I. CHANNEL CODING

We consider the following channel coding problem:

Prior to Shannon results, it was assumed that the error probability of the channel communication in Figure 1, grows as $R$ grows, where $R$ is the rate transmitted through the channel, i.e., the number of bits transmitted through the channel per one usage of the channel.

According to Shannon theorem, as long as we allow a delay such that the encoding is done in blocks of size $n$ (where $n$ can be as large as we want), the error probability is arbitrary low for $R \leq C$, and is 1 for $R > C$, where $C$ is the channel capacity. This is illustrated in Figure 2.

**Assumption 1 (Memoryless)** The channel is memoryless, i.e.,

$$P(y_i|x_{i-1}, y_{i-1}) = P(y_i|x_{i}) \ .$$ (1)

**Assumption 2 (Discrete time)** The channel is Discrete time.

We will denote a discrete time memoryless channel as: DMC.

**Lemma 1 (Memoryless channel without feedback)** For a memoryless channel without feedback, i.e., $P(x_i|x_{i-1}, y_{i-1}) = P(x_i|x_{i-1})$:

$$P(y^n|x^n) = \prod_{i=1}^{n} P(y_i|x_i) \ .$$ (2)
Note: When no feedback is available, a memoryless channel can also be defined by (2).

Proof:

\[
P(y^n | x^n) = \frac{P(y^n, x^n)}{P(x^n)}
\]

\[
= \frac{\prod_{i=1}^{n} P(y_i | x_i, y^{i-1}, x^{i-1})}{P(x^n)}
\]

\[
\overset{(a)}{=} \prod_{i=1}^{n} P(x_i | y^{i-1}, x^{i-1}) P(y_i | x_i, y^{i-1}, x^{i-1})
\]

\[
\overset{(b)}{=} \prod_{i=1}^{n} P(x_i | x^{i-1}) P(y_i | x_i)
\]

\[
\overset{(c)}{=} \frac{P(x^n) \prod_{i=1}^{n} P(y_i | x_i)}{P(x^n)}
\]

\[
= \prod_{i=1}^{n} P(y_i | x_i)
\]  

Where:

(a) - Follows from the chain rule.
(b) - Follows from the memoryless and no feedback properites.

(c) - Follows from the probability chain rule.

Definition 1 (Code)  An \( (n, 2^{nR}) \) code for the channel \( (\mathcal{X}, P_{Y|X}, \mathcal{Y}) \) consists of the following:

1) An index message \( \{1, \ldots, 2^{nR}\} \)

2) An encoding function \( f : \{1, 2, \ldots, 2^{nR}\} \rightarrow \mathcal{X}^n \), \( (4) \)
yielding codewords \( x^n(1), x^n(2), \ldots, x^n(2^{nR}) \). The set of codewords is called the codebook.

3) A decoding function \( g : \mathcal{Y}^n \rightarrow \{1, 2, \ldots, 2^{nR}\} \), \( (5) \)
which is a deterministic rule that assigns a guess to each possible received vector.

Definition 2 (Maximal probability of error)  The maximal probability of error, \( P_{\text{max}}^{(n)} \), for an \( (n, 2^{nR}) \) code is defined as:

\[
P_{\text{max}}^{(n)} = \max_m P \left( M \neq \hat{M} | M = m \right), \quad (6)
\]

Note that \( \hat{M} \) depends on the code since \( \hat{M} = g(Y^n) \) and \( Y^n \) is the output of the channel where the input is \( X^n = f(M) \).

Definition 3 (Average probability of error)  The average probability of error, \( P_{\text{er}}^{(n)} \), for an \( (n, 2^{nR}) \) code is defined as:

\[
P_{\text{er}}^{(n)} = \Pr \left( M \neq \hat{M} \right) \quad (7)
\]

Definition 4 (Achievable rate)  the rate \( R \) is achievable if there exists a sequence of \( (2^{nR}, n) \) codes such that:

\[
\lim_{n \rightarrow \infty} P_{\text{er}}^{(n)} = 0. \quad (8)
\]

Definition 5 (Capacity)  Capacity, denoted by \( C \), is the suprimum over all achievable rates.
Note that Def. 5 is an operational definition, namely, it rises from the operational definition of communication problem. The next theorem relates the operational definition of capacity to a mathematical quantity and can be calculated.

