1. Channels, Block Codes and Shannon's Theorem

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. **Channels and Information.** The notions of Conditional Entropy and Mutual Information find a natural application in this context.

Example 7. *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.

Here is another example with a Binary Erasure Channel:

Example 8. *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 $H(x|y) \leq H(x)$ 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 next definition identifies extreme cases:

Definition 9. *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 10. *Verify the equivalences stated in the definition.*

Another useful definition is the following:

Definition 11. *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 12. *For a row symmetric channel, $H(y|x)$ is independent of the input distribution.*

Proof. (**HW**). □

Theorem 13. *In a column symmetric channel, a uniform input distribution gives rise to a uniform output distribution.*

Proof. (**HW**). □

Exercise 14. *find a row-symmetric but not column-symmetric and a column-symmetric but not row-symmetric channel matrix.*

Remark 15. *The matrix of a symmetric channel is not, in general, a symmetric matrix.*

Following a previous definition, we have

Definition 16. *The **mutual information** of a channel is*

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

$I(x;y)$ measures 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 17. *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 18. *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. □

Exercise 19. *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.

1.2. **Block codes.**

Definition 20. *Given two strings x and y of the same length m over an alphabet A, the **distance** $d(x, y)$ is defined to be the number of coordinates where they differ. More precisely,*

$$\text{dist} : A \times A \to \mathbb{Z}, \qquad \text{dist}(a, b) = \begin{cases} 0 & \text{if } a = b \\ 1 & \text{if } a \neq b \end{cases}$$

is extended to sequences of length m, $x\{= x_1 \cdots x_m\}$ and $y = \{y_1 \cdots y_m\}$ as $\text{dist}(x, y) = \sum_{i=1}^{m} d(x_i, y_i)$.

It is clear that dist defines a distance in A^m for any $m \in \mathbb{N}$. This distance, called Hamming distance, is essential for what follows.

Exercise 21. *Prove that Hamming distance is translation invariant.*

Definition 22. *A q-block code C of **length** n is a subset of A^n. $M = |C|$, the number of codewords, is the **size** of the code.*
*The **information rate** of C is defined to be $\frac{\log_q(M)}{n}$.*
*The **distance** of the code is defined to be $d(C) = \min\{\text{dist}(c, c') : c, c' \in C; c \neq c'\}$.*

The channel encoding problem consists basically in the following: when a codeword $c \in C$ is sent, a word x is received; we will use the notation $c \rightsquigarrow x$. x may differ from c because of interferences in the transmission. We want to construct codes that allow efficient transmission, that we identify with high information rates, and capability of error detection and/or correction:

Definition 23. *C is u-error detecting if it is possible to identify the existence of, at least, u errors in the transmission of any codeword.*
C is t-error correcting if it is possible to correct at least t errors in the transmission.
C is said to be exactly u-error detecting if it is u-error detecting but not $u+1$-error detecting. A similar definition applies to error correcting.

Implicit in these definitions lies a concept of decoding:

Definition 24. ***Minimal Distance Decoding** consists in decoding each received x as the codeword c that minimizes $\text{dist}(x, c)$. In case there is more than one codeword that minimizes that distance, there is the option of not decoding (**incomplete decoding**) or of choosing one of the codewords (**complete decoding**).*

Example 25 (Repetition Codes)**.** *A simple example of a code is given by encoding each symbol $a \in \{0,1\}$ as the "constant" word $a \cdots a$ of length $2r+1$, for some $r \in \mathbb{N}$. The information rate is clearly $\frac{1}{2r+1}$, while the distance is $2r+1$ so this repetition code corrects r errors.*
In this case there are only two codewords and Minimal Distance Decoding consists simply in choosing the symbol that occurs in the majority of the entries (that is why the length is odd).
Of course, repetition codes may also be defined for larger alphabets A.

Example 26. *A second example, that will be much explored later, is the following: we consider again the alphabet $\{0,1\}$, but we identify it with the finite field $\mathbb{F}_2 = \mathbb{Z}/2\mathbb{Z}$; the message to be sent is decomposed in strings of length 4 and (x_1, x_2, x_3, x_4) is encoded as $(c_i)_{\in}\mathbb{F}_2^7$ as*

$$\begin{cases} c_i = x_i \, \forall \, 1 \leq i \leq 4 \\ c_5 = c_2 + c_3 + c_4 \\ c_6 = c_1 + c_3 + c_4 \\ c_7 = c_1 + c_2 + c_4 \end{cases}$$

In particular, this code has information rate $\frac{4}{7}$.
If a vector $(u_i) \in \mathbb{F}_2^7$ is received, the decoding procedure consists in computing

$$\begin{cases} s_1 = u_5 - (u_2 + u_3 + u_4) \\ s_2 = u_6 - (u_1 + u_3 + u_4) \\ s_3 = u_7 - (u_1 + u_2 + u_4) \end{cases}$$

