1. Channels, Vector random variables and Codes

This set of notes gives an overview of the application of the model of a channel, discussed earlier, and the related notions of conditional entropy, mutual information, etc., to the process of encoding, transmission and decoding of messages using a block code.
We deduce, as a motivation, a relation betwen 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 This result is superseded by the main result, Shannon's theorem.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 $p(c) = p(u)$. If the probability distribution on the alphabet is uniform, we obtain an also uniform probability distribution on C: $p(c) = M^{-1}$, 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 1. *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 generalize directly to the case of vector random variables.
However, for this generalization 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 particularly 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 2 (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 3. *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 4. *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

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

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 5. *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 6. *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 7. *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 $\mathrm{dist}(x, y)$ between two vectors $x, y \in \mathbb{F}_q^n$ is equal to the number of coordinates where the two vectors are different.

Definition 8. *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

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

is maximized by minimizing $\mathrm{dist}(u, c)$.
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 na exercise.

Example 9. *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.

3. Probability of error and Code parameters

As it was seen before, under appropriate conditions Maximum Likelywood Decoding coincides with Minimal Distance Decoding. 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 - not necessarily linear- with length n, size M and minimal distance d) is said to be **optimal** if it is not contained in a $(n, M+1, d)$ code.
This is a good point to introduce another parameter for block codes: the sphere of radius r around $x \in \mathbb{F}_q^n$ is

$$N(x, r) = \{y \in \mathbb{F}_q^n : \text{dist}(x, y) \leq r\};$$

Definition 10. *the* ***covering radius*** *of the linear code C is*

$$cov(C) = \min\{s : \cup_{c\in C} N(c, s) = \mathbb{F}_q^n\}.$$

With this definition, a (n, M, d) code is optimal if and only if $cov(C) < d$ (**HW**).

We have also the following characterizations of the covering radius for linear codes:

Lemma 11. *The covering radius of a linear code C is equal to*

i) $\max\{i : \alpha_i > 0\}$*, where α_i denotes the number of unique coset leaders with weight i;*
ii) *the smallest integer s such that any $v \in \mathbb{F}_q^{n-k}$ is a linear combination of some s columns of the parity check matrix of C.*

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 $a \in \mathbb{F}_q$, $\rho < 0.5$.

We then have

Proposition 12. *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 13. *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 $n \to +\infty$.

3.1. **Probability of error and Syndrome Decoding.** Let C be a $[n,k,d]$ linear code over $\mathbb{F}_q$. Under syndrome decoding, error patterns are corrected if they are unique coset leaders. Let, as above, α_i denote the number of unique coset leaders with weight i. Then the probability of correct decoding is

$$\sum_{i=0}^{n} \alpha_i \left(\frac{\rho}{q-1}\right)^i (1-\rho)^{n-i}$$

or equivalently

Proposition 14. *The probability of decoding error is*

$$1 - \sum_{i=0}^{n} \alpha_i \left(\frac{\rho}{q-1}\right)^i (1-\rho)^{n-i}.$$

If the received word u is a codeword (different from c) then not only the error is not corrected but is undetected. This will happen if and only if the error pattern $u - c$ is also a codeword. So we have

Proposition 15. *The probability of an error pattern to be undetected is*

$$p_{ed} = \sum_{i=1}^{n} A_i \left(\frac{\rho}{q-1}\right)^i (1-\rho)^{n-i}$$

where A_i denotes the number of codewords with weight i.

3.2. **Example: The Binary Symmetric Channel.** In this subsection, we illustrate the computation of probability of error, probability distributions and entropy for three codes used in the encoding of messages from $\mathbb{F}_2^4$.

3.2.1. *Probability of error.* Obviously, different codes encode the same message with different probability of decoding error. The details of the example below, namely the deduction of the various explicit formulas for the error, are left as an exercise (**HW**).

