1. Discrete Memoryless Channels and Conditional Entropy

In order to study in detail the problem of channel encoding, we must start by defining precisely the mathematical model of a channel of communication.

Definition 1. *A **Discrete Memoryless Channel** consists of an **input alphabet** $I = \{x_0, \cdots, x_{m-1}\}$, an **output alphabet** $O = \{y_0, \cdots, y_{v-1}\}$, and a **channel matrix** $M = [p(y_j|x_i)]$ of **forward channel probabilities** satisfying*

$$\forall i, j \; p(y_j|x_i) \geq 0; \qquad \forall i \sum_j p(y_j|x_i) = 1.$$

The adjectives discrete and memoryless have the meaning that symbols are communicated through the channel one by one and that each symbol received depends only on the corresponding symbol that was sent. From now on these properties are implicitly assumed to hold and a Discrete Memoryless Channel will be named simply a channel.
The forward channel probability $p(y_j|x_i)$ is to be interpreted as the conditional probability of receiving y_j given that x_i was sent.

Example 2. *The fundamental example is the **Binary Symmetric Channel** with matrix*

$$\begin{bmatrix} 1-p & p \\ p & 1-p \end{bmatrix}.$$

In this case both input and output alphabets consist of two symbols and we may assume $I = O = \{0, 1\}$.

In most cases the input and output alphabet will be the same, e.g. a finite field, but the possibility of distinct alphabets should not be ruled out in coding theory. A simple example is motivated by the property of forward probabilities

$$\forall i \sum_j p(y_j|x_i) = 1;$$

this means that a sent symbol is not "lost" in the communication; however, it is desirable to consider the possibility not only of a symbol being changed into another but also of being erased or become illegible by the receiving device. For this reason, coding schemes may include an extra output symbol **?**.

Example 3. *The simplest example is the **Binary Erasure Channel** with matrix*

$$\begin{bmatrix} 1-\lambda-\mu & \mu & \lambda \\ \lambda & \mu & 1-\lambda-\mu \end{bmatrix}$$

where the input alphabet is $\{0, 1\}$ and the output alphabet is $\{0, ?, 1\}$. We have $p(1|0) = \lambda$ and $p(?|0) = \mu$, and similarly for $x = 1$.

Remark 4. *We may also add artificially, the new symbol ? to the input alphabet, fixing $p(?) = 0$ and $p(y|?) = 0$ for any $y \neq 0$. In the example above, we would have with input and output alphabets $\{0, ?, 1\}$, the matrix*

$$\begin{bmatrix} 1-\lambda-\mu & \mu & \lambda \\ 0 & 1 & 0 \\ \lambda & \mu & 1-\lambda-\mu \end{bmatrix}$$

Due to the type of codes we are discussing, we will always assume that the input and output alphabet are the same, either a finite set I or $I \cup \{?\}$. Later we will use finite sets with some extra structure.

Given a **input probability distribution** $p(x_i)$ in I, the channel determines an **output probability distribution** in O by

$$p(y_j) = \sum_i p(y_j|x_i)p(x_i).$$

$$p(x_i, y_j) = p(y_j|x_i)p(x_i)$$

is the **joint distribution** of the the input and output variables.

When the problem of decoding a message is considered, it is natural to ask for the **backward channel probabilities**: given that some y_j is received we would like to know the probability that some x_i was sent. These are defined, of course only in the case that $p(y_j) > 0$, by

$$p(x_i|y_j) = \frac{p(y_j|x_i)p(x_i)}{p(y_j)}.$$

Exercise 5. *Deduce the formula for the output probability distribution and for the backward conditional probabilities in the case of the Symmetric Binary Channel and of the Binary Channel with Erasure, described above.*
Confirm that adding the erasure symbol ? to the input does not change the output probabilities as well as the backward probabilities (with $p(x =?|y = y_j) = 0$, for any y_j).