It is clear that the vector (s_1, s_2, s_3) is the zero vector if and only if (u_i) is a codeword. Moreover, if it is not and we assume that an error occured in transmission in a single coordinate, we are able to correct it because if

$$u_i = c_i + 1, \qquad u_j = c_j \, \forall \, j \neq i,$$

each possible nonzero vector (s_1, s_2, s_3) is obtained for each value of $1 \leq i \leq 7$.
In other words, for each vector (u_i) there is a unique codeword at minimal distance from it and that distance is 1 unless, of course, if (u_i) is a codeword.

Let's consider the situation where one of these codes is used for transmission through a Binary Symmetric Channel with cross-over probability of error p:

If the repetition code described above is used with this channel, the decoding delivers a wrong codeword if and only if more than r coordinates are changed. So the probability of decoding error is

$$P_{de} = \sum_{k=r+1}^{2r+1} \binom{2r+1}{k} p^k (1-p)^{2r+1-k} = \binom{2r+1}{r+1} p^{r+1} + \text{ terms with higher powers of } p.$$

On the other hand, if the code described in the second example is used, the decoding procedure delivers a wrong codeword if and only if more than 1 coordinate

is changed, since the codeword chosen is at distance 0 or 1 from the received vector. So

$$P_{de} = \sum_{k=2}^{7} \binom{7}{k} p^k (1-p)^{7-k} = \binom{7}{2} p^2 + \text{ terms with higher powers of } p.$$

We now confirm that error detecting/correcting capability is closely related to the distance of the code:

Proposition 27. *If a block code C with minimal distance d is used and minimal distance decoding is applied, then*

a) *C is u-error detecting if and only if $d > u$;*
b) *C is t-error correcting if and only if $d \geq 2t + 1$.*

Proof. (**HW**). □

Remark 28. *It is important to clarify that error detection capability refers to detection of the existence of errors and not to the identification of the errors. In fact, if the codeword c is sent and the received vector x satisfies* $\operatorname{dist}(x, c) = u < d$, *then we know that x is not a codeword and so detect the presence of an error; but minimal distance decoding may lead to correct x to a different codeword c' and so, in a certain sense, the true error was not detected. On the other hand, if* $\operatorname{dist}(x, c) \geq d$, *the received vector may be a codeword and in that case the existence of errors is not even detected.*

In other words, if we use the code for error detection only (ie, we merely identify the existence of errors in transmission) then we are sure to detect all cases where less than d errors occur; if we use it for error correction we are sure to correct all cases where less than $\frac{d-1}{2}$ errors occur.
*In certain cases, we may use the code to simultaneously detect and correct errors: suppose that the minimal distance of the code is even, say $d = 2t + 2$, and that x is received. If there exists a unique codeword c at minimal distance from x we decide to correct x to c, otherwise we declare the existence of errors; in this way, if t (or less) errors occur in the transmission, we decode correctly, while if $t+1$ errors occur we still detect correctly. Notice that (**HW**) if d is odd this is no longer true.*

1.3. **Channels, Vector random variables and Codes.** We start the discussion of the application of the model of a channel, discussed earlier, and the notions of conditional entropy, mutual information, etc., to the process of encoding, transmission and decoding of messages using a block code.
We deduce a relation between the information rate of a code, the capacity of the channel and the probability of error in that process. This result is superseded by the main result, Shannon's theorem.

We consider a block code of length n and size M as a subset C of $\mathbb{F}_q^n$ with $|C| = M$. The notions of input, output and forward and backward conditional probabilities are generalized in the natural way:

> if $y = (y_1, \cdots, y_n) \in O^n$, $p(y|c) = \prod_i p(y_i|c_i)$;
> the output probability distribution is given by $p(y) = \sum_{c \in C} p(y|c)p(c)$, depending on a probability distribution on the code C;

and similarly for the backward probabilities and joint probabilities. As for individual symbols, codewords and output strings are values of vector random variables $X = (X_1, \cdots, X_n)$ and $Y = (Y_1, \cdots, Y_n)$.
In our typical application codewords $c \in \mathbb{F}_q^n$ are used to encode messages $u \in \mathbb{F}_q^k$. If $p()$ is a probability distribution on the alphabet $\mathbb{F}_q$, the coordinates u_i of a message u are values of independent random variables U_i, so we have a probability distribution on $\mathbb{F}_q^k$ given by $p(u) = \prod_{i=1}^k p(u_i)$ and we may define the probability distribution on the code by $p(c) = p(u)$.
In this overview, and in the applied examples discussed later, we will assume always that the probability distribution on the alphabet is uniform; as a consequence, the probability distribution on C is also uniform: $p(c) = M^{-1}$, for all $c \in C$.