Consider a Binary Symmetric Channel with crossover probability ρ. The probability of error in the decoding of a message (x_1, x_2, x_3, x_4), if Minimal Distance Decoding is applied, is

i) $1-(1-\rho)^4$ if the message is not encoded (equivalently, if the trivial code $\mathbb{F}_2^4$ is used);
ii) $1-(1-\rho)^7-7\rho(1-\rho)^6$ if the message is encoded with the Hamming $[7,4,3]$ code;
iii) and finally,

$$1-\sum_{j=0}^{4}\binom{4}{j}\left(3\rho(1-\rho)^2\right)^j(1-\rho)^{12-3j}=\sum_{j=1}^{4}\binom{4}{j}(3\rho^2-2\rho^3)^j(3\rho(1-\rho)^2+(1-\rho)^3)^{4-j},$$

if the message is encoded with the 3 repetition code, ie, if the encoded message is

$$(x_1,x_1,x_1,x_2,x_2,x_2,x_3,x_3,x_3,x_4,x_4,x_4).$$

For $\rho=0.01$ the approximate values of these probabilities are respectively

$$0.03094, \qquad 0.00203, \qquad 0.00119;$$

for $\rho=0.1$ we get, respectively,

$$0.344, \qquad 0.15, \qquad 0.107.$$

The information rates are

$$1, \qquad 4/7, \qquad 1/3.$$

As expected, the improvement in error-correcting capability is obtained at the cost of lower information rates.

3.3. **Conditional Entropy.** For the remainder of this section, we fix the channel matrix with $\rho=0.1$,

$$\begin{bmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{bmatrix}$$

and consider now the computation of the conditional entropy. If the input probability distribution is uniform, we know the same happens for the output; so $H(x)=H(y)=1$.
The backward conditional probabilites are equal to the forward ones and the conditional probabilities are

$$H(x|0)=H(x|1)=H(x|y)=H(\rho)\approx 0.469.$$

3.3.1. *The trivial code.* Consider first the no-coding case:
If the messages are, as above, vectors (x_1, x_2, x_3, x_4) and we use the trivial code $\mathbb{F}_2^4$, the probability of each codeword c is 2^{-4} and, for any output vector u, $p(Y = u|X = c) = \rho^s(1-\rho)^{4-s}$ where $s = \mathrm{dist}(u, c)$.
Using the Binomial theorem, we confirm that the output probability distribution is uniform as well: $p(u) = 2^{-4}$ for any $u \in \mathbb{F}_2^4$, and so the backward probabilities are $p(X = c|Y = u) = p(Y = u|X = c)$.

This implies that, for any u,

$$H(X|u) = -\sum_c p(c|u)\log(p(c|u)) = -\sum_{s=0}^{4}\binom{4}{s}\rho^s(\rho)^{4-s}\log\left(\rho^s(\rho)^{4-s}\right),$$

which, if computed with the base 2^4 for the logarithms is exactly equal to $H_2(\rho)$ (**HW**).
So, if suitably normalized, $H(X|Y) = H(x|y)$.

3.3.2. *Direct Sum codes.* This observation may be generalized: the trivial $[4, 4]$ code $\mathbb{F}_2^4$ is the direct sum of four copies of the trivial $[1, 1]$ code. We consider the general problem of computation of the probability distributions, conditional probabilities and conditional entropies for the direct sum of codes, starting with the case of two codes C_1 and C_2, with sizes respectively M_1 and M_2, over $\mathbb{F}_q$. The direct product has size $M = M_1M_2$.
Because we will not need to refer explicitly to vector coordinates, we may denote a vector of the direct product as v_1v_2, with $v_i \in C_i$. We will use a similar notation for random variables, etc.
Assuming the uniform probability on the alpabet, we have uniform probability distribution in the codes and, in particular, $p(c_1c_2) = p(c_1)p(c_2)$. It follows that (**HW**)

$$p(u_1u_2|c_1c_2) = p(u_1|c_1)p(u_2|c_2), \qquad p(u_1u_2) = p(u_1)p(u_2) \text{ and } p(c_1c_2|u_1u_2) = p(c_1|u_1)p(c_2|u_2).$$