Remark 6. *Both input and output data are more precisely modeled by random variables x and y and all probability distributions refer to values of these variables: $p(y_j|x_i) = p(y = y_j|x = x_i)$, and so on. But we will use whenever possible the simplified notation and keep the reference to Probability Theory to a minimum.*

1.1. **Conditional Entropy.** Entropy was used in the modeling of source encoding as a measure, given a probability distribution, of the uncertainty on the knowledge of symbols. Similarly,

Definition 7. *Given a channel and an input probability distribution, the **conditional entropy** of the input, given that $y = y_j$, is defined as*

$$H(x|y_j) = -\sum_i p(x_i|y_j) \log\left(p(x_i|y_j)\right);$$

*the **conditional entropy of** x **given** y is then*

$$H(x|y) = \sum_j H(x|y_j)p(y_j).$$

The definition depends on the choice of a basis for the logarithmic function; the most usual choice is $\log_m()$, where $m = |I|$.
The conditional entropy of the channel may be interpreted as the uncertainty on the knowledge of the value of x after observing y.
The conditional entropy of y given x is defined in a similar way:

$$H(y|x) = -\sum_{i,j} p(y_j|x_i)\log(p(y_j|x_i))p(x_i).$$

Example 8. *For a binary symmetric channel with crossover probability $p < 0.5$, if the input probability is uniform ($p(x = 0) = p(x = 1) = 0.5$), then the output probability is also uniform (a fact to be generalized later) and $H(x|y) = H(y|x) = H(p)$. For example, taking $p = 0.01$, we have*

$$H(x) = H(y) = 1, \qquad H(x|y) = H(y|x) \approx 0.08,$$

while with $p = 0.1$, $H(x|y) = H(y|x) \approx 0.469$.
If the input distribution is not uniform then, for the same channel, the results are different: suppose that $p = 0.1$ and that the input probability distribution is

$$p(x = 0) = 0.6, \qquad p(x = 1) = 0.4.$$

Then the output distribution is

$$p(y = 0) = 0.58, \qquad p(y = 1) = 0.42;$$

the forward conditional entropy is still $H(y|x) \approx 0.469$, but

$$H(x) \approx 0.971, \qquad H(y) \approx 0.9815 \text{ and } H(x|y) \approx 0.4587.$$

Notice how the channel increased the uncertainty on the value of the output compared to that of the input. Also, $H(y|x)$ does not depend on the probability distribution in the input.

A nonsymmetric binary channel may act differently: suppose the matrix of forward channel probabilities is

$$\begin{bmatrix} 0.9 & 0.1 \\ 0.4 & 0.6 \end{bmatrix}.$$

For the uniform probability distribution in the input, we obtain $H(y|x) \approx 0.72$ and

$$p(y = 0) = 0.65, \qquad p(y = 1) = 0.35, \qquad H(y) \approx 0.934.$$

The backward conditional probabilities are

$$p(0|0) \approx 0.692,\ p(0|1) \approx 0.142,\ p(1|0) \approx 0.307,\ p(1|1) \approx 0.857,$$

and

$$H(x|0) \approx 0.89, \qquad H(x|1) \approx 0.59, \qquad H(x|y) \approx 0.785.$$

If, on the other hand,

$$p(x = 0) = 0.6, \qquad p(x = 1) = 0.4,$$

with $H(x) \approx 0.971$, we have $H(y|x) \approx 0.709$; the output probabilities and entropy are

$$p(y = 0) = 0.70, \qquad p(y = 1) = 0.30, \qquad H(y) \approx 0.881.$$

The backward conditional probabilities are

$$p(0|0) \approx 0.771,\ p(0|1) \approx 0.2,\ p(1|0) \approx 0.23,\ p(1|1) \approx 0.8,$$

and

$$H(x|0) \approx 0.777, \qquad H(x|1) \approx 0.722, \qquad H(x|y) \approx 0.761.$$

Here is another example with a Binary Erasure Channel:

Example 9. *Let $I = \{0, 1\}$ and $O = \{0, ?, 1\}$. The input probabilities are*

$$p(0) = 1/4,\ p(1) = 3/4,$$

and the matrix of forward channel probabilities is

$$\begin{bmatrix} 1/2 & 1/2 & 0 \\ 0 & 1/3 & 2/3 \end{bmatrix}.$$

The output probability distribution is then

$$p(0) = 1/8,\ p(?) = 3/8,\ p(1) = 1/2,$$

and the backward probabilities $p(x_i|y_j)$ are

$$\begin{array}{lll} p(0|0) = 1 & p(0|?) = 1/3 & p(0|1) = 0 \\ p(1|0) = 0 & p(1|?) = 2/3 & p(1|1) = 1 \end{array}.$$

The interpretation of these values is useful for the understanding of the notion of conditional probability in coding: for example, the value $p(0|0) = 1$ is obvious from the observation of the forward probabilities,as $y = 0$ may be only (in probabilistic terms) the result of $x = 0$.
In this case $H(x) \approx 0.811$ and

$$H(x|0) = H(x|1) = 0, \qquad H(x|?) \approx 0.918,$$

and finally $H(x|y) \approx 0.344$.
It is interesting to observe that although, as we will confirm, $H(x|y) \leq H(x)$ always holds, for particular values of the output it may happen that $H(x|y_j) > H(x)$. In the example, the uncertainty on the value of X increases in consequence of ? being received.

The entropy of the joint distribution is $H(x, y) = -\sum_{i,j} p(x_i, y_j) \log(p(x_i, y_j))$. An application of the defining formulas gives

Proposition 10. *$H(x|y) = H(x, y) - H(y)$.*

Proof. (**HW**). □

Remark 11. *This proposition shows that conditional entropy may be defined for any random variables, independently from its relation with communication channels.*

We have also, as a simple consequence of Gibbs Lemma,

Lemma 12. *For any random variables x and y,*

$$H(x, y) \leq H(x) + H(y),$$

with equality if and only if x and y are independent.

Proof. (**HW**). □

Corollary 13. *$H(x|y) \leq H(x)$, with equality if and only if x and y are independent.*

The next definition identifies extreme cases:

Definition 14. *A channel is*

i) ***lossless*** *if for any j such that $p(y_j) > 0$ there exists i such that $p(x_i|y_j) = 1$; equivalently, $H(x|y) = 0$.*

ii) ***deterministic*** *if*
$$\forall i \, \exists j : p(y_j|x_i) = 1,$$
or equivalently $H(y|x) = 0$.
iii) ***noiseless*** *if it is both lossless and deterministic.*
iv) ***useless*** *if, for any input probability distribution* x *and* y *are independent random variables, i.e.,* $H(x|y) = H(x)$.

Exercise 15. *Verify the equivalences stated in the definition.*

Another useful definition is the following:

Definition 16. *A channel is* ***row symmetric*** *(respectively,* ***column symmetric****) if each row (respectively, column) is obtained by a permutation of the entries of the first row (respectively, column).*
The channel is said to be symmetric if it is both row and column symmetric.

Theorem 17. *For a row symmetric channel,* $H(y|x)$ *is independent of the input distribution.*

Proof. (**HW**). □

Theorem 18. *In a column symmetric channel, a uniform input distribution gives rise to a uniform output distribution.*

Proof. (**HW**). □

Exercise 19. *find a row-symmetric but not column-symmetric and a column-symmetric but not row-symmetric channel matrix.*

Remark 20. *The matrix of a symmetric channel is not, in general, a symmetric matrix.*

Given the information interpretation of conditional entropy, it is natural to relate it to the probability of error in the transmission, ie the probability that $x \neq y$. The next inequality does just this.

Proposition 21 (Fano's inequality)**.** *If* x *and* y *are random variables taking values in the same set* $\{x_1, \cdots, x_m\}$ *and* $p_e = p(x \neq y)$*, then*
$$H(x|y) \leq H(p_e) + p_e \log(m-1).$$

This will be an easy corollary from the next theorem. We start with the statement of a very useful inequality, a particular case of Jensen's inequality, which will be used frequently:

Lemma 22 (Jensen's inequality (particular case))**.** *Let* x *and* y *be random variables taking values in finite sets* $\{x_i :\leq i \leq m\}$ *and* $\{y_j : 1 \leq j \leq v\}$*, respectively,* $f(x, y)$ *a joint probability distribution and* $g(x, y)$ *a positive function. We have*
$$\sum_{i,j} f(x_i, y_j) \log(g(x_i, y_j)) \leq \log\left(\sum_{i,j} f(x_i, y_j) g(x_i, y_j)\right),$$
with equality if and only if there exist x_0, y_0 *such that* $f(x_0, y_0) = 1$.