Similarly, the final decoding of the output y will be a $v \in \mathbb{F}_q^k$, which is the value of another random variable $V = (V_1, \cdots, V_k)$.

Exercise 29. *Show that the above formulas for $p(y|c)$ and $p(y)$ define forward conditional probabilities and a probability distribution on the output strings.*

The definitions and properties of entropy, conditional entropy and mutual information generalise directly to the case of vector random variables.
However, for this generalisation to be coherent we must use the base M for the logarithm. This will be particularly relevant when the entropy and mutual information related to different codes are compared.

In particular, we get a version of Fano's inequality. This is relevant because, as we'll confirm later, the exact computation of the conditional entropy associated with a code may be difficult, depending on a detailed knowledge of its structure.

Proposition 30 (Fano's inequality)**.** *If X and Y are the random variables associated, respectively, to the input and output of the transmission of codewords from a size M code, and $p_e = p(X \neq Y)$, then*

$$H(X|Y) \leq H(p_e) + p_e \log(M-1).$$

We have also that the sequence of random variables

$$U \longrightarrow X \longrightarrow Y \longrightarrow V$$

corresponding to the three steps coding-transmission-decoding, form a Markov chain, and we may conclude (**HW**) that $I(U;V) \leq I(X;V) \leq I(X;Y)$.

On the other hand, the first and last of these mutual informations may be compared to the ones of their coordinates.

Proposition 31. *Let $U = (U_1, \cdots, U_k)$ and $V = (V_1, \cdots, V_k)$ be random vectors, with the U_i (resp. the V_i) taking values in the same set I (resp. O), $|I| = q$, such that the components U_i are independent. Then*

$$\frac{1}{\log_q(M)} \sum_i I(U_i; V_i) \leq I(U; V).$$

Proof. To simplify the notation, we will denote the values of the U_i by x and the ones of the V_i by y. In the same way the values of U and V will be denoted u and v. Notice that the probability distributions of the U_i (as those of the V_i) may be distinct. We will denote $p(U_i = x)$ by $p_i(x)$, $p(V_i = y|U_i = x)$ by $p_i(y|x)$, and so on.
First, according to the definition of mutual information,

$$\sum_i I(U_i; V_i) = \sum_i \sum_{x,y} p_i(x, y) \log_q \left(\frac{p_i(x|y)}{p_i(x)} \right).$$

In the same way,

$$I(U; V) = \sum_{u,v} p(u, v) \log_M \left(\frac{p(u|v)}{p(u)} \right) = \frac{1}{\log_q(M)} \sum_{u,v} p(u, v) \log_q \left(\frac{p(u|v)}{p(u)} \right),$$

and the hypothesis of independence of the U_i implies that, for $u = (u_1, \cdots, u_k)$,

$$p(u) = \prod_i p_i(u_i) = \prod_i p(U_i = u_i).$$

To compare the two quantities, we view the U_i and V_i, and so also the functions $f_i(x, y) = \log_q \left(\frac{p_i(x|y)}{p_i(x)} \right)$ as being defined in the same sample space of $U \times V$, ie, $I^k \times O^k = (I \times O)^k$: if $u = (u_1, \cdots, u_q) \in I^k$ and $v = (v_1, \cdots, v_k) \in O^k$, $f_i(u, v) = f_i(u_i, v_i)$.
Because mutual information is an expected value, we compute now the expected value, over $I^k \times O^k$ of $\sum_i f_i(u, v)$:

$$\sum_{u,v} p(u, v) \sum_i \log_q \left(\frac{p_i(x|y)}{p_i(x)} \right) = \sum_i \sum_{u,v} p(u, v) \log_q \left(\frac{p_i(x|y)}{p_i(x)} \right) =$$

$$= \sum_i \sum_{x,y} \sum_{\substack{u:u_i=x \\ v:v_i=y}} p(u, v) \log_q \left(\frac{p_i(x|y)}{p_i(x)} \right) = \sum_i \sum_{x,y} p_i(x, y) \log_q \left(\frac{p_i(x|y)}{p_i(x)} \right) = \sum_i I(U_i; V_i).$$

So,

$$\sum_i I(U_i; V_i) = \sum_{u,v} p(u, v) \sum_i \log_q \left(\frac{p_i(x|y)}{p_i(x)} \right) =$$

$$= \sum_{u,v} p(u, v) \log_q \left(\prod_i \frac{p_i(x|y)}{p_i(x)} \right).$$

To make the notation clear, for each $u = (u_1, \cdots, u_k)$ and $v = (v_1, \cdots, v_k)$,

$$\left(\prod_i \frac{p_i(x|y)}{p_i(x)} \right) = \left(\prod_i \frac{p(U_i = u_i|V_i = v_i)}{p(U_i = u_i)} \right).$$

