Introduction to Information Theory

Lecture 6

Lecturer: Haim Permuter *Scribe: Yoav Eisenberg and Yakov Miron*

I. Channel Coding

We consider the following channel coding problem given in Fig. 1. This is a fundamental problem in digital communication and storage of sending a message (set of bits) through a noisy channel.

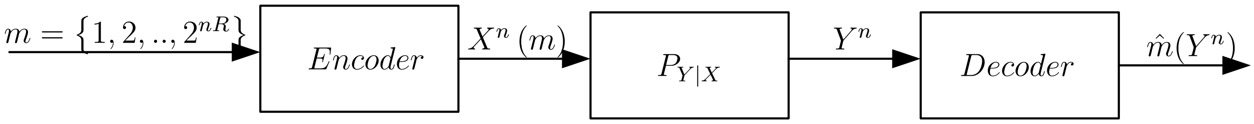

Fig. 1. Channel coding setting.

Prior to Shannon results, it was assumed that the error probability of the channel communication in Fig. 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, 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 Fig. 2.

Assumption 1 (Discrete time) The transmission via the channel happens at discrete time, i.e., 1,2,3,4,5,.... .

Assumption 2 (Memoryless property) The channel is memoryless , i.e.,

$$P\left(y_i|y^{i-1}, x^i, m\right) = P\left(y_i|x_i\right), \quad i = 1, 2, \ldots, n. \tag{1}$$

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

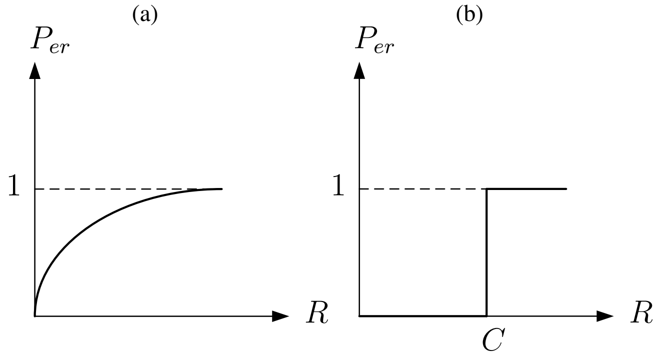

Fig. 2. (a) $P_{er}(R)$ as it was believed prior to Shannon theorem (b) $P_{er}(R)$ as derived from Shannon's theorem on capacity of channels

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

1) A message M that is distributed uniformly on $\left\{1, \ldots, 2^{nR}\right\}$.
2) An encoding function
$$f : \left\{1, 2, \ldots, 2^{nR}\right\} \to \mathcal{X}^n, \tag{2}$$
yielding codewords $x^n(1), x^n(2), \ldots, x^n\left(2^{nR}\right)$. The set of codewords is called the *codebook*.
3) A decoding function
$$g : \mathcal{Y}^n \to \left\{1, 2, \ldots, 2^{nR}\right\}, \tag{3}$$
which is a deterministic rule that assigns a guess to all possible output sequences..

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

$$P^{(n)}_{\max} = \max_m P\left(M \neq \hat{M} | M = m\right), \tag{4}$$

Note that $\hat{M}$ depends on the chosen codebook since $\hat{M} = g(Y^n)$ and $X^n = f(M)$. Therefore, we abuse notation when we omit the dependence on the codebook.

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

$$P^{(n)}_{er} = \Pr\left(M \neq \hat{M}\right) \tag{5}$$

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

$$\lim_{n\to\infty} P_{er}^{(n)} = 0. \tag{6}$$

Definition 5 (Capacity) The *capacity* is denoted by C, and is defined as the supremum over all achievable rates.

In contrast to lossless compression, where one is interested in minimizing the rate, here, higher rate corresponds to more information.

Note that Def. 5 is an operational definition, namely, it arises from the communication definition. The next theorem relates the operational definition of capacity to a mathematical quantity and can be calculated.