**Theorem 1 (Channel capacity)** For DMC $(\mathcal{X}, P_{Y|X}, \mathcal{Y})$ the capacity satisfies

\[ C = \max_{P_X} I(X;Y) \]  \hspace{1cm} (9)

where $I(X;Y)$ is the mutual information and $P_X$ is an input distribution of the channel.

II. **Examples**

A. **BSC - Binary symmetric channel**

**Example 1 (BSC - Binary symmetric channel)** Consider the BSC, shown in Figure 3:

![BSC diagram]

Fig. 3. BSC - Binary symmetric channel. $C = 1 - H_b(\delta)$

The output $Y$, can be written as:

\[ Y = X \oplus Z \]

Where $Z \sim Bernoulli(\delta)$ and $Z \perp X$, i.e., $Z$ and $X$ are independent. If $Z = 1$, a flip occurs, and if $Z = 0$, there’s no flip. The mutual information is given by:

\[ I(X;Y) = H(Y) - H(Y|X) \]

\[ \overset{(a)}{=} H(Y) - H(Y \oplus X|X) \]

\[ \overset{(b)}{=} H(Y) - H(Z|X) \]

\[ \overset{(c)}{=} H(Y) - H(Z) \]
\[ \leq 1 - H_b(\delta) \]  \hspace{1cm} (10)

Where:

- \(H_b(\delta)\) is the binary entropy, i.e.:
  \[
  H_b(\delta) \triangleq -\delta \log(\delta) - (1 - \delta) \log(1 - \delta)
  \]  \hspace{1cm} (11)

- (a) - Given \(X\), there is a one to one mapping between \(X\) and \(Y\).
- (b) - Follows from the fact that \(Y \oplus X = Z\).
- (c) - Follows from the fact that \(Z \perp X\).
- (d) - Follows from the fact that \(H(Y)\) is bounded by \(\log |Y|\).

Note that if one chooses \(X \sim Ber(0.5)\), equality in (d) holds, i.e. \(I(X; Y) = 1 - H_b(\delta)\).

One more example, called the erasure channel, is explained in details in the appendix of the lecture.

### III. Joint Weak Typicality

**Lemma 2** (Weak typicality)

\[
A^n_\epsilon = \{(x^n, y^n) \in X^n \times Y^n : | - \frac{1}{n} \log p(x^n) - H(X) | \leq \epsilon, \;
| - \frac{1}{n} \log p(y^n) - H(Y) | \leq \epsilon, \;
| - \frac{1}{n} \log p(x^n, y^n) - H(X, Y) | \leq \epsilon \}.
\]  \hspace{1cm} (12)

**Theorem 2** Let \(X^n, Y^n\) be i.i.d. \(\sim P_{XY}(x, y)\) then:

1) \(\lim_{n \to \infty} \Pr\{(X^n, Y^n) \in A^n_\epsilon\} = 1\)

2) \(|A^n_\epsilon| \leq 2^{n(H(X,Y) + \epsilon)}\)

3) If \(\left(\tilde{X}^n, \tilde{Y}^n\right) \sim \prod_{i=1}^n P_X(x_i) P_Y(y_i)\), then:

\[
\Pr\left\{ \left(\tilde{X}^n, \tilde{Y}^n\right) \in A^n_\epsilon \right\} \leq 2^{-n(I(X,Y) - 3\epsilon)}
\]  \hspace{1cm} (15)

**Proof:**

1) follows from the Law of large numbers.
2) follows from:

\[
1 = \sum_{x^n, y^n} P_{X^n, Y^n}(x^n, y^n) \geq \sum_{(x^n, y^n) \in A^n_\epsilon} P_{X^n, Y^n}(x^n, y^n) \geq \sum_{(x^n, y^n) \in A^n_\epsilon} 2^{-n(H(X,Y)+\epsilon)} = |A^n_\epsilon| \cdot 2^{-n(H(X,Y)+\epsilon)} \quad (16)
\]

Where:

(a) - Summing over less elements reduces the probability.

(b) - Follows from (14).