Proof. This is a consequence of the convexity of $-\log$ (**HW**). □

Theorem 23. *Let x, y, z denote discrete random variables, taking values in finite sets, and $a(z) = \sum_{i,j} p(y_j)p(z_k|x_i, y_j)$. Then*

$$H(x|y) \leq H(z) + \sum_k p(z_k) \log(a(z_k)).$$

Proof. We have

$$H(x|y) = \sum_{i,j} p(x_i, y_j) \log\left(\frac{1}{p(x_i|y_j)}\right) = \sum_{i,j,k} p(x_i, y_j, z_k) \log\left(\frac{1}{p(x_i|y_j)}\right) =$$

$$= \sum_k p(z_k) \sum_{i,j} \frac{p(x_i, y_j, z_k)}{p(z_k)} \log\left(\frac{1}{p(x_i|y_j)}\right).$$

For each z_k, $\frac{p(x_i,y_j,z_k)}{p(z_k)}$ determines a probability distribution on x, y; applying Jensen's inequality, we get

$$H(x|y) \leq \sum_k p(z_k) \log\left(\frac{1}{p(z_k)} \sum_{i,j} \frac{p(x_i, y_j, z_k)}{p(x_i|y_j)}\right) =$$

$$= H(z) + \sum_k p(z_k) \log\left(\sum_{i,j} \frac{p(x_i, y_j, z_k)}{p(x_i|y_j)}\right).$$

We notice now that

$$\frac{p(x_i, y_j, z_k)}{p(x_i|y_j)} = p(y_j)p(z_k|x_i, y_j).$$

□

Exercise 24. *Check the details of the proof.*

The proof of Fano's inequality follows taking $z = 0$ if $x = y$ and $z = 1$ if $x \neq y$ (**HW**).

1.2. **Mutual Information and channel capacity.** From the previous definitions and their interpretation we may, following Shannon, define the measure of information that gets through the channel in the following way:

Definition 25. *The **mutual information** of a channel is*

$$I(x; y) = H(x) - H(x|y).$$

And the following result justifies the adjective mutual:

Proposition 26. *If the random variables x and y take values $\{x_1, \cdots, x_n\}$ and $\{y_1, \cdots, y_m\}$, respectively, then*

$$I(x; y) = \sum_{i,j} p(x_i, y_j) \log\left(\frac{p(x_i|y_j)}{p(x_i)}\right),$$

ie, $I(x; y)$ is the expected value, in the sample space of the pair (x, y), of the function $\log\left(\frac{p(x_i|y_j)}{p(x_i)}\right)$.

Proof. (**HW**). □

And we have the consequence

Corollary 27.

$$I(x;y) = I(y;x).$$

Remark 28. *We may also interpret $I(x;y)$ as a measure of the decrease of the uncertainty on the knowledge of an unknown x produced by the knowledge of y. Accordingly, for a lossless channel $I(x;y) = H(x)$, while for a useless channel $I(x;y) = 0$.*

The mutual information depends on the channel forward probabilities but also on the input probability distribution. We get finally to the crucial notion of capacity of a channel:

Definition 29. *The **capacity** of a channel is*

$$Cap = \max_{P(x)} I(x;y)$$

where the maximum is over all probability distributions $P(x)$ on the input.

Although the capacity of a channel may be hard to compute, in the special case of a symmetric channel we have an explicit formula:

Theorem 30. *The capacity of a symmetric channel is*

$$Cap = \log(v) + \sum_j p(y_j|x_i)\log(p(y_j|x_i))$$

for any i, where $v = |O|$. Moreover, this value is achieved as $I(x;y)$ for the uniform probability distribution on I.

