1 Basic Definitions

Definition 1. Entropy: *We can restrict the surprise to the sample space and consider it to be a function from the sample space to the real numbers. The expected value of the surprise is the entropy of the probability distribution. If the sample space is $S = \{s_1, s_2, \cdots, s_N\}$, with probability distribution P, the entropy of the probability distribution is given by*

$$H(P) = -\sum_{i=1}^{N} P(s_i)\log(P(s_i)). \tag{1}$$

If we have a sample space with N elements, the maximum value of the entropy function is $\log(N)$.

Definition 2. Bayes Theorem: *If S is a sample space with a probability distribution function P, and E and F are events in S, then*

$$P(E|F) = \frac{P(F|E)P(E)}{P(F)} \tag{2}$$

Theorem 1. *The entropy of a Gaussian distribution with mean μ and variance σ^2 is $\ln(\sqrt{2\pi e}\sigma)$ in natural units.*

2 Information Channel

The following assumptions will apply in our modeling of an information channel:

- **Stationary:** The statistical nature of the channel and noise do not change with time.
- **Memoryless:** The behaviour of the channel and the effect of the noise at time t will not depend on the behaviour of the channel or the effect of noise at any previous time.

We now formally define a mathematical structure for an information channel.

Definition 3. Information channel: An information channel is a triple $\{A, B, P\}$, where A is the input alphabet, B is the output alphabet and P is the set of channel probabilities. $A = \{a_i : i = 1, 2, ...r\}$ is a discrete set of $r = |A|$ symbols (where $|A|$ is the size of the input alphabet), and $B = \{b_j : j = 1, 2, ...s\}$ is a discrete set of $s = |B|$ symbols. The transmission behaviour of the channel is described by the probabilities in $P = \{P(b_j|a_i) : i = 1, 2, ...r; j = 1, 2, ...s\}$, where $P(b_j|a_i)$ is the probability that the output symbol b_j will be received if the input symbol a_i is transmitted.

Note 1. *The input alphabet represents the symbols transmitted into the channel and the output alphabet represents the symbols received from the channel. The definition of the channel implies that the input and output symbols may be different. In reality, one would expect that the received symbols are the same as those transmitted. However the effect of noise may introduce "new" symbols and thus we use different input and output alphabets to cater for such cases. For more general applications the channel models the matching of the input symbols to prescribed output symbols or classes which are usually different. In statistical applications the input and output symbols arise from two random variables and the channel models the joint relationship between the variables.*

The conditional probabilities that describe an information channel can be represented conveniently using a matrix representation:

$$P = \begin{bmatrix} P(b_1|a_1) & P(b_2|a_1) & \cdots & P(b_s|a_1) \\ P(b_1|a_2) & P(b_2|a_2) & \cdots & P(b_s|a_2) \\ \vdots & \vdots & \ddots & \vdots \\ P(b_1|a_r) & P(b_2|a_r) & \cdots & P(b_s|a_r) \end{bmatrix} \tag{3}$$

where P is the channel matrix and for notational convenience we may sometimes rewrite this as:

$$P = \begin{bmatrix} P_{11} & P_{12} & \cdots & P_{1s} \\ P_{21} & P_{22} & \cdots & P_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ P_{r1} & P_{r2} & \cdots & P_{rs} \end{bmatrix} \tag{4}$$

where we have defined $P_{ij} = P(b_j|a_i)$.
A graphical representation of an information channel is given in Figure 1.
The channel matrix exhibits the following properties and structure:

- Each row of P contains the probabilities of all possible outputs from the same input to the channel.
- Each column of P contains the probabilities of all possible inputs to a particular output from the channel.

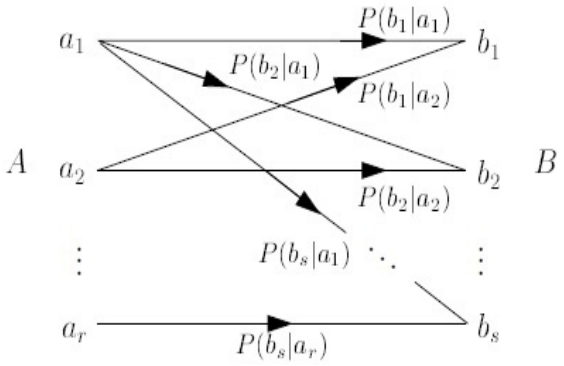


Figure 1: Graphical representation of an information channel.