3) We need to upper bounds the probability that \( (\tilde{X}^n, \tilde{Y}^n) \) is in the set \( A^{(n)}_\epsilon \). We do so by summing over the probability of the elements in \( A^{(n)}_\epsilon \) according to the distribution of \( (\tilde{X}^n, \tilde{Y}^n) \).

\[
P\left[ (\tilde{X}^n, \tilde{Y}^n) \in A^n_\epsilon \right] = \sum_{(x^n, y^n) \in A^n_\epsilon} P_{\tilde{X}^n, \tilde{Y}^n}(x^n, y^n)
= \sum_{(x^n, y^n) \in A^n_\epsilon} P_{X^n}(x^n) P_{Y^n}(y^n)
\leq |A^n_\epsilon| \cdot 2^{-n(H(X)-\epsilon)} \cdot 2^{-n(H(Y)-\epsilon)}
\leq 2^{n(H(X,Y)+\epsilon)} \cdot 2^{-n(H(X)-\epsilon)} \cdot 2^{-n(H(Y)-\epsilon)}
\leq 2^{-n(I(X;Y)-3\epsilon)} \quad (17)
\]

Where:

- (a) - Using the bound on \( P_{X^n}(x^n) = \) and \( P_{Y^n}(y^n) \) according to (12) and (13).
- (b) - Using the bound on the size of the set, i.e., \(|A^n_\epsilon| \leq 2^n(H(X,Y)+\epsilon)\).
- (c) - Follows from the definition of mutual information.

IV. PROOF OF THE CAPACITY THEOREM

Proof of the converse part of Theorem 1 : In the converse part (upper bound on \( C \)) we need to prove that
if $R$ is an achievable rate then

$$R \leq C^I = \max_P I (X; Y)$$  \hspace{1cm} (18)$$

Fix a code $(n, 2^{nR})$ with a probability of error $P_e^{(n)}$. Denote by $M$ the message, which is distributed uniformly over $\{1, ..., 2^{nR}\}$, to be sent. Thus, we have:

$$nR \overset{(a)}{=} H (M)$$

$$= H (M) + H (M|Y^n) - H (M|Y^n)$$

$$\overset{(b)}{=} I (M; Y^n) + H (M|Y^n)$$

$$\overset{(c)}{=} I (M; Y^n) + H \left(M|Y^n, \hat{M}\right)$$

$$\leq I (M; Y^n) + H \left(M|\hat{M}\right)$$

$$\overset{(d)}{=} I (M; Y^n) + H \left(M|\hat{M}\right)$$

$$\leq I (M; Y^n) + (1 + P_e \cdot nR)$$

$$\overset{(f)}{=} I (M; Y^n) + n \cdot \epsilon_n$$

$$= H (Y^n) - H (Y^n|M) + n \cdot \epsilon_n$$

$$\overset{(g)}{=} \sum_{i=1}^{n} \left[ H (Y_i|Y^{i-1}) - H (Y_i|Y^{i-1}, M, X^i) \right] + n \cdot \epsilon_n$$

$$\overset{(h)}{=} \sum_{i=1}^{n} \left[ H (Y_i|Y^{i-1}) - H (Y_i|X_i) \right] + n \cdot \epsilon_n$$

$$\overset{(i)}{=} \sum_{i=1}^{n} \left[ H (Y_i) - H (Y_i|X_i) \right] + n \cdot \epsilon_n$$

$$= \sum_{i=1}^{n} I (Y_i; X_i) + n \cdot \epsilon_n$$

$$\overset{(j)}{\leq} n \cdot C^I + n \cdot \epsilon_n$$  \hspace{1cm} (23)$$

Where:

(a) - Message number is uniform distributed, i.e., $M \sim uniform (1, 2, ..., 2^{nR})$. Thus: $H (M) = \log |\mathcal{M}|$.

(b) - Definition of Mutual Information
(c) - Since \( \hat{M} = f(Y^n) \) is a deterministic function of \( Y^n \).

(d) - Conditioning reduces entropy.

(e) - \( P_{er} = P(M \neq \hat{M}) \) and using Fano’s inequality: \( H(M|\hat{M}) \leq 1 + P(M \neq \hat{M}) \log |\mathcal{M}| \).

(f) - \( \epsilon = 1/n + P_e \times R \)

(g) - \( X^n \) is a function of the Message \( M \).

(h) - Follows from the property of a memoryless channel, as described in (2).

(i) - Removing given information from entropy increases entropy.

(j) - According to the definition of capacity: \( C = \max_{P(X)} I(X;Y) \).

Now, dividing both sides by \( n \), we get

\[
R \leq C_I + \epsilon_n \quad (24)
\]

and since we assumed that \( R \) is an achievable rate, then there exists a sequence of codes \((n, 2^{nR})\) such that \( P_{err}^{(n)} \to 0 \) as \( n \to \infty \) and this implies also that \( \epsilon_n \to 0 \) as \( n \to \infty \). Therefore we obtained that if \( R \) is achievable, then \( R \leq C_I \), and that satisfies the proof.