Proof. We know that, as a consequence of symmetry, $H(y|x)$ is independent of the probability distribution $P(x)$ and

$$H(y|x) = -\sum_j p(y_j|x_i)\log(p(y_j|x_i))$$

for any i. So

$$C = \max_{P(x)} I(y;x) = \max_{P(x)} H(y) - H(y|x).$$

On the other hand, the maximum $\log(v)$ of $H(y)$ is achieved with the uniform distribution $p(y_j) = 1/v$ for all j. But by theorem 12 uniform distribution on the output is obtained if we have uniform distribution on the input. □

We end with some properties of mutual information that will be relevant to its interpretation and application in the context of communication channels. Their proofs are almost direct applications of Jensen's inequality. Notice that the same base for logarithms is being used everywhere.

Theorem 31. *If x, y, z are discrete random variables with values in finite sets,*

$$I((x,y);z) \geq I(y;z),$$

with equality if and only if $p(z_k|x_i, y_j) = p(z_k|y_j)$ for all values with $p(x_i, y_j, z_k) > 0$.

Proof. (**HW**): we have

$$I((x,y);z) = \sum_{x_i,y_j,z_k} p(x_i,y_j,z_k) \log\left(\frac{p(z_k|(x_i,y_j))}{p(z_k)}\right),$$

and

$$I(y;z) = \sum_{y_j,z_k} p(y_j,z_k) \log\left(\frac{p(z_k|y_j)}{p(z_k)}\right) = \sum_{x_i,y_j,z_k} p(x_i,y_j,z_k) \log\left(\frac{p(z_k|y_j)}{p(z_k)}\right).$$

Apply Jensen's inequality. □

Although the inequality is not surprising, given the interpretation of mutual information, the condition for equality is more important: it means that the sequence (x,y,z) is a **Markov chain**; informally, z depends on x only through the dependence of y on x. This implies (**HW**)

that the reversed sequence (z,y,x) is also a Markov chain and, consequently,

Corollary 32. *If (x,y,z) is a Markov chain, then*

$$I(x;z) \leq I(x;y), \qquad I(x;z) \leq I(y;z).$$

Proof. (**HW**).

□

We will come to this point later when we consider the coding and communication process.

The next two results state that the mutual information between two random variables has convexity properties:

Proposition 33. *Let the forward channel probabilities $p(y|x)$ be fixed and $p_1(x)$ and $p_2(x)$ be probability distributions on input random variables X_1 and X_2, respectively; let Y_1 and Y_2 be the corresponding output random variables. For $0 \leq t \leq 1$, let X be the input random variable with probability distribution $p(x) = tp_1(x) + (1-t)p_2(x)$ and Y the corresponding output. Then*

$$tI(X_1;Y_1) + (1-t)I(X_2;Y_2) \leq I(X;Y).$$

Proof. (**HW**). □

Proposition 34. *Let $p(x)$ be a fixed probability distribution on the input variable X, and $p_1(y|x)$ and $p_2(y|x)$ be two sets of forward channel probabilities; denote as Y_1 and Y_2 the corresponding output variables. Then, for $0 \leq t \leq 1$ the forward channel probabilities*

$$p(y|x) = tp_1(y|x) + (1-t)p_2(y|x)$$

with output Y satifies

$$I(X;Y) \leq tI(X;Y_1) + (1-t)I(X;Y_2).$$

Proof. (**HW**). □

Exercise 35. *Determine the capacity of the Binary Erasure Channel with matrix*

$$\begin{bmatrix} 1-\lambda-\mu & \mu & \lambda \\ 0 & 1 & 0 \\ \lambda & \mu & 1-\lambda-\mu \end{bmatrix}$$

and the optimal input probability distribution.

Exercise 36. *Let* $\{p_1, \cdots, p_m\}$ *be a probability distribution and* $p^* = \max_i\{p_i : 1 \leq i \leq m\}$. *Prove that*

a) $H_2(p_1, \cdots, p_m) \geq H_2(p^*)$;
b) $H_2(p_1, \cdots, p_m) \geq -\log_2(p^*)$;
c) $H_2(p_1, \cdots, p_m) \geq 2(1-p^*)$.