- If we transmit the symbol a_i we must receive an output symbol with probability 1, that is:

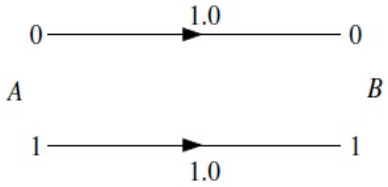
$$\sum_{j=1}^{s} P(b_j|a_i) = 1 \quad for\ i = 1, 2, ...r \tag{5}$$

that is, the probability terms in each row must sum to 1.

Example 1. *Consider a binary source and channel with input alphabet* $\{0, 1\}$ *and output alphabet* $\{0, 1\}$.

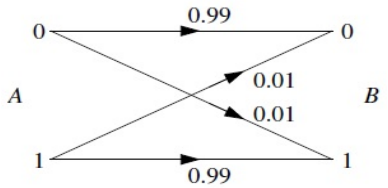
Proof.

- **Noiseless:** If the channel is noiseless there will be no error in transmission, the channel matrix is given by $P = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ and the channel is



A typical input-output sequence from this channel could be:
input: 0 1 1 0 0 1 0 1 1 0
output: 0 1 1 0 0 1 0 1 1 0

- **Noisy:** Say the channel is noisy and introduces a bit inversion 1% of the time, then the channel matrix is given by $P = \begin{bmatrix} 0.99 & 0.01 \\ 0.01 & 0.99 \end{bmatrix}$ and the channel is



A typical input-output sequence from this channel could be:
input: 0 1 1 0 0 1 0 1 1 0
output: 0 1 1 0 0 1 1 1 1 0

□

3 BSC and BEC Channels

In digital communication systems the input to the channel will be the binary digits $\{0, 1\}$ and this set will be the input alphabet and, ideally, also be the output alphabet. Furthermore, the effect of noise will not depend on the transmission pattern, that is, the channel is assumed *memoryless*. Two possible scenarios on the effect of noise are possible.
Ideally if there is no noise a transmitted 0 is detected by the receiver as a 0, and a transmitted 1 is detected by the receiver as a 1. However, in the presence of noise the receiver may produce a different result.
The most common effect of noise is to force the detector to detect the wrong bit (bit inversion), that is, a 0 is detected as a 1, and a 1 is detected as a 0. In this case the information channel that arises is called a *Binary Symmetric Channel* or BSC where $P(b = 1|a = 0) = P(b = 0|a = 1) = q$ is the probability of error (also called bit error probability, Bit Error Rate (BER), or "crossover" probability) and the output alphabet is also the set of binary digits $\{0, 1\}$. The parameter q fully defines the behaviour of the channel. The BSC is an important channel for digital communication systems as noise present in physical transmission media (fibre optic cable, copper wire, etc.) typically causes bit inversion errors in the receiver.
A BSC has channel matrix $P = \begin{bmatrix} p & q \\ q & p \end{bmatrix}$ where $p = 1 - q$ and is depicted in Figure 2.

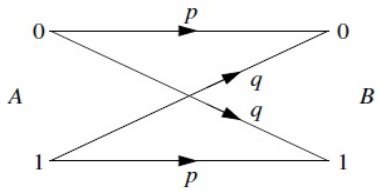


Figure 2: Binary symmetric channel.

Another effect that noise (or more usually, loss of signal) may have is to prevent the receiver from deciding whether the symbol was a 0 or a 1. In this case the output alphabet includes an additional symbol, ?, called the "erasure" symbol that denotes a bit that was not able to be detected. Thus for binary input $\{0, 1\}$, the output alphabet consists of the three symbols, $\{0, ?, 1\}$. This information channel is called a *Binary Erasure Channel* or BEC where $P(b =?|a = 0) = P(b =?|a = 1) = q$ is the probability of error (also called the "erasure" probability). Strictly speaking a BEC does not model the effect of bit inversion; thus a transmitted bit is either received correctly (with probability $p = 1 - q$) or is received as an "erasure" (with probability q). A BEC is becoming an increasingly important model for wireless mobile and satellite communication channels, which suffer mainly from dropouts and loss of signal leading to the receiver failing to detect any signal.
A BEC has channel matrix $P = \begin{bmatrix} p & q & 0 \\ 0 & q & p \end{bmatrix}$ and is depicted in Figure 3.

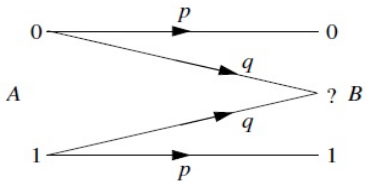


Figure 3: Binary erasure channel.