We now prove that if \( R < C_I \) then for a DMC (discrete memoryless channel) there exists a sequence of codes \((2^n, n)\) such that \( P_{err}^{(n)} \to 0 \) as \( n \to \infty \). The proof is based on a random coding algorithm. We generate a code randomly, and show that expectation (over all codebooks) the probability of error goes to zero. Since the probability of error goes to zero when we average over all codes, then there exists at least one code for which the probability of error goes to zero.

**Proof of achievability part of Theorem 1:**

**Design of the code:** We fix \( P_X(x) \) and a rate \( R \), and generate the codebook \( \mathcal{C} \), with entries \( X^n(i) \), where \( X^n(i) \) is the codeword associated with message \( i \), and: \( X^n(i) \stackrel{i.i.d.}{\sim} P_X(x), i = 1, \ldots, 2^{nR} \). Reveal the codebook to the encoder and the decoder.

**Encoder:** Encode message \( i \) into codeword \( i \) from the codebook \( \mathcal{C} \), i.e., \( X^n(i) \).
Decoder: The decoder receives \( Y^n \), and looks for \( X^n \in C \) such that: \((X^n, Y^n) \in A^n_{\epsilon}\). If such \( X^n \) exists, then the message associated with \( X^n \) is decoded. Meaning: the decoder goes through every \( X^n \) in the codebook, and checks whether or not \((X^n, Y^n)\) are jointly typical. If so, it decodes the message associated with \( X^n \).

Analysis of error:

\[
P_{er} = P \left( M \neq \hat{M} \right) = \sum_{m=1}^{2^nR} P(M = m)P \left( M \neq \hat{M} | M = m \right) = \frac{1}{2^nR} \sum_{m=1}^{2^nR} P \left( M \neq \hat{M} | M = m \right)
\]

(25)

Where the last equality holds since we assume all messages are equally probable. Because of symmetry we may assume that \( m = 1 \), and analyzes only \( P \left( M \neq \hat{M} | M = 1 \right) \).

An error is defined if \((X^n(1), Y^n) \notin A^n_{\epsilon}\) or if there exists a \( j \neq 1 \) such that \((X^n(j), Y^n) \in A^n_{\epsilon}\). Hence, there are two types of error. We need to show that both error goes to zero as the block-length \( n \) goes to infinity.

- \((X^n(1), Y^n) \notin A^n_{\epsilon}\). This error goes to zero as \( n \to \infty \), since following the first property in Theorem 2 we have \( P \left[ (X^n(1), Y^n) \notin A_{\epsilon} \right] \to \infty 0.\)

- \( \exists j \neq 1 \) s.t. \((X^n(j), Y^n) \in A^n_{\epsilon}\). This error also goes to 0 as \( n \to \infty \) since:

\[
P_{\text{Error}2} = P \left( \bigcup_{j=2}^{2^nR} E_j \right) \leq \sum_{j=2}^{2^nR} P(E_j)
\]

\[
\leq 2^{nR}\sum_{j=2}^{2^nR} P(E_j)
\]

\[
\leq 2^{nR}2^{-n(I(X;Y) - 3\epsilon)}
\]

\[
= 2^{n(R - I(X;Y) + 3\epsilon)}
\]

(26)

Where:

- (a) follows from the union bound, e.g. , \( P(A \cup B) \leq P(A) + P(B).\)

- (b) follows from (17).

Hence if \( R < I(X;Y) \), then \( P_{er} \to 0 \) as \( n \to \infty \).

Having showed that there exists a code s.t. the average error probability, defined in (3), tends to 0, as \( n \to \infty \), we will now show that a small average probability of error implies a small maximal probability of error, defined in (2), at essentially the same rate.
**Theorem 3 (Half Codewords)** Assume that $P_{er} = \epsilon$. There exists a set of codewords that is half of the size, i.e.:

$$
\frac{2^nR}{2} = 2^n(R - \frac{1}{n})
$$

and $P_{max} \leq 2\epsilon$, where $P_{er}$ and $P_{max}$ are defined in (3) and (2), respectively.