For any output u_1u_2, we have then (**HW**)

$$H(X|u_1u_2) = H(X_1|u_1) + H(X_2|u_2).$$

However, this equality holds with the same choice of logarithm in both sides, and, following the definition, we should use $\log_M$ in the lefthand side and $\log_{M_i}$ in each of the summands of the righthand side. But

$$\log_M(a) = (1 + \log_{M_1}(M_2)\log_{M_1}(a),$$

and so we obtain the general formula, where the subscript in H indicates the choice of base:

$$H_M(X|u_1u_2) = \frac{H_{M_1}(X_1|u_1)}{1 + \log_{M_1}(M_2} + \frac{H_{M_2}(X_2|u_2)}{1 + \log_{M_2}(M_1}.$$

It is easy (**HW**) to generalize this to a direct sum of codes $C_1, \cdots, C_t$, with sizes $M_1, \cdots, M_t$ respectively: in a more symmetric form, and denoting $M = \prod_i M_i$, for any $u = u_1 \cdots u_t$,

$$H_M(X|u) = \sum_{i=1}^{t}\frac{H_{M_i}(X_i|u_i)}{\sum_j \log_{M_i}(M_j)} = \sum_{i=1}^{t}\frac{H_{M_i}(X_i|u_i)}{\log_{M_i}(M)},$$

i.e., the conditional entropy of the direct sum of codes is a weighted average of the conditional entropies of the summands.

Finally, we obtain a similar formula

$$H_M(X|Y) = \sum_{i=1}^{t} \frac{H_{M_i}(X_i|Y_i)}{\log_{M_i}(M)}.$$

Exercise 16. *Prove the last formula.* ***Hint:*** *Consider first the case $t = 2$ and apply induction.*

In the case that all the codes C_i have the same size M, this simplifies to

$$H_{M^t}(X|u) = \frac{1}{t}\sum_{i=1}^{t} H_{M_i}(X_i|u_i), \qquad H_{M^t}(X|Y) = \frac{1}{t}\sum_{i=1}^{t} H_{M_i}(X_i|Y_i).$$

3.3.3. *Repetition Code.* We consider now the $[3, 1]$ repetition code:

If $p(0) = p(1) = 0.5$ then also $p(000) = p(111) = 0.5$. The forward conditional probabilities for the code would now form a 2×8 matrix. We have

$$p(000|000) = p(111|111) = (1-p)^3 = 0.729,\ p(000|111) = p(111|000) = p^3 = 0.001;$$

other values of $p(u|c)$ are determined by the distance; so $p(100|000) = (1-p)^2 p = 0.081$ while $p(100|111) = (1-p)p^2 = 0.009$, and so on.
We obtain the output probabilities $p(000) = p(111) = \frac{(1-p)^3+p^3}{2} = 0.365$, while all other vectors have probability

$$\frac{(1-p)^2 p + (1-p)p^2}{2} = \frac{(1-p)p}{2} = 0.045.$$

The backward conditional probabilities are

$$p(000|u) = \begin{cases} \frac{(1-p)^3}{(1-p)^3+p^3} \approx 0.9986 & u = 000 \\ 1-p = 0.9 & w(u) = 1 \\ p = 0.1 & w(u) = 2 \\ \frac{p^3}{(1-p)^3+p^3} \approx 0.00137 & w(u) = 3 \end{cases}$$

and these values are reversed for $p(111|u)$.

It turns out that the conditional entropies are

$$H(X|000) = H(X|111) \approx 0.015,\ H(X|u) = H(\rho) \approx 0.467 \text{ for other } u,$$

and finally, $H(X|Y) \approx 0.137$. As expected the uncertainty on X remaining after the knowledge of Y is much smaller than with no coding.

If we apply this to the encoding of messages $(x_1, x_2, x_3, x_4) \in \mathbb{F}_2^4$, we are using a direct sum of four copies of the repetition $[3, 1]$ code, and we may apply the results from the last paragraph. In particular, we confirm that the conditional entropy $H(X|Y)$, suitably normalized by the size of the code, is approximately equal to 0.137.