Theorem 1 (Channel capacity) For a memoryless channel $(\mathcal{X}, P_{Y|X}, \mathcal{Y})$, the capacity is given by

$$C = \max_{P_X} I\left(X; Y\right), \tag{7}$$

where the joint distribution is $P_X P_{Y|X}$.

Remarks:

- The capacity satisfy $C \leq \log_2 |\mathcal{X}|$.
- The capacity satisfy $C \leq \log_2 |\mathcal{Y}|$.
- The capacity is concave in P_X. Thus, the maximum can be computed efficiently.

II. Examples

A. Binary clean channel

The binary clean channel has binary input and output alphabets, $\mathcal{X} = \mathcal{Y} = \{0, 1\}$. The channel is given by $Y = X$ and is described in Figure 3:

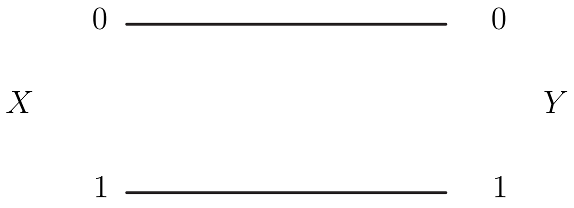

Fig. 3. Binary clean channel

It is easy to note that the maximal achievable rate is 1[bit/Channel use]. The channel coding coding theorem asserts this simple observation with:

$$\begin{aligned} C &= \max_{P_X} I(X;Y) \\ &= \max_{P_X} H(X) \\ &= 1. \end{aligned} \tag{8}$$

B. Noisy channel with non-overlapping outputs

The channel has a input alphabets, $\mathcal{X} = \{0, 1\}$, and quadratic channel output alphabet $\mathcal{Y} = \{0, 1, 2, 3\}$. The channel is described in Figure 4: The capacity is at most $C \leq \log |\mathcal{X}| = 1$[bit/Channel use]. Moreover, we can achieve the upper bound with $p(x) \sim$ Bern(0.5). Formally, from the channel coding coding theorem:

$$C = \max_{P_X} I(X;Y) \tag{9}$$

$$= \max_{P_X} H(X) - H(X|Y) \tag{10}$$

$$= \max_{P_X} H(X) \tag{11}$$

$$= 1. \tag{12}$$

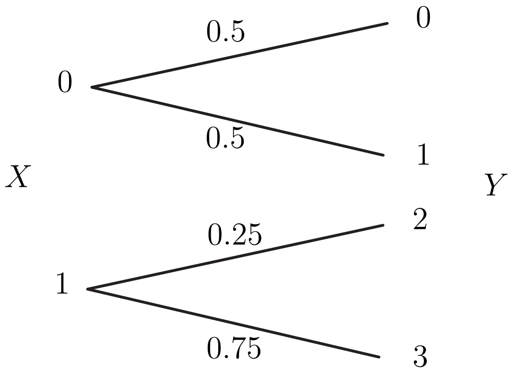

Fig. 4. Noisy channel with non-overlapping outputs.

C. BSC - Binary symmetric channel

Consider the BSC, shown in Figure 5:

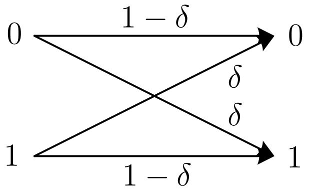

Fig. 5. BSC - Binary symmetric channel. $C = 1 - H_b\left(\delta\right)$

The output Y, can be written as:

$$Y = X \oplus Z,$$

where $Z \sim Bernoulli\left(\delta\right)$ and $Z \perp X$ (Z and X are independent). Note that if $Z = 1$, an error occurred, and $Z = 0$ means that there is no error. The mutual information is given by:

$$\begin{aligned} I\left(X;Y\right) &= H\left(Y\right) - H\left(Y|X\right) \\ &\overset{(a)}{=} H\left(Y\right) - H\left(Y \oplus X|X\right) \\ &\overset{(b)}{=} H\left(Y\right) - H\left(Z|X\right) \end{aligned}$$

$$\begin{aligned} &\overset{(c)}{=} \quad H(Y) - H(Z) \\ &\overset{(d)}{\leq} \quad 1 - H_b(\delta) \end{aligned} \tag{13}$$

Where:

- $H_2(\delta)$ is the binary entropy function, i.e.:

$$H_2(\delta) \overset{\triangle}{=} -\delta \log(\delta) - (1-\delta)\log(1-\delta) \tag{14}$$

- (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 $H(Y) \leq \log|\mathcal{Y}|$.