4 Mutual Information

To fully specify the behaviour of an information channel it is necessary to specify the characteristics of the input as well as the channel matrix. We will assume that the input characteristics are described by a probability distribution over the input alphabet, with $P(a_i)$ denoting the probability of symbol a_i being input to the channel. Then a channel will be fully specified if the input source probabilities, $P(A) = \{P(a_1), P(a_2), \cdots, P(a_r)\}$, and channel probabilities, $P(B) = [P(b_j|a_i)]_{j,i=1,1}^{s,r}$, are given.
If a channel is fully specified then the output probabilities, $P(B) = \{P(b_1), P(b_2), \cdots, P(b_s)\}$, can be calculated by:

$$P(b_j) = \sum_{i=1}^{r} P(b_j|a_i)P(a_i) \tag{6}$$

The probabilities $P(b_j|a_i)$ are termed the *forward probabilities* where forward indicates that the direction of channel use is with input symbol a_i being transmitted and output symbol b_j being received (i.e., a_i then b_j or b_j given a_i). We can similarly define the backward probabilities as $P(a_i|b_j)$ indicating the channel is running backwards: output

symbol b_j occurs first followed by input symbol a_i (i.e., b_j then a_i, or a_i given b_j). The backward probabilities can be calculated by application of Bayes Theorem as follows:

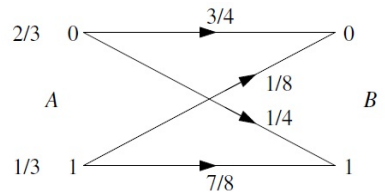
$$P(a_i|b_j) = \frac{P(a_i, b_j)}{P(b_j)} = \frac{P(b_j|a_i)P(a_i)}{P(b_j)} = \frac{P(b_j|a_i)P(a_i)}{\sum_{i=1}^{r} P(b_j|a_i)P(a_i)} \tag{7}$$

where $P(a_i, b_j)$ is the joint probability of a_i and b_j.

Example 2. *Consider the binary information channel fully specified by:*

$$P = \begin{bmatrix} 3/4 & 1/4 \\ 1/8 & 7/8 \end{bmatrix} \text{ and } \begin{matrix} P(a=0) = 2/3 \\ P(a=1) = 1/3 \end{matrix} \tag{8}$$

which is usually represented diagrammatically as:



Proof. The output probabilities are calculated as follows:

$$\begin{aligned} P(b=0) &= P(b=0|a=0)P(a=0) + P(b=0|a=1)P(a=1) \\ &= \frac{3}{4} \times \frac{2}{3} + \frac{1}{8} \times \frac{1}{3} = \frac{13}{24} \\ P(b=1) &= 1 - P(b=0) = \frac{11}{24} \end{aligned}$$

and the backward probabilities by:

$$\begin{aligned} P(a=0|b=0) &= \frac{P(b=0|a=0)P(a=0)}{P(b=0)} = \frac{(\frac{3}{4})(\frac{2}{3})}{(\frac{13}{24})} = \frac{12}{13} \\ P(a=1|b=0) &= 1 - P(a=0|b=0) = \frac{1}{13} \\ P(a=1|b=1) &= \frac{P(b=1|a=1)P(a=1)}{P(b=1)} = \frac{(\frac{7}{8})(\frac{1}{3})}{(\frac{11}{24})} = \frac{7}{11} \\ P(a=0|b=1) &= 1 - P(a=1|b=1) = \frac{4}{11} \end{aligned}$$

□

Conceptually we can characterise the probabilities as *a priori* if they provide the probability assignment *before* the channel is used (without any knowledge), and as *a posteriori* if the probability assignment is provided *after* the channel is used (given knowledge of the channel response). Specifically:

- $P(b_j)$ a priori probability of output symbol b_j if we *do not know* which input symbol was sent.
- $P(b_j|a_i)$ a posteriori probability of output symbol b_j if *we know* that input symbol a_i was sent.
- $P(a_i)$ a priori probability of input symbol a_i if *we do not know* which output symbol was received.
- $P(a_i|b_j)$ a posteriori probability of input symbol a_i if *we know* that output symbol b_j was received.