3.3.4. *Hamming Code.* Now consider the $[7, 4]$ Hamming code with uniform input probability, $p(c) = 2^{-4}$ for every c.
Again, if we want to normalize entropy, we must use base 2^4 logarithms, i.e., we may use base 2 and divide everything by 4. With this convention, $H(X) = 1$. The forward conditional probabilities will be

$$p(u|c) = (1-p)^{7-s}p^s$$

where $s = \text{dist}(u, c)$, and so

$$p(u) = 2^{-4}\sum_c p(u|c);$$

We know that the code contains 1 vector with weight 0, 7 with weight 3, 7 with weight 4 and 1 with weight 7; this implies that the output probability of the zero vector is

$$p(Y = 0) = 2^{-4}((1-p)^7 + 7(1-p)^4p^3 + 7(1-p)^3p^4 + p^7) \approx 0.03;$$

translation invariance of Hamming distance implies the output probability of any codeword c is equal to this:

$$p(Y = c) = 2^{-4}\sum_{d\in C}(1-p)^{7-s}p^s,$$

where $s = \text{dist}(c, d)$; but $\text{dist}(c, d) = \text{dist}(0, d - c)$ and $d \to d - c$ is a bijection of the code.
This reasoning shows that we need to compute $p(Y = u)$ only for one element in each coset, e.g., the coset leader. It happens that, for each coordinate i, there exist exactly 3 codewords of weight 3 and 4 codewords of weight 4 that are nonzero at i. With this fact, it is then easy to compute for any vector e with weight 1

$$p(Y = e) = 2^{-4}((1-p)^6p+3(1-p)^5p^2+4(1-p)^4p^3+4(1-p)^3p^4+3(1-p)^2p^5+p^6) \approx 0.0046.$$

We are now able to compute backward conditional probabilities $p(c|u)$: if c and d are distinct codewords,

$$p(X = c|Y = d) = \frac{p(Y = d|X = c)p(X = c)}{p(Y = d)} \approx p(Y = d|X = c) \times a,$$

where $a = 2.083$. So

$$p(X = c|Y = d) = \begin{cases} a(1-p)^7 \approx 0.996 & \text{if } \text{dist}(c, d) = 0 \\ a(1-p)^4p^3 \approx 0.00137 & \text{if } \text{dist}(c, d) = 3 \\ a(1-p)^3p^4 \approx 1.52 \times 10^{-4} & \text{if } \text{dist}(c, d) = 4 \\ ap^7 = 2.083 \times 10^{-7} & \text{if } \text{dist}(c, d) = 7 \end{cases}$$

If u is not a codeword, and

$$b = \frac{p(X = c)}{p(Y = u)} = \frac{2^{-4}}{4.6 \times 10^{-3}} \approx 13.587,$$

then

$$p(X = c|Y = u) = \begin{cases} b(1-p)^6p \approx 0.720 & \text{if } \text{dist}(c, u) = 1 \\ b(1-p)^5p^2 \approx 0.082 & \text{if } \text{dist}(c, u) = 2 \\ b(1-p)^4p^3 \approx 0.009 & \text{if } \text{dist}(c, u) = 3 \\ b(1-p)^3p^4 \approx 0.001 & \text{if } \text{dist}(c, u) = 4 \\ b(1-p)^2p^5 \approx 1.1 \times 10^{-4} & \text{if } \text{dist}(c, u) = 5 \\ bp^6 = 1 \times 1.36 \times 10^{-5} & \text{if } \text{dist}(c, u) = 6 \end{cases}$$

If c is a codeword, $H(X|c) \approx 0.02762$, while for other vectors u, $H(X|u) \approx 0.3795$. Finally, $H(X|Y) \approx 0.2088$.

4. Shannon's Second 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 arbitrarly small:

Theorem 17 (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 summarize in the next theorem:

Theorem 18. *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 19. *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.*

5. Supplementary Results and Problems

Problem 20. *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.

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.

5.1. **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 21. *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 22. *It is only necessary to prove* $2 \implies 1$.

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

Lemma 23. *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 24. *Prove the lemma.*

Exercise 25. *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.

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

Theorem 26 (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 27. *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 28. $H(\rho_1, \cdots, \rho_M) = H(1-\rho_1) + (1-\rho_1)H\left(\frac{\rho_2}{1-\rho_1}, \cdots, \frac{\rho_M}{1-\rho_1}\right)$.

Exercise 29. *Apply the result in the exercise to get*

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

Exercise 30. *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 31. *Finish the proof of the theorem.*

□