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

Special cases: Denote by $C(\delta)$, the BSC capacity

- When $\delta = 0$, the channel is clean and $C(0) = 1$.
- When $\delta = 0.5$, for all input distributions, the input and the output and are independent. Therefore, the mutual information and the capacity are equal to zero.
- When $\delta = 1$, each input is always flipped. However, the decoder is able to produce X by inverting the output Y. Thus, $C(1) = 1$.

Puzzle: Show from the operational definitions only that the capacity of the BSC satisfy $C(\delta) = C(1-\delta)$.

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

III. Proof of the capacity theorem

In this section we prove the capacity theorem, i.e., Theorem 1. In order to so we need to prove two directions. First, that a rate larger then $\max_{P_X} I(X;Y)$ is not achievable (this is called converse). In other words we prove that the capacity is upper bounded by $\max_{P_X} I(X;Y)$. This is done in subsection III-A. Then we show that a rate that is lower then $I(X;Y)$ is achievable, in other words, we prove that the capacity is lower bounded by $I(X;Y)$. This is called achvability proof and is described in Subsection III-C. These two parts yield Theorem 1. For the converse part the main tool is Fano's inequality and for the achievability part we use joint weak typicality which is explained in III-B.

A. Proof of the converse for the capacity theorem

For the converse proof of the Coding Theorem, we will need a technical lemma that describes the probabilistic relations between inputs and outputs.

Lemma 1 (Memoryless channel without feedback) For a memoryless channel (without feedback),

$$P(y^n|x^n,m)=\prod_{i=1}^{n}P(y_i|x_i) \tag{15}$$

Remark: When no feedback is available, a memoryless channel can also be defined by (15).

Proof of Lemma 1: Consider the following chain of equalities,

$$\begin{aligned} P(y^n|x^n,m) &= \frac{P(y^n,x^n,m)}{P(x^n,m)} \\ &= \frac{P(m)\prod_{i=1}^{n}P(y_i,x_i|y^{i-1},x^{i-1},m)}{P(x^n,m)} \\ &\overset{(a)}{=} \frac{P(m)\prod_{i=1}^{n}P(x_i|y^{i-1},x^{i-1},m)P(y_i|x_i,x^{i-1},y^{i-1},m)}{P(x^n,m)} \\ &\overset{(b)}{=} \frac{P(m)\prod_{i=1}^{n}P(x_i|x^{i-1},m)P(y_i|x_i)}{P(x^n,m)} \\ &\overset{(c)}{=} \frac{P(m)P(x^n|m)\prod_{i=1}^{n}P(y_i|x_i)}{P(x^n,m)} \\ &= \prod_{i=1}^{n}P(y_i|x_i) \end{aligned} \tag{16}$$

Where:

(a) Follows from the chain rule.

(b) Follows from the memoryless property in (2) and the Markov chain $X_i-(X^{i-1},M)-Y^{i-1}$.

(c) Follows from the probability chain rule. ■

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

$$R \leq C^I \triangleq \max_{P(X)} I\left(X;Y\right). \tag{17}$$

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:

$$\begin{aligned}
nR &\overset{(a)}{=} H\left(M\right) \\
&= H\left(M\right) + H\left(M|Y^n\right) - H\left(M|Y^n\right) \\
&\overset{(b)}{=} I\left(M;Y^n\right) + H\left(M|Y^n\right) \\
&\overset{(c)}{=} I\left(M;Y^n\right) + H\left(M|Y^n, \hat{M}\right) \\
&\overset{(d)}{\leq} I\left(M;Y^n\right) + H\left(M|\hat{M}\right) \\
&\overset{(e)}{\leq} I\left(M;Y^n\right) + \left(1 + P_e \cdot nR\right) \\
&\overset{(f)}{=} I\left(M;Y^n\right) + n \cdot \epsilon_n \\
&\overset{(g)}{=} H\left(Y^n\right) - H\left(Y^n|X^n, M\right) + n \cdot \epsilon_n \\
&\overset{(h)}{=} \sum_{i=1}^{n} \left[H\left(Y_i|Y^{i-1}\right) - H\left(Y_i|X_i\right)\right] + n \cdot \epsilon_n \qquad (18) \\
&\overset{(i)}{\leq} \sum_{i=1}^{n} \left[H\left(Y_i\right) - H\left(Y_i|X_i\right)\right] + n \cdot \epsilon_n \qquad (19) \\
&= \sum_{i=1}^{n} I\left(Y_i;X_i\right) + n \cdot \epsilon_n \qquad (20) \\
&\overset{(j)}{\leq} n \cdot C^I + n \cdot \epsilon_n \qquad (21)
\end{aligned}$$

Where:

(a) The message distributed uniformly on $\left(1, 2, ..., 2^{nR}\right)$. Thus, $H\left(M\right) = \log|\mathcal{M}| = nR$.

(b) Definition of Mutual Information

(c) Since $\hat{M} = g\left(Y^n\right)$ is a deterministic function of Y^n.

(d) Conditioning reduces entropy.

(e) Follows from Fano's inequality: $H\left(M|\hat{M}\right) \leq 1 + P\left(M \neq \hat{M}\right)\log|\mathcal{M}| = 1 + P_e nR$.

(f) follows by defining $\epsilon_n = \frac{1}{n} + P_e R$

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

(h) Follows from Lemma 1.

(i) Follows from conditioning reduced entropy.

(j) Follows by taking the maximum over $P(x_1)P(x_2),\ldots,P(x_n)$. The optimization problem is then equal to C^I for all $i = 1,\ldots,n$.

Now, dividing both sides of the equation by n, we have

$$R \leq C^I + \epsilon_n. \tag{22}$$

If R is an achievabale rate, then there exists a sequence of codes $(n, 2^{nR})$ such that $P_e^{(n)} \rightarrow 0$ which implies $\epsilon_n \rightarrow 0$. Therefore we obtained that if R is achievable, then $R \leq C^I$, and this completes the proof. ■

We will now give a brief subsection on joint typicality. This tool will be used in the achievability part, where we show the existence of a code with rate that is arbitrarily close to C^I and attains vanishing probability of error.

B. Joint weak typicality

Definition 6 (Weak typicality) The set of $\epsilon-$jointly typical sequences with respect to $P_{X,Y}$ is defined by

$$A_\epsilon^{(n)} = \{(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n :| -\frac{1}{n}\log p\left(x^n\right) - H\left(X\right)| \leq \epsilon, \tag{23}$$

$$| -\frac{1}{n}\log p\left(y^n\right) - H\left(Y\right)| \leq \epsilon, \tag{24}$$

$$| -\frac{1}{n}\log p\left(x^n, y^n\right) - H\left(X, Y\right)| \leq \epsilon\}, \tag{25}$$

where $p(x^n, y^n) = \prod_{i=1}^n P_{X,Y}(x_i, y_i)$

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

1) $\lim_{n\rightarrow\infty}\Pr\{(X^n, Y^n) \in A_\epsilon^n\} = 1$

2) $|A_\epsilon^n| \leq 2^{n\cdot(H(X,Y)+\epsilon)}$

3) If $\tilde{X^n}, \tilde{Y^n} : \tilde{X^n} \overset{(i.i.d)}{\sim} P_X(x), \tilde{Y^n} \overset{(i.i.d)}{\sim} P_Y(y), \left(\tilde{X^n}, \tilde{Y^n}\right) \sim P_{X^n}(x^n) \cdot P_{Y^n}(y^n)$, then

$$\Pr\left\{(\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^n\right\} \leq 2^{-n(I(X,Y)-3\cdot\epsilon)}. \tag{26}$$