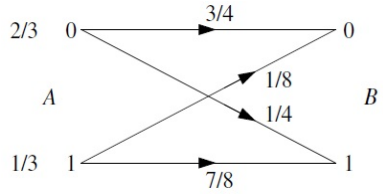
We can similarly refer to the *a priori entropy of A*:

$$H(A) = \sum_{a \in A} P(a) \log \frac{1}{P(a)} \tag{9}$$

as the average uncertainty we have about the input *before* the channel output is observed and the *a posteriori entropy of A given* b_j:

$$H(A|b_j) = \sum_{a \in A} P(a|b_j) \log \frac{1}{P(a|b_j)} \tag{10}$$

as the average uncertainty we have about the input *after* the channel output b_j is observed.
How does our average uncertainty about the input change after observing the output of the channel? Intuitively, we expect our uncertainty to be reduced as the channel output provides us with knowledge and knowledge reduces uncertainty. However, as we will see in the following example the output can sometimes increase our uncertainty (i.e., be more of a hindrance than a help!).



Example 3. *Consider the binary information channel from Example 2.*
What is our uncertainty of the input that is transmitted through the channel before we observe an output from the channel? This is provided by the entropy based on the given a priori input probabilities, $P(a=0)=\frac{2}{3}$ *and* $P(a=1)=\frac{1}{3}$*, yielding the a priori entropy of A,* $H(A)=0.918$.

Proof. What is our uncertainty of the input that is transmitted through the channel after we observe an output from the channel? Say we observe an output of $b=0$, then the a posteriori input probabilities, $P(a=0|b=0)=\frac{12}{13}$ and $P(a=1|b=0)=\frac{1}{13}$, yield an a posteriori entropy of A, $H(A|b=0)=0.391$. Thus we reduce our uncertainty once we observe an output of $b=0$. But what if we observe an output of $b=1$? Then the a posteriori input probabilities become $P(a=1|b=1)=\frac{7}{11}$ and $P(a=0|b=1)=\frac{4}{11}$ and the a posteriori entropy of A, $H(A|b=1)=0.946$. Our uncertainty is in fact increased! The reason is that our high expectation of an input of 0, from the a priori probability $P(a=0)=\frac{2}{3}$, is negated by receiving an output of 1. Thus $P(a=0|b=1)=\frac{4}{11}$ is closer to equi-probable than $P(a=0)=\frac{2}{3}$ and this increases the a posteriori entropy.
Notwithstanding the fact that $H(A|b=1)>H(A)$ even though $H(A|b=0)<H(A)$ we can show that if we average across all possible outputs the channel will indeed reduce our uncertainty, that is:

$$\begin{aligned} H(A|B) &= \sum_{j=1}^{2} P(b_j)H(A|b_j) \\ &= P(b=0)H(A|b=0)+P(b=1)H(A|b=1) \\ &= 0.645 \end{aligned}$$

and thus $H(A|B)<H(A)$. □

The average of the a posterior entropies of A, $H(A|B)$, calculated in Example 3 is sometimes referred to as the *equivocation of A with respect to B* where equivocation is used to refer to the fact that $H(A|B)$ measures the amount of uncertainty or equivocation we have about the input A when observing the output B. Together with the a priori entropy of A, $H(A)$, we can now establish a measure of how well a channel transmits information from the input to the output. To derive this quantity consider the following interpretations:

- $H(A)$ average uncertainty (or surprise) of the input to the channel *before* observing the channel output;
- $H(A|B$ average uncertainty (or equivocation) of the input to the channel *after* observing the channel output;
- $H(A)-H(A|B)$ reduction in the average uncertainty of the input to the channel *provided or resolved* by the channel.

Definition 4. Mutual Information: *For input alphabetA and output alphabet B the term*

$$I(A;B)=H(A)-H(A|B) \tag{11}$$

is the mutual information between A and B.

The mutual information, $I(A;B)$, indicates the information about A, $H(A)$, that is provided by the channel minus the degradation from the equivocation or uncertainty, $H(A|B)$. The $H(A|B)$ can be construed as a measure of the "noise" in the channel since the noise directly contributes to the amount of uncertainty we have about the channel input, A, given the channel output B.
Consider the following cases:

- **Noisefree:** If $H(A|B)=0$ this implies $I(A;B)=H(A)$ which means the channel is able to provide all the information there is about the input, i.e., $H(A)$. This is the best the channel will ever be able to do.
- **Noisy:** If $H(A|B)>0$ but $H(A|B)<H(A)$ then the channel is noisy and the input information, $H(A)$, is reduced by the noise, $H(A|B)$, so that the channel is only able to provide $I(A;B)=H(A)-H(A|B)$ amount of information about the input.
- **Ambiguous:** If $H(A|B)=H(A)$ the amount of noise totally masks the contribution of the channel and the channel provides $I(A;B)=0$ information about the input. In other words the channel is useless and is no better than if the channel was not there at all and the outputs were produced independently of the inputs!