But then

$$\sum_i I(U_i;V_i) - \log_q(M) I(U;V) = \sum_{u,v} p(u,v) \log_q \left(\prod_i \frac{p_i(x|y)}{p_i(x)} \right) - \sum_{u,v} p(u,v) \log_q \left(\frac{p(u|v)}{\prod_i p_i(x)} \right) =$$

$$= \sum_{u,v} p(u,v) \log \left(\frac{\prod_i p_i(x|y)}{p(u|v)} \right) \leq \log_q \left(\sum_{u,v} \frac{p(u,v)}{p(u|v)} \prod_i p_i(x|y) \right),$$

by Jensen's inequality. As $\frac{p(u,v)}{p(u|v)} = p(v)$, we get

$$\sum_{u,v} p(v) \prod_i p_i(x|y) = \sum_v p(v) \sum_u \prod_i p_i(x|y);$$

but for a fixed v,

$$\sum_u \prod_i p_i(x|y) = \sum_{u_1} \sum_{u_2} \cdots \sum_{u_k} \prod_i p_i(x|y) = \sum_{u_1} p_1(x|y) \sum_{u_2} p_2(x|y) \cdots \sum_{u_k} p_k(x|y) = 1;$$

therefore

$$\sum_v p(v) \sum_u \prod_i p_i(x|y) = \sum_v p(v) = 1,$$

and

$$\sum_i I(U_i;V_i) - \log_q(M) I(U;V) \leq 0.$$

□

Proposition 32. *Let $X = (X_1, \cdots, X_n)$ and $Y = (Y_1, \cdots, Y_n)$ be the input and output of a memoryless channel (ie, $p((y_1, \cdots, y_n)|(x_1, \cdots, x_n)) = \prod_{i=1}^n p(y_i|x_i)$). Then*

$$I(X;Y) \leq \frac{1}{\log_q(M)} \sum_{i=1}^n I(X_i;Y_i).$$

Proof. (**HW**). □

Let's consider now the consequences of these results to our coding-transmission-decoding process sketched above. We will restrict ourselves to the simplest non-trivial case: we'll assume the channel is binary, and that the source messages $u = (u_1, \cdots, u_k)$ are the values of a random vector $U = (U_1, \cdots, U_k)$ where the U_i are independent and uniformly distributed. The reasoning applies, essentially, in the general case.

Suppose we want to guarantee that this process satisfies $p_e = p(U_i \neq V_i) < \varepsilon$ for some small ε.

We know that

$$I(U;V) \geq \frac{1}{\log_q(M)} \sum_i I(U_i;V_i) = \frac{1}{\log_q(M)} \sum_i (H(U_i) - H(U_i|V_i)),$$

and Fano's inequality gives us

$$H(U_i|V_i) \leq H(p_e) + p_e \log(q-1) \leq H(\varepsilon).$$

Since the distribution of the U_i is uniform, we get

$$I(U;V) \geq \frac{k}{\log_q(M)} (1 - H(\varepsilon)).$$

On the other hand, $I(X;Y) \leq \frac{1}{\log_q(M)} \sum_i I(X_i;Y_i) \leq \frac{n}{\log_q(M)} Cap$, where Cap denotes the channel's capacity. We arrive at

Proposition 33. *Suppose that a binary code with information rate k/n is used for transmission through a channel with capacity Cap, and that the probability of decoding error is bounded above by ε. THen*

$$k(1 - H(\varepsilon)) \leq nCap \Leftrightarrow \frac{k}{n} \leq \frac{Cap}{1 - H(\varepsilon)}.$$

This shows, roughly speaking, that if our code has rate larger than the channel capacity, the decoding error is bounded below away from zero. On the other hand, a fixed small ε implies an upper bound on the rate of the code. In another section, we'll see how Shannon's theorem answers the question of how close can we get to that upper bound.
Before that, we are going to study the problem of bounding the probability of error from a different point of view.

2. Probability of error, Ideal observers, and Maximum Likelywood Decision

We want to understand the properties and relation with probability of error of a general decoding procedure.
Considering the decoding procedure, it is necessary to admit the possibility that some received strings can not be decoded, according to the decoding criteria used. We formalize that possibility by adding a new codeword:

A **decision scheme** (or **decoding scheme**) is a function $f : \mathbb{F}_q^n \to C \cup \{*\}$. $f(u) = *$ occurs if the output u is not decoded, and corresponds in practice to ask instead for a retransmission or simply report an error. The definition of backward conditional probabilities is generalized putting $p(*|u) = 0$ for any $u \in \mathbb{F}_q^n$.
The function f determines a partition of $\mathbb{F}_q^n$ in sets

$$B_c = \{u \in \mathbb{F}_q^n : f(u) = c\}$$

together, eventually, with the set $B_* = \{u \in \mathbb{F}_q^n : f(u) = *\}$ of undecodable strings.