*Proof:* If we throw away the worst half of the codewords, with the highest error probabilities, we will remain with a codebook, consisting of the best half of the codewords. The remaining codewords must have a maximal probability of error less than $2\epsilon$ (Otherwise, these codewords themselves would contribute more than $2\epsilon$ to the sum, and $P_{er}$ would be greater than $\epsilon$). If we reindex these codewords, we have $2^nR-1$ codewords. Throwing out half the codewords has changed the rate from $R$ to $R - \frac{1}{n}$, which is negligible for large $n$.

This implies that if achievability holds for the average error probability, it holds for the maximum error probability as well.

**Example 2 (Illustration of proof of Theorem 3)** Assume that the average score in a specific class is 30, follows that *at least* half of the class, scored less than 60.

Proof: If half of the class scored *exactly* 60, and the other half scored *exactly* 0, then the average score is 30. But if more than 50 percent score above 60, the average score is above 30 and we get contradiction.

In our example, the score 30 represents $\epsilon$ and the score 60 represents $2 \times \epsilon$.

**APPENDIX A**

**Binary Erasure Channel (BEC)**

A Binary Erasure Channel (BEC) is a common communications channel model used in coding theory and information theory. In this model, a transmitter sends a bit (a zero or a one), and the receiver either receives the bit or it receives a symbol '?' that the represents that the bit was erased, namely, the receiver knows that a bit was sent but it does not know which one. For instance, in an internet protocol, '?' may represent that the packet received was corrupted and $\{0, 1\}$ may represent tow possible packets.
1) **No-Feedback channel:** The capacity of the channel is given by \( C = \max_{P_X} I(X; Y) \), and \( C \) is known to be the upper bound of the achievable rates. So, we try to find the channel capacity:

\[
I(X; Y) = H(X) - H(X|Y) \tag{28}
\]

\[
= H(X) - \sum_{\psi \in Y} P(y = \psi)H(X|y = \psi) \tag{29}
\]

\[
= H(X) - P(y = 0)H(X|y = 0) - P(y = 1)H(X|y = 1) - P(y = \text{?})H(X|y = \text{?}) \tag{30}
\]

Where '?' stands for an erased bit (See Figure 4).

Since the model of the channel suggests that for a successfully received bit we know the sent bit with probability of 1, meaning \( X \) is determined by \( Y \) (given \( y \neq \text{?} \)), we get \( H(X|y = 0) = H(X|y = 1) = 0 \). Also, if the bit was erased, by the symmetry of the channel we have no additional information regarding the value of the transmitted bit. Therefore \( H(X|Y = \text{?}) = H(X) \), and \( P(y = \text{?}) = e \) by symmetry. Substituting this into Eq. 28:
\[ I(X; Y) = H(X) - P(y = '?) H(X|y = '?) = H(X) - e \cdot H(X) = (1 - e)H(X) \quad (31) \]

Finally, we need to find the supremum of the mutual information over all the possible distributions of \( X \).

\[ C = \sup_{P_X} I(X; Y) = (1 - e) \sup_{P_X} H(X) = 1 - e \quad (32) \]

Where the last equation holds for \( X \sim Bernoulli(\frac{1}{2}) \). Therefore an upper bound on the achievable rate is indeed \( R \leq C = 1 - e \), as suggested.

2) **Code achieving capacity that uses feedback:** Assume the transmitter at time \( i \) knows the previous outputs of the channel, i.e., \( y^{i-1} \), so we can re-transmit the “erased” bit. In order to successfully receive \( n \) bits, we must transmit \( \frac{n}{1 - e} \) bits: First we transmit \( n \) bits. Since \( P_e = e \) and we have feedback, we know that \( e \cdot n \) bits were erased, so we need to re-transmit them. This time, \( e \cdot en \) bits were erased, so once again we re-transmit them, and so on. All together, we have transmitted:

\[ n + en + e^2n + \cdots = \sum_{j=0}^{\infty} e^j n = n \cdot \sum_{j=0}^{\infty} e^j = n \cdot \frac{1}{1 - e} = \frac{n}{1 - e} \quad (33) \]

In order to successfully receive \( n \) bits.

**Definition 6 (Code Rate)** In telecommunication and information theory, the code rate \( R \) of a channel code is the proportion of the data-stream that is useful (non-redundant). That is, if the code rate is \( R = k/n \), for every \( k \) bits of useful information, the coder generates totally \( n \) bits of data, of which \( n-k \) are redundant.

In the feedback erasure channel scenario, the ratio of useful information to total sent information was \( R = \frac{n}{n/(1-e)} = 1 - e \).