An alternative expression to Equation 11 for the mutual information can be derived as follows:

$$\begin{aligned} I(A;B) &= H(A) - H(A|B) \\ &= \sum_{a\in A} P(a)\log\frac{1}{P(a)} - \sum_{a\in A}\sum_{b\in B} P(a,b)\log\frac{1}{P(a|b)} \\ &= \sum_{a\in A}\sum_{b\in B} P(a,b)\log\frac{1}{P(a)} - \sum_{a\in A}\sum_{b\in B} P(a,b)\log\frac{1}{P(a|b)} \\ &= \sum_{a\in A}\sum_{b\in B} P(a,b)\log\frac{P(a|b)}{P(a)} \end{aligned} \tag{12}$$

Using Equation 7 the mutual information can be expressed more compactly:

Result 1. Alternative expressions for Mutual Information:

$$I(A;B) = \sum_{a\in A}\sum_{b\in B} P(a,b)\log\frac{P(a,b)}{P(a)P(b)} = \sum_{a\in A}\sum_{b\in B} P(a)P(b|a)\log\frac{P(b|a)}{P(b)} \tag{13}$$

4.1 Importance of Mutual Information

The mutual information has been defined in the context of measuring the information carrying capacity of communication channels. However the concept of mutual information has had far-reaching effects on solving difficult estimation and data analysis problems in biomedical applications, image processing and signal processing. In these applications the key in using mutual information is that it provides a measure of the independence between two random variables or distributions.
In image processing and speech recognition the use of the *maximum mutual information* or MMI between the observed data and available models has yielded powerful strategies for training the models based on the data in a discriminative fashion. In signal processing for communication systems the idea of minimizing the mutual information between the vector components for separating mutually interfering signals has led to the creation of a new area of research for signal separation based on the idea of *independent component analysis* or ICA.

5 Channel Capacity: Maximum Mutual Information

Consider an information channel with input alphabet A, output alphabet B and channel matrix P_{AB} with conditional channel probabilities $P(b_j|a_i)$. The mutual information:

$$I(A;B) = \sum_{a\in A}\sum_{b\in B} P(a,b)\log\frac{P(a,b)}{P(a)P(b)} \tag{14}$$

which, if we now assume the logarithm is base 2, indicates the amount of information the channel is able to carry in bits per symbol transmitted. The maximum amount of information carrying capacity for the channel is $H(A)$, the amount of information that is being transmitted through the channel. But this is reduced by $H(A|B)$, which is an indication of the amount of noise present in the channel.
The expression for mutual information depends not only on the channel probabilities, $P(b_j|a_i)$, which uniquely identify a channel, but also on how the channel is used, the input or source probability assignment, $P(a_i)$. As such $I(A;B)$ cannot be used to provide a unique and comparative measure of the information carrying capacity of a channel since it depends on how the channel is used. One solution is to ensure that the same probability assignment (or input distribution) is used in calculating the mutual information for different channels. The questions then is: which probability assignment should be used? Obviously we cant use an input distribution with $H(A) = 0$ since that means $I(A;B) = 0$ for whatever channel we use! Intuitively an input distribution with maximum information content (i.e., maximum $H(A)$) makes sense. Although this allows us to compare different channels the comparison will not be fair since another input distribution may permit certain channels to exhibit a higher value of mutual information.
The usual solution is to allow the input distribution to vary for each channel and to determine the input distribution that produces the maximum mutual information for that channel. That is we attempt to calculate the maximum amount of information a channel can carry in any single use (or source assignment) of that channel, and we refer to this measure as the capacity of the channel.

Definition 5. Channel Capacity: *The maximum average mutual information, $I(A;B)$, in any single use of a channel defines the channel capacity. Mathematically, the channel capacity, C, is defined as:*

$$C = \max_{\{P_A(a)\}} I(A;B) \tag{15}$$

that is, the maximum mutual information over all possible input probability assignments,$P_A(a)$.