If c is sent, u is received and $f(u) \neq c$ we have a **decision error**. The probability of a decision error, given that c is sent, is

$$p(error|c) = \sum_{u \notin B_c} p(u|c).$$

Averaging over all codewords, we have

$$p_e = \sum_{c \in C} p(error|c)p(c) = \sum_{c \in C} \sum_{u \notin B_c} p(u|c)p(c).$$

This depends on the input distribution as well as on the decision scheme.
In principle, a good decision scheme is one that minimizes p_e. In order to identify more clearly how to achieve this, we take the point of view of the decoder and rewrite the error probability in terms of the output: given that u is received, a correct decision happens if $f(u) = c$, so

$$p(error|u) = 1 - p(f(u)|u);$$

averaging over all u

$$p_e = \sum_{u \in \mathbb{F}_q^n} p(error|u)p(u) = 1 - \sum_{u \in \mathbb{F}_q^n} p(f(u)|u)p(u)$$

and this is minimized by maximizing the sum on the right. But the factors $p(u) = \sum_{c \in C} p(u|c)p(c)$ do not depend on the decision scheme. So the choice is to maximize $p(f(u)|u)$ for each u.

Definition 34. *A decision scheme f such that*

$$\forall u \, p(f(u)|u) = \max_{c \in C} p(c|u)$$

*is called an **ideal observer**.*

So an ideal observer chooses for each output string u the codeword most likely to have been sent, given that u was received.

The definition of an ideal observer depends not only on the channel forward probabilities but also on the input probability distribution. This dependence may be avoided by choosing not to minimize the average probability of error but the maximum probability of error

$$p_e^{max} = \max_{c \in C} p(error|c).$$

This has the advantage of not depending on the input probability distribution and giving a uniform bound on the error probability for any distribution. Unfortunately, no general method to choose decision schemes that minimize p_e^{max} is known.
Another way to avoid that dependence is suggested by the observation that, if we fix the uniform probability distribution $p(c) = \frac{1}{M}$ for every $c \in C$, then

$$p_e = \frac{1}{M} \sum_{c \in C} p(error|c);$$

This last expression is sometimes called the uniform probability of error and denoted by p_e^u. But for the uniform probability distribution

$$p(c|u) = \frac{p(u|c)p(c)}{p(u)} = \frac{1}{Mp(u)} p(u|c),$$

and so

$$\max_{c \in C} p(c|u) = \max_{c \in C} \frac{1}{Mp(u)} p(u|c) =$$

$$= \max_{c \in C} \left(\frac{1}{\sum_{c \in C} p(u|c)} p(u|c) \right) = \frac{1}{\sum_{c \in C} p(u|c)} \max_{c \in C} p(u|c).$$

We define, regardless of the input probability distribution,

Definition 35. *f is a **Maximum Likelihood Decision scheme**, or a MLD scheme, if it satisfies*

$$\forall u \, p(u|f(u)) = \max_{c \in C} p(u|c).$$

So in a Maximum Likelihood Decision scheme, for each u, $f(u)$ is the codeword c such that u is most likely of being received given that c is sent.

Remark 36. *If $p(c) = \frac{1}{M}$ for every $c \in C$, a MLD scheme is the same thing as an ideal observer.*

We relate now this decision schemes with the decoding by minimal distance. Recall that the Hamming distance $\text{dist}(x, y)$ between two vectors $x, y \in \mathbb{F}_q^n$ is equal to the number of coordinates where the two vectors are different.

Definition 37. *A discrete memoryless channel is* ***strongly symmetric*** *if the forward probabilities satisfy*

$$p(y_j|x_i) = \begin{cases} 1-\rho & \text{if } y_j = x_i \\ \frac{\rho}{q-1} & \text{if } y_j \neq x_i \end{cases}$$

for some $0 \leq \rho < 1/2$.

Suppose that we have a strongly symmetric channel . Then

$$p(u|c) = \prod_{i=1}^{n} p(u_i|c_i) = \left(\frac{\rho}{q-1}\right)^{\text{dist}(u,c)} (1-\rho)^{n-\text{dist}(u,c)} = (1-\rho)^n \left(\frac{\rho}{(1-\rho)(q-1)}\right)^{\text{dist}(u,c)}$$

is maximized by minimizing $\text{dist}(u, c)$.
So, under these conditions, a MLD scheme is equivalent to minimal distance decoding.

However, this equivalence does not hold for other channels, even symmetric. The computational details of the following example are left as an exercise; we use the concepts of Linear Codes, presented in another set of notes.

