Information Theory (5XSE0)
Ch.4: Channel Capacity

Hamdi Joudeh

TU/e (Q3 2020-2021)

Reading

- Cover & Thomas, Ch. 7 (excluding 7.13).

Before we start

- Recall that a sequence $(X_1, X_2, \ldots, X_n)$ is denoted by X^n. A sub-sequence of X^n comprising the first i elements is denoted by X^i. The sub-sequence X^{i-1} is empty when $i = 1$.
- Recall the joint AEP. For any (X^n, Y^n) drawn according to $p(x^n, y^n) = \prod_{i=1}^n p(x_i, y_i)$, we have

$$\mathbb{P}\left\{(X^n, Y^n) \in \mathcal{A}_\epsilon^{(n)}\right\} \geq 1 - \epsilon, \text{ for sufficiently large } n \tag{1}$$

  where $\mathcal{A}_\epsilon^{(n)}$ is the jointly typical set. On the other hand, for any independent pair of sequences $(\tilde{X}^n, \tilde{Y}^n) \sim p(x^n)p(y^n)$ with the same marginal distributions as (X^n, Y^n), we have

$$\mathbb{P}\left\{(\tilde{X}^n, \tilde{Y}^n) \in \mathcal{A}_\epsilon^{(n)}\right\} \leq 2^{-n(I(X;Y)-3\epsilon)}. \tag{2}$$

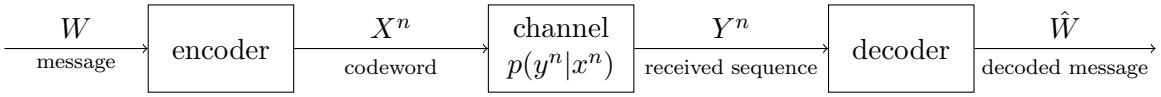
- Recall the data processing inequality. For a Markov chain $W \to X^n \to Y^n \to \hat{W}$, we have

$$I(W; \hat{W}) \leq I(W; Y^n) \leq I(X^n; Y^n). \tag{3}$$

- Recall Fano's inequality. Suppose that we wish to estimate W, taking values on $\mathcal{W}$, from $\hat{W}$. Let $P_e = \mathbb{P}\{W \neq \hat{W}\}$ be the expected probability of error. A bound on P_e is given by

$$H(W|\hat{W}) \leq 1 + P_e \log |\mathcal{W}|. \tag{4}$$

In this chapter we study the problem of communication over a noisy channel (see the below diagram).



On one end of a noisy channel, we have a message W chosen at random from an index set $\{1, 2, \ldots, M\}$. We wish to communicate the value of W over the channel and recover it at the other end. The channel takes a sequence of symbols X^n as an input and produces an output sequence Y^n. Given a specific input sequence x^n, the channel produces one of possibly many output sequences according to a distribution $p(y^n|x^n)$. This probabilistic mapping models uncertainty in the transmission due to noise. W is *encoded* into X^n on one end, and the corresponding Y^n is *decoded* into $\hat{W}$ on the other end. Communication is successful if $W = \hat{W}$, and a decoding error occurs otherwise. The efficiency of the communication (or the rate) is given by $R = \frac{\log M}{n}$, measured in bits per channel symbol.

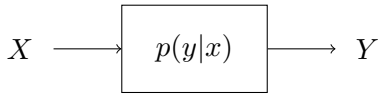
We wish to increase efficiency by making R as large as possible. At the same time, we wish to make the decoding error probability small to guarantee reliable communication. Given a negligibly small decoding error probability, how large can R be? This question is answered in this chapter.

1 Discrete Memoryless Channels and Information Capacity

In our treatment of the above described problem, we focus on a class of channels known as *discrete memoryless channels* (DMCs). A DMC is a system characterized by three main components:

- A discrete input alphabet $\mathcal{X}$.
- A discrete output alphabet $\mathcal{Y}$.
- A collection of conditional pmfs $p(y|x)$, one for each input $x \in \mathcal{X}$. Note that $p(y|x) \geq 0$ for every $(x, y) \in \mathcal{X} \times \mathcal{Y}$, and $\sum_{y \in \mathcal{Y}} p(y|x) = 1$ for every $x \in \mathcal{X}$.

When used once (i.e. one channel symbol), a DMC takes an input random variable X defined on $\mathcal{X}$, and produces an output random variable Y defined on $\mathcal{Y}$. The mapping from X to Y is probabilistic and is governed by the *transition probabilities* $p(y|x)$. These probabilities are fixed, e.g. chosen by nature.