The channel capacity has the following properties:

1. $C \geq 0$ since $I(A;B) \geq 0$
2. $C \leq min\{\log|A|, \log|B|\}$ since $C = max\ I(A;B)$ and $max\ I(A;B) \leq max\ H(A) = \log|A|$ when considering the expression $I(A;B) = H(A) - H(A|B)$, and $max\ I(A;B) \leq max\ H(B) = \log|B|$ when considering the expression $I(A;B) = H(B) - H(B|A)$

The calculation of C involves maximization of $I(A;B)$ over $r = \log|A|$ independent variables (the input probabilities, $\{P(a_i) : i = 1, 2, \cdots, r\}$ subject to the two constraints:

1. $P(a_i) \geq 0 \quad \forall a_i$
2. $\sum_{i=1}^{r} P(a_i) = 1$

5.1 Channel Capacity of a BSC

For a BSC, $I(A;B) = H(B) - H(B|A)$. We want to examine how the mutual information varies with different uses of the same channel. For the same channel the channel probabilities, p and q, remain constant. However, for different uses the input probability, $\omega = P(a = 0)$ varies from 0 to 1. The maximum mutual information occurs at $P(a = 0) = \frac{1}{2}$. For $\omega = P(a = 0) = \frac{1}{2}$ the mutual information expression simplifies to:

$$I(A;B) = 1 - \left(p \log \frac{1}{p} + q \log \frac{1}{q}\right) \tag{16}$$

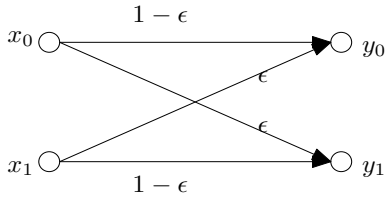
Since this represents the maximum possible mutual information we can then state:

$$C_{BSC} = 1 - \left(p \log \frac{1}{p} + q \log \frac{1}{q}\right) \tag{17}$$

How does the channel capacity vary for different error probabilities, q? The following observations can be made:

- When $q = 0$ or 1, that is, no error or 100% bit inversion, the BSC channel will provide its maximum capacity of 1 bit
- When $q = 0.5$ the BSC channel is totally ambiguous or useless and has a capacity of 0 bits

How to derive the channel capacity of the binary-symmetric channel (BSC) with crossover probability ϵ ?



Solution:
Denoting $p(x' = 0) = u$ and $p(x = 1) = 1 - u$, we obtain

$$\begin{aligned}
p(y) &= \sum_{x'=0}^{1} p(x')p(y|x') \\
&= p(x' = 0)p(y|x' = 0) + p(x' = 1)p(y|x' = 1) \\
&= \begin{cases} p(x' = 0)p(y = 0|x' = 0) + p(x' = 1)p(y = 0|x' = 1), & y = 0 \\ p(x' = 0)p(y = 1|x' = 0) + p(x' = 1)p(y = 1|x' = 1), & y = 1 \end{cases} \\
&= \begin{cases} u(1-\epsilon) + (1-u)\epsilon, & y = 0 \\ u\epsilon + (1-u)(1-\epsilon), & y = 1 \end{cases}
\end{aligned}$$

and

$$
\begin{aligned}
H(Y|X) &= \sum_{x=0}^{1}\sum_{y=0}^{1} p(x)p(y|x)\log_2\frac{1}{p(y|x)} \\
&\Big(= \sum_{x=0}^{1}\sum_{y=0}^{1} p(x)\cdot H(Y|X=x)\Big) \\
&= p(x=0)p(y=0|x=0)\log_2\frac{1}{p(y=0|x=0)} + p(x=0)p(y=1|x=0)\log_2\frac{1}{p(y=1|x=0)} \\
&+ p(x=1)p(y=0|x=1)\log_2\frac{1}{p(y=0|x=1)} + p(x=1)p(y=1|x=1)\log_2\frac{1}{p(y=1|x=1)} \\
&= u(1-\epsilon)\log_2\frac{1}{(1-\epsilon)} + u\epsilon\log_2\frac{1}{\epsilon} + (1-u)\epsilon\log_2\frac{1}{\epsilon} + (1-u)(1-\epsilon)\log_2\frac{1}{(1-\epsilon)} \\
&= \epsilon\log_2\frac{1}{\epsilon} + (1-\epsilon)\log_2\frac{1}{(1-\epsilon)} \\
&= H_b(\epsilon) \qquad \text{(This is called the binary entropy function.)}
\end{aligned}
$$