Example 38. *Consider the linear code over* $\mathbb{F}_3$ *with generator*

$$G = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 2 \\ 0 & 1 & 0 & 2 & 1 & 0 \\ 0 & 0 & 1 & 0 & 2 & 1 \end{bmatrix}.$$

Let $u = (0, 0, 0, 1, 1, 1)$ *be a received word. It has syndrome* $(1, 1, 1)$. *The error patterns in the corresponding coset with minimal weight are*

$$e_1 = (1, 2, 0, 0, 0, 0),\ e_2 = (2, 0, 1, 0, 0, 0),\ e_3 = (0, 1, 2, 0, 0, 0),$$

corresponding, respectively, to decoding into the codewords

$$c_1 = (2, 1, 0, 1, 1, 1),\ c_2 = (1, 0, 2, 1, 1, 1),\ c_3 = (0, 2, 1, 1, 1, 1).$$

Suppose the channel has matrix of forward probabilities

$$\begin{bmatrix} 3/5 & 3/10 & 1/10 \\ 1/10 & 3/5 & 3/10 \\ 3/10 & 1/10 & 3/5 \end{bmatrix};$$

Then

$$p(u|c_i) = \frac{1}{6}\frac{3}{10}\left(\frac{3}{5}\right)^4$$

but

$$p(u, 0) = \left(\frac{3}{10}\right)^3 \left(\frac{3}{5}\right)^3$$

which is larger.
So, in this case, the MLD scheme does not correspond to minimal distance decoding.

2.0.1. *Probability of error and Code parameters.* We discuss now the relation between code parameters and estimates on the probability of decoding error:
A (n, M, d) code over $\mathbb{F}_q$ (ie, a code with length n, size M and minimal distance at least d) is said to be **optimal** if it is not contained in a $(n, M+1, d)$ code.

Suppose that the following hypothesis are satisfied:

i) The input probability distribution is uniform;
ii) The channel is strongly symmetric, with $p(a|a) = 1 - \rho$ for every $a \in \mathbb{F}_q$, $\rho < 0.5$.

As it was seen above, under these conditions, Maximum Likelihood Decision coincides with Minimal Distance Decoding.
We then have

Proposition 39. *If C is an optimal (n, M, d) code over $\mathbb{F}_q$, the probability of decision error, under MLD, satisfies*

$$\sum_{j=d}^{n} \binom{n}{j} \rho^j (1-\rho)^{n-j} \leq p_e \leq 1 - \sum_{j=0}^{t} \binom{n}{j} \rho^j (1-\rho)^{n-j}$$

where $t = \lfloor \frac{d-1}{2} \rfloor$.

Proof. The first inequality follows from the observation that, if c is sent and u received and $\mathrm{dist}(u, c) \geq d$, then there exists some $c' \neq c$ such that $\mathrm{dist}(u, c') < \mathrm{dist}(u, c)$ and so u is incorrectly decoded; the second inequality is a consequence of C being t-error correcting. The details are left as an exercise (**HW**). □

The following theorem follows from the first inequality, under the same hypothesis on the input and channel probabilities. We omit the proof, which depends on estimates on binomial coefficients that follow essentially from Stirling's formula.

Theorem 40. *Let C_n be a family of (n, M_n, d_n) codes. If, for some $s < \rho$, and all sufficiently large n*

$$\frac{d_n - 1}{n} < s$$

then the probability of decoding error of C_n approaches 1 as $n \to +\infty$.

2.1. **Shannon's Noisy Channel Theorem.** Shannon's second theorem, also called the Noisy Channel Theorem, tells us that, as long as the information rate is kept below the channel's capacity, the probability of error may be made arbitrarily small:

Theorem 41 (Shannon)**.** *Consider a discrete memoryless channel with capacity Cap. For any $R < Cap$ there exists a sequence C_n of q-ary codes with decision schemes f_n such that*

i) *C_n is a (n, M) code with $M \geq \lceil q^{nR} \rceil$);*
ii) *$p_e^{\max}(n)$, the maximum probability of error of C_n approaches 0 as $n \to +\infty$.*

Proof. We present only a sketch of a proof: we fix a large n (to be specified later), and define

$$\Omega = \{(x, y) \in \mathbb{F}_q^n \times \mathbb{F}_q^n\}$$

to be the pairs of possible inputs and outputs of the channel. This becomes a probability space defining $p(x)$ to be the product of the probabilities of the coordinates,

for a fixed probability distribution on $\mathbb{F}_q$; $p(y|x)$ is defined in a similar way and $p(y) = \sum_{x \in \mathbb{F}_q^n} p(y|x)p(x)$.
Let R' satisfy $R < R' < Cap$, and consider the subset

$$T = \{(x, y) \in \Omega : \log_2 \left(\frac{p(y|x)}{p(y)} \right) \geq nR'\}.$$