Proof:

1) follows from the weak law of large numbers.
2) follows from:

$$\begin{aligned} 1 &= \sum_{x^n, y^n} P_{X^n,Y^n}(x^n, y^n) \\ &\overset{(a)}{\geq} \sum_{(x^n,y^n)\in A_\epsilon^n} P_{X^n,Y^n}(x^n, y^n) \\ &\overset{(b)}{\geq} \sum_{(x^n,y^n)\in A_\epsilon^n} 2^{-n\cdot(H(X,Y)+\epsilon)} \\ &= |A_\epsilon^n| \cdot 2^{-n\cdot(H(X,Y)+\epsilon)} \end{aligned} \tag{27}$$

(a) follows from decreasing the number of summed elements;

(b) Follows from (25).

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

$$\begin{aligned} P\left[\left(\tilde{X}^n, \tilde{Y}^n\right) \in A_\epsilon^n\right] &= \sum_{(x^n,y^n)\in A_\epsilon^n} P_{\tilde{X}^n,\tilde{Y}^n}(x^n, y^n) \\ &= \sum_{(x^n,y^n)\in A_\epsilon^n} P_{X^n}(x^n) P_{Y^n}(y^n) \\ &\overset{(a)}{\leq} |A_\epsilon^n| \cdot 2^{-n(H(X)-\epsilon)} \cdot 2^{-n(H(Y)-\epsilon)} \\ &\overset{(b)}{\leq} 2^{n(H(X,Y)+\epsilon)} \cdot 2^{-n(H(X)-\epsilon)} \cdot 2^{-n(H(Y)-\epsilon)} \\ &\overset{(c)}{=} 2^{-n(I(X;Y)-3\epsilon)} \end{aligned} \tag{28}$$

Where:

- (a) - Using the bound on $P_{X^n}(x^n) =$ and $P_{Y^n}(y^n)$ according to (23) and (24).

- (b) - Using the bound on the size of the set, i.e., $|A_\epsilon^n| \leq 2^{n\cdot(H(X,Y)+\epsilon)}$.
- (c) - Follows for the definition of mutual information.

■

C. Proof of the Achievability

We now prove that if $R < C^I$ then for a DMC there exists a sequence of codes $(2^n, n)$ such that $P_{er}^{(n)} \underset{n\rightarrow\infty}{\longrightarrow} 0$. The proof is based on a *random coding*; the code is generated randomly, and then we show that the expectation (over all codebooks) of the error probability goes to zero. Since the expected value goes to zero, there exists at least one code for which the probability of error goes to zero.

Proof of the 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 $X^n\left(i\right)$, where $X^n\left(i\right)$ is the codeword associated with message i, and: $X^n\left(i\right) \overset{i.i.d}{\sim} P_X(x), i = 1, \dots, 2^{nR}$. Reveal the codebook to the encoder and the decoder.

Encoder: For a message i, transmit $X^n(i)$, that is, the ith codeword in $\mathcal{C}$.

Decoder: The decoder receives Y^n, and looks for all $X^n \in \mathcal{C}$ such that: $(X^n, Y^n) \in A_\epsilon^n$. If there is a unique codeword, $X^n(i)$ that is typical, it declares $\hat{m} = i$. In case no message or more than one message are jointly typical with Y^n, it declares an error.

Analysis of error: Consider

$$\begin{aligned} P_e &= P\left(M \neq \hat{M}\right) \\ &= \sum_{m=1}^{2^{nR}} P\left(M = m\right) P\left(M \neq \hat{M} | M = m\right) \\ &= \frac{1}{2^{nR}} \sum_{m=1}^{2^{nR}} P\left(M \neq \hat{M} | M = m\right), \end{aligned}$$

where the last equality holds since the message is distritbuted uniformly. Because of the symmetry in the code construction (with respect to messages), we may assume that $m = 1$, and analyze $P\left(M \neq \hat{M} | M = 1\right)$.