Hence,

$$
\begin{aligned}
I(X;Y) &= H(Y) - H(Y|X) \\
&= [u(1-\epsilon)+(1-u)\epsilon]\log_2\frac{1}{u(1-\epsilon)+(1-u)\epsilon} + [u\epsilon+(1-u)(1-\epsilon)]\log_2\frac{1}{u\epsilon+(1-u)(1-\epsilon)} - H_b(\epsilon),
\end{aligned}
$$

and

$$
\begin{aligned}
C &= \max_{p(x)=(u,1-u)} I(X;Y) \\
&= \max_{0\le u\le 1}\Big([u(1-\epsilon)+(1-u)\epsilon]\log_2\frac{1}{u(1-\epsilon)+(1-u)\epsilon} + [u\epsilon+(1-u)(1-\epsilon)]\log_2\frac{1}{u\epsilon+(1-u)(1-\epsilon)} - H_b(\epsilon)\Big) \\
&= -\min_{0\le u\le 1}\Big([u(1-\epsilon)+(1-u)\epsilon]\log_2[u(1-\epsilon)+(1-u)\epsilon] + [u\epsilon+(1-u)(1-\epsilon)]\log_2[u\epsilon+(1-u)(1-\epsilon)] + H_b(\epsilon)\Big).
\end{aligned}
$$

Taking the derivative of the term inside parentheses with respect to u yields:

$$
\begin{aligned}
C' &= (1-2\epsilon)\log_2[u(1-\epsilon)+(1-u)\epsilon] + [u(1-\epsilon)+(1-u)\epsilon]\frac{1-2\epsilon}{\log(2)[u(1-\epsilon)+(1-u)\epsilon]} \\
&+ (2\epsilon-1)\log_2[u\epsilon+(1-u)(1-\epsilon)] + [u\epsilon+(1-u)(1-\epsilon)]\frac{2\epsilon-1}{\log(2)[u\epsilon+(1-u)(1-\epsilon)]} \\
&= (1-2\epsilon)\Big(\log_2[u(1-\epsilon)+(1-u)\epsilon] - \log_2[u\epsilon+(1-u)(1-\epsilon)]\Big).
\end{aligned}
$$

As a result, the above derivative equals zero if, and only if,

$$
u(1-\epsilon)+(1-u)\epsilon = u\epsilon+(1-u)(1-\epsilon)
$$

which gives $u^* = 1/2$. Finally,

$$
\begin{aligned}
C &= -\Big([u^*(1-\epsilon)+(1-u^*)\epsilon]\log_2[u^*(1-\epsilon)+(1-u^*)\epsilon] + [u^*\epsilon+(1-u^*)(1-\epsilon)]\log_2[u^*\epsilon+(1-u^*)(1-\epsilon)] + H_b(\epsilon)\Big) \\
&= 1 - H_b(\epsilon)
\end{aligned}
$$

6 Information Capacity Theorem

In Section 5 we defined the channel capacity as the maximum of the mutual information over all possible input distributions. Of importance to communication engineers is the channel capacity of a band-limited, power-limited Gaussian channel. This is given by the following maximization problem:

$$
\begin{aligned}
C &= \max_{\{f_X(x):E[X^2]\le P\}} I(X;Y) \\
&= \max_{\{f_X(x):E[X^2]\le P\}} \{H(Y) - H(Y|X)\}
\end{aligned}
\tag{18}
$$

We now provide the details of deriving an important and well-known closed-form expression for C for Gaussian channels. The result is the *Information Capacity Theorem* which gives the capacity of a Gaussian communication channel in terms of the two main parameters that confront communication engineers when designing such systems: the signal-to-noise ratio and the available bandwidth of the system.

Claim 1. *If $Y = X + N$ and X is uncorrelated with N then:*

$$H(Y|X) = H(N) \tag{19}$$

Proof. We note from Bayes Theorem that $f_{XY}(x, y) = f_Y(y|x)f_X(x)$. Also since $Y = X + N$ and since X and N are uncorrelated we have that $f_Y(y|x) = f_N(y - x)$. Using these in the expression for $H(Y|X)$ gives:

$$\begin{aligned} H(Y|X) &= -\int_{XY} f_{XY}(x,y) \log f_Y(y|x) dy\ dx \\ &= -\int_X f_X(x) \Big\{ \int_Y f_Y(y|x) \log f_Y(y|x) dy \Big\} dx \\ &= -\int_X f_X(x) \Big\{ \int_N f_N(n) \log f_N(n) dn \Big\} dx \\ &= H(N) \int_X f_X(x) dx \\ &= H(N) \end{aligned} \tag{20}$$

□

For a Gaussian random variable the differential entropy attains the maximum value of $\log\left(\sqrt{2\pi e}\sigma\right)$; so for the Gaussian random variable, N, we know that:

$$H(N) = \log\left(\sqrt{2\pi e}\sigma\right) = \frac{1}{2}\log\left(2\pi e\sigma_N^2\right) \tag{21}$$