Suppose now that $C \subset \mathbb{F}_q^n$ is a code with size M, also to be chosen later; we choose as decoding scheme the following: for each output y, if

$$S(y) = \{x : (x, y) \in T\};$$

contains exactly one codeword c we put $f(y) = c$; otherwise, we put $f(y) = *$.
If we denote, for $c \in C$, $P_e(c)$ to be the probability that a decoding error occurs when c is transmitted, we may give the following rough estimate: denoting

$$\Lambda(x, y) = \begin{cases} 1 & \text{if } (x, y) \in T \\ 0 & \text{if } (x, y) \notin T \end{cases}$$

$$P_e(c) \leq \sum_y (1 - \Lambda(c, y))p(y|c) + \sum_{x \in C \setminus c} \sum_y \Lambda(x, y)p(y|x) = Q_c.$$

Notice that Q_c is in fact a function on C, which is virtually impossible to compute or even estimate for large or complicated codes. The approach is then to estimate its average over all possible (n, M) codes. For this, we turn the space of these codes into a probability space, putting $p(C) = \prod_{c \in C} p(c)$. This corresponds to the informal idea of randomly choosing the codewords.
The estimates on the expected values of the summands of Q_c (seen as random variables on the space of all codes) is the most technical point in the proof and we omit all the details. It turns out that the expected value of the first summand above is

$$p((x, y) \notin T) = p(\log_2 \left(\frac{p(y|x)}{p(y)} \right) < nR'),$$

and that it follows from the weak law of large numbers that this approaches zero as $n \to +\infty$.
On the other hand, for each $x \in C$, the corresponding term in the second summand has expected value bounded above by $2^{-nR'}$. So the expected value for the summand is bounded above by $M2^{-nR'}$.
This is the point where we choose $M = 2^{1+\lceil nR \rceil}$ implying that $M2^{-nR'}$ can be made arbitrarly small, by choosing n sufficiently large. Putting all together, we may claim that, given ε, we have that the expected value of Q_c is, for sufficiently large n, bounded above by $\varepsilon/2$.
The last step is to define a global error function

$$P_e(C) = \frac{1}{M} \sum_{c_i} P_e(c_i),$$

where each summand is already a function of all the codewords in C. The estimates above imply that the expected value of the random variable P_e is (always for large n) bounded by $\varepsilon/2$, and so there must exist a code C with size M such that $P_e(C) < \varepsilon/2$. This code may not satisfy the conditions of the theorem, because it my contain codewords c for which $p_e(c) > \varepsilon$. But this may occur at most for half of the codewords. Discarding these we obtain the desired code. □

This proof is difficult at some points (the ones omitted above) but its version for the Binary Symmetric Channel, and uniform probability distribution on $\mathbb{F}_2$, may be a good exercise to grasp its fundamental ideas.
However, the crucial observation is that the proof relies on a nonconstructive existence argument. To this day, no family of codes with the above properties is known. And it should also be noticed that, from a practical point of view, the codes in a family fulfilling those conditions (for some R arbitrarily close to the capacity of the channel) may be too long or have too complicated decision schemes to make them useful for encoding.

Shannon's Theorem has also converse statements, which we summarise in the next theorem:

Theorem 42. *Consider a discrete memoryless channel with capacity C. Let C_n be a sequence of q-ary $(n, \lceil q^{nR} \rceil)$ codes and corresponding decision schemes f_n with uniform probability of decision error $p_e^u(n)$.*
If $R > C$ then

i) *there exists $\delta > 0$ such that $p_e^u(n) > \delta$ for all n;*
ii) $\lim_n p_e^u(n) = 1$.

Since it is very difficult to obtain explicit codes satisfying the conditions of Shannon's Theorem, the next best thing to ask for is a family of codes such that neither the size nor the distance become too small, compared to the length:

Definition 43. *A family C_n of codes is **asymptotically good** if it contains a subset C_{n_i} with parameters $[n_i, k_i, d_i]$ satisfying:*

i) $\lim_{i\to+\infty} n_i = +\infty$;
ii) $\liminf_{i\to+\infty} \frac{k_i}{n_i} > 0$;
iii) $\liminf_{i\to+\infty} \frac{d_i}{n_i} > 0$.

*A family is **asymptotically bad** if it does not contain such a subfamily.*

2.2. **Supplementary Results and Problems.**

Problem 44. *Consider the channel with channel matrix*

$$\begin{bmatrix} 1/6 & 1/3 & 1/2 \\ 1/3 & 1/2 & 1/6 \\ 1/2 & 1/6 & 1/3 \end{bmatrix}.$$

Given the input distribution

$$p(x = 0) = 0.5, \qquad p(x = 1) = p(x = 2) = 0.25,$$

find the best decision scheme (for transmission with no coding) and the associated average and maximum probabilities of error.

