration combiner). Is there any loss of information here?
To prove this, we construct the variable $\tilde{\mathbf{Y}}$ from the following invertible transformation.

$$\tilde{\mathbf{Y}} = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} \tilde{Y_1} \\ \tilde{Y_2} \end{bmatrix}$$

The question is that with this process is the capacity will be the same? First of all the processing cannot increase the capacity. Since the process is invertible there is no information loss. It means that $I(X; \mathbf{Y}) = I(X; \tilde{\mathbf{Y}})$.

$$\tilde{\mathbf{Y}} = \begin{bmatrix} Y_1 + Y_2 \\ -Y_1 + Y_2 \end{bmatrix} = \begin{bmatrix} Z \\ -Z_1 + Z_2 \end{bmatrix} = \begin{bmatrix} 2X + Z_1 + Z_2 \\ -Z_1 + Z_2 \end{bmatrix}$$

We can show that $Z_1 + Z_2$ is independent of $-Z_1 + Z_2$. That is because the two are Gaussian and uncorrelated:

$$\mathrm{E}((Z_2 + Z_1)(Z_2 - Z_1)) = 0$$

So we can write:

$$I(X; \mathbf{Y}) = I(X; \tilde{\mathbf{Y}}) = I(X; \tilde{Y_1}, \tilde{Y_2}) = I(X; \tilde{Y_1}) + I(X; \tilde{Y_2}|\tilde{Y_1})$$

and

$$\begin{aligned} I(X; \tilde{Y_2}|\tilde{Y_1}) &= I(X; -Z_1 + Z_2|2X + Z_1 + Z_2) \\ &= H(-Z_1 + Z_2|2X + Z_1 + Z_2) - H(-Z_1 + Z_2|2X + Z_1 + Z_2, X) \\ &= 0 \end{aligned}$$

This is true because on independence of $Z_1 + Z_2$ and $Z_2 - Z_1$. Therefore;

$$I(X; \mathbf{Y}) = I(X; Z)$$

So Z is sufficient statistic for X.

9 Exercise (to be completed)

1. Let

$$X = \begin{cases} a & \texttt{with probability 1/2,} \\ b & \texttt{with probability 1/4,} \\ c & \texttt{with probability 1/8,} \\ d & \texttt{with probability 1/8,} \end{cases}$$

What is entropy of X ? (answer: 7/4 bits)
Suppose we wish to determine the value of X with the minimum number of binary question. What is an efficient questions ? What is the expectation value of the number of questions ? (answer: 1.75)