For random variable $Y = X + N$ where X and N are uncorrelated we have that:

$$H(Y) \leq \frac{1}{2}\log\left(2\pi e\sigma_Y^2\right) = \frac{1}{2}\log\left[2\pi e\Big(\sigma_X^2 + \sigma_N^2\Big)\right] \tag{22}$$

with the maximum achieved when Y is a Gaussian random variable. Thus:

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H(Y) - H(N) \\ &\leq \frac{1}{2}\log\left[2\pi e\Big(\sigma_X^2 + \sigma_N^2\Big)\right] - \frac{1}{2}\log\left(2\pi e\sigma_N^2\right) \\ &\leq \frac{1}{2}\log\Big(\frac{\sigma_X^2 + \sigma_N^2}{\sigma_N^2}\Big) \\ &= \frac{1}{2}\log\Big(1 + \frac{\sigma_X^2}{\sigma_N^2}\Big) \end{aligned} \tag{23}$$

If X is chosen to be a Gaussian random variable with $\sigma_X^2 = P$ then Y will also be a Gaussian random variable and the maximum mutual information or channel capacity will be achieved:

$$C = \frac{1}{2}\log\Big(1 + \frac{P}{\sigma_N^2}\Big) \quad \textit{bits per transmission} \tag{24}$$

Since the channel is also band-limited to W Hz then there can be no more than $2W$ symbols transmitted per second and $\sigma_N^2 = N_o W$. This provides the final form of the channel capacity, stated as Shannons most famous Information Capacity Theorem, which is also known as the Shannon-Hartley Law in recognition of the early work by Hartley.

Result 2. Information Capacity Theorem *The information capacity of a continuous channel of bandwidth W Hz, perturbed by $AWGN$ of power spectral density $N_o/2$ and bandlimited also to W Hz, is given by:*

$$C = W\log\Big(1 + \frac{P}{N_o W}\Big) \quad \textit{bits per second} \tag{25}$$

where P is the average transmitted power and $P/N_o W$ is the signal-to-noise ratio or SNR.

Equation 25 provides the theoretical capacity or upper bound on the bits per second that can be transmitted for error-free transmission through a channel for a given transmitted power, P, and channel bandwidth, W, in the presence of $AWGN$ noise with power spectral density, $N_o/2$. Thus the information capacity theorem defines a fundamental limit that confronts communication engineers on the rate for error-free transmission through a power-limited, band-limited Gaussian channel.

Example 4. *What is the minimum signal-to-noise ratio that is needed to support a* $56k$ *modem?*

Proof. A $56k$ modem requires a channel capacity of $56,000$ bits per second. We assume a telephone bandwidth of $W = 3600\ Hz$. From Equation 25 we have:

$$\begin{aligned} 56,000 &= 3600 \log(1 + SNR) \\ \Rightarrow SNR &= 48159 \end{aligned}$$

or

$$\begin{aligned} SNR_{dB} &= 10 \log_{10}(48159) \\ SNR_{dB} &= 47\ dB \end{aligned}$$

Thus a SNR of at least $48\ dB$ is required to support running a $56k$ modem. In real telephone channels other factors such as crosstalk, co-channel interference, and echoes also need to be taken into account. □

Suppose the Discrete Memoryless Source (DMS) has the source alphabet A and entropy $H(A)$ bits per source symbol. Let the source generate a symbol every T_s seconds. Then the average information rate of the source is $\frac{H(A)}{T_s}$ bits per second.Let us assume that the channel can be used once every T_c seconds and the capacity of the channel is C bits per channel use. Then the channel capacity per unit time is $\frac{C}{T_c}$ bits per second. We now state Shannon's second theorem known as the *Noisy Channel Coding Theorem* or simply the *Channel Coding Theorem.*

Theorem 2. Noisy Channel Coding Theorem

1. *Let a DMS with an alphabet A have entropy $H(A)$ and produce symbols every T_s seconds. Let, a discrete memoryless channel have a capacity C and be used once every T_c seconds. Then if*

$$\frac{H(A)}{T_s} \leq \frac{C}{T_c} \tag{26}$$

   there exists a coding scheme for which the source output can be transmitted over the noisy channel and be reconstructed with an arbitrarily low probability of error.

2. *Conversely if*

$$\frac{H(A)}{T_s} > \frac{C}{T_c} \tag{27}$$

   it is not possible to transmit information over the channel and reconstruct it with an arbitrarily small probability error.

The parameter $\frac{C}{T_c}$ is called the **Critical Rate**.