Problem 45. *Consider the channel over $\mathbb{F}_5$ defined by the forward condiotional probabilities*

$$p(i|i) = p(i|i-1) = 0.5, \ \forall i \in \mathbb{F}_5; \qquad p(i|j) = 0, \ \textit{otherwise}\ .$$

a) *Determine the capacity of the channel;*
b) *Find a code of length 2 with zero probability of decoding error.*

Problem 46. *Let C be the binary code*

$$C = \{0000, 0011, 1100, 1111\}.$$

Knowing that these codewords are used with probabilities, respectively, $1/2$, $1/8$, $1/8$ *and* $1/4$,*and transmitted through a Binary Symmetric Channel*

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

determine the optimal decoding rule for each value of $0 < p < 0.5$.

The following two subsections include, in the form of a sequence of exercises, two results related to the material in these notes that were not included in the main text.

2.3. **Maximal and uniform probability of error.** Shannon's Theorem and its converses give results either on the maximal probability of error $p_e^{\max}$ or on the uniform probability of error p_e^u. The following proposition shows that we may use either of them in the statement:

Proposition 47. *Consider a discrete memoryless channel with capacity C. The following are equivalent:*

1- *For any $R < C$, there exists a sequence C_n of q-ary $(n, \lceil q^{nR} \rceil)$ codes, with decision schemes f_n, such that*

$$\lim_{n \to +\infty} p_e^{\max}(C_n) = 0.$$

2- *For any $R' < C$, there exists a sequence D_n of q-ary $(n, \lceil q^{nR'} \rceil)$ codes, with decision schemes g_n, such that*

$$\lim_{n \to +\infty} p_e^{u}(D_n) = 0.$$

Exercise 48. *It is only necessary to prove* $2 \implies 1$.

In order to prove $2 \implies 1$, we need the following

Lemma 49. *Suppose $0 < R < C$. There exists R' satisfying, for sufficiently large n,*

$$R + \frac{\log_q(2)}{n} + \frac{1}{n} \leq R' < C, \text{ and } \frac{1}{2}\lceil q^{nR'} \rceil \geq \lceil q^{nR} \rceil.$$

Exercise 50. *Prove the lemma.*

Exercise 51. *Complete the proof of the proposition as follows: assume* 2 *and let $R < C$. For a fixed $\varepsilon > 0$, justify the existence of a sequence D_n of $(n, \lceil q^{nR'} \rceil)$ q-ary codes such that $p_e^u(D_n) < \frac{\varepsilon}{2}$.*
Show that, as a consequence, at least half of the codewords $d \in D_n$ satisfy $p(error|d) < \varepsilon$.
Conclude the proof of 1.

2.4. **Fano's Inequality.** We state again Fano's inequality:

Theorem 52 (Fano's Inequality)**.** *For any code C with size M, and any decision scheme f, and for any probability distribution on the codewords, if p_e denotes the probability of decision error, then*

$$H(X|Y) \leq H(p_e) + \log(M-1)p_e.$$

Remark 53. *In the inequality, all logarithms have the same base. If we use* $\log_2$ *on the righthand side and the normalized entropy on the left, the formula is*

$$H_M(X|Y) \leq \frac{H_2(p_e) + \log_2(M-1)p_e}{\log_2(M)}.$$

Fano's inequality plays an important role in the deduction of the converse of Shannon's Theorem.

Proof. Fix $u \in \mathbb{F}_q^n$ and assume, without loos of generality, that $f(u) = c_1$. Denote $\rho_i = p(c_i|u)$. The following exercise contains a general fact about entropy functions. Recall that $H(s)$ denotes the entropy function $H(s, 1-s) = -(s\log(s) + (1-s)\log(1-s))$.

Exercise 54. $H(\rho_1, \cdots, \rho_M) = H(1-\rho_1) + (1-\rho_1)H\left(\frac{\rho_2}{1-\rho_1}, \cdots, \frac{\rho_M}{1-\rho_1}\right)$.

Exercise 55. *Apply the result in the exercise to get*

$$H(X|Y=u) \leq H(p(error|u)) + p(error|u)\log(M-1).$$

Exercise 56. *Prove that for any $s_1, \cdots, s_m$ with $0 \leq s_i \leq 1$ and non-negative $t_1, \cdots, t_m$ such that $\sum_i t_i = 1$,*

$$\sum_i t_i H(s_i) \leq H\left(\sum_i t_i s_i\right).$$

Exercise 57. *Finish the proof of the theorem.*

□