Recall that an error occurs if :

1) $E_1 = (X^n(1), Y^n) \notin A_\epsilon^n$

2) $E_j = (X^n(j), Y^n) \in A_\epsilon^{(n)}$.

- The probability of E_1 tends to zero from Theorem 2.
- Consider the probability of the second error event,

$$\begin{aligned} P_{E_2} &= P\left(\bigcup_{j=2}^{2^{nR}} E_j\right) \\ &\overset{(a)}{\leq} \sum_{j=2}^{2^{nR}} P\left(E_j\right) \\ &\overset{(b)}{\leq} 2^{nR} 2^{-n(I(X;Y)-3\epsilon)} = \\ &= 2^{n(R-I(X;Y)+3\epsilon)} \end{aligned} \tag{29}$$

Where:

(a) Follows from the union bound, e.g. , $P\left(A \cup B\right) \leq P\left(A\right) + P\left(B\right)$.

(b) Follows from Theorem 2.

Hence, if $R < I\left(X;Y\right)$, then $P_{Error2} \overrightarrow{{}_{n\to\infty}} 0$. Note that we showed that $\mathbb{E}_{\mathcal{C}}[P_e^{(n)}] \to 0$, but we need to show that there exists a single code code with $P_e \to 0$. This is settled from the fact that if the expectation of a R.V. goes to zero, then there exists an instance of the R.V. that goes to zero as well. ■

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.:

$$\frac{2^{nR}}{2} = 2^{n\cdot\left(R-\frac{1}{n}\right)} \tag{30}$$

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 2ϵ (Otherwise, these codewords themselves would contribute more than 2ϵ to the sum, and P_{er} would be greater than ϵ). 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 1 (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 ϵ 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.

Where '?' stands for an erased bit.

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:

$$I(X;Y) = H(X) - H(X|Y) \tag{31}$$

$$= H(X) - \sum_{\psi \in Y} P(y = \psi) H(X|y = \psi) \tag{32}$$

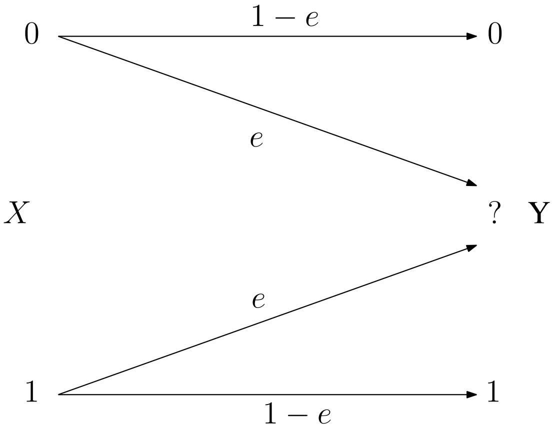

Fig. 6. Binary Erasure Channel

$$= H(X) - P(y=0)H(X|y=0) - P(y=1)H(X|y=1) - P(y='?')H(X|y='?') \tag{33}$$

Where '?' stands for an erased bit (See Figure 6).

Since the model of the channel suggests that for a successfully recieved bit we know the sent bit with probability of 1, meaning X is determined by Y (given $y \neq '?'$), 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 $H(X|Y='?') = H(X)$, and $P(y='?') = e$ by symmetry. Substituting this into Eq. 31:

$$I(X;Y) = H(X) - P(y='?')H(X|y='?') = H(X) - e \cdot H(X) = (1-e)H(X) \tag{34}$$

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

$$C = \sup_{P_X} I(X;Y) = (1-e)\sup_{P_X} H(X) = 1-e \tag{35}$$

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 recieve 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:

$$n + en + e^2n + \cdots = \sum_{j=0}^{\infty} e^j n = n \cdot \sum_{j=0}^{\infty} e^j = n \cdot \frac{1}{1-e} = \frac{n}{1-e} \tag{36}$$

In order to successfully recieve n bits.

Definition 7 (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 $R = \frac{n}{n/(1-e)} = 1 - e$