Transition probabilities may create confusion about the input at the output, capturing the effects of noise. To see this, consider a binary channel with $\mathcal{X} = \mathcal{Y} = \{0, 1\}$, and first let's assume that

$$p(y|x) = \begin{cases} 1 & y = x \\ 0, & y \neq x. \end{cases}$$

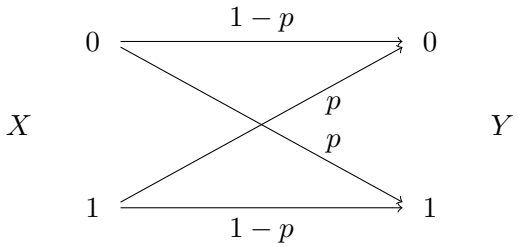
In this case, the input is produced exactly at the output, i.e. $Y = X$. This is an example of a noiseless channel where no confusion about the input occurs when observing the output. A noisy version of this channel is known as the binary symmetric channel, defined as follows.

Definition 1. A binary symmetric channel (BSC) is characterized by binary input and output alphabets $\mathcal{X} = \mathcal{Y} = \{0, 1\}$, and transition probabilities given by

$$p(y|x) = \begin{cases} 1 - p, & y = x \\ p, & y \neq x \end{cases} \tag{5}$$

where $p \in [0, 1]$ is a parameter of the BSC known as the *crossover probability*.

Given an input x, a BSC produces an equal output $y = x$ with probability $1 - p$, and flips the input to produce an output[1] $y = x \oplus 1$ with probability p. The BSC is illustrated as follows



Note that by taking $p = 0$, the BSC becomes the noiseless channel discussed above.

In the BSC, input bits are corrupted by noise which occasionally flips a 1 into a 0 and a 0 into a 1. In some real-world channels, bits are not necessarily corrupted but are instead completely lost. A simple channel exhibiting this *erasure* behavior is defined as follows.

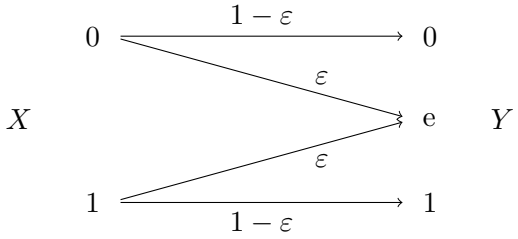
[1]The operation $\oplus$ denote the modulo 2 addition (XOR). That is, $0 \oplus 0 = 1 \oplus 1 = 0$ and $0 \oplus 1 = 1 \oplus 0 = 1$.

Definition 2. A binary erasure channel (BEC) is characterized by alphabets $\mathcal{X} = \{0, 1\}$ and $\mathcal{Y} = \{0, \mathrm{e}, 1\}$, where the outcome e is known as an *erasure*. Transition probabilities are given by

$$p(y|x) = \begin{cases} 1-\varepsilon, & y = x \\ \varepsilon, & y = \mathrm{e} \end{cases}$$

where $\varepsilon \in [0, 1]$ is a parameter of the BEC known as the *erasure probability*.

Given an input x, a BEC produces an equal output $y = x$ with probability $1-\varepsilon$, and an erasure $y = \mathrm{e}$ with probability ε. This illustrated as follows



Note that there is no crossover in the BEC, and when an erasure occurs the input bit is lost.

1.1 Information capacity

How much information can be transferred over a DMC? Suppose that we have an input X with a pmf of $p(x)$, referred to as an *input distribution*. This input induces an output Y with an *output distribution* given by $p(y) = \sum_{x \in \mathcal{X}} p(y|x)p(x)$. The mutual information between X and Y is given by

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X).$$

Recall that $I(X;Y)$ is interpreted as the amount of information about the input X contained in the output Y (or the reduction in uncertainty about X due to the observation of Y). Therefore, $I(X;Y)$ can be thought of as the amount of information transferred (or communicated) over the channel.

For instance, consider the scenario where a transmitter has a piece of information and wishes to communicate it to a receiver through a medium modeled by a DMC. The transmitter sends an input signal X, while on the other end the receiver observes a noisy output signal Y and wishes to learn the value of X from Y. Before observing the output, a receiver's uncertainty about the input is $H(X)$. This is reduced to $H(X|Y)$ after observing the output Y. The difference $H(X) - H(X|Y) = I(X;Y)$ is the amount of information that has been communicated over the channel. Since the channel $p(y|x)$ is fixed by nature and cannot be controlled, $I(X;Y)$ in this case can be increased only by controlling the input distribution $p(x)$. The information capacity is defined as the maximum possible amount of transferred information $I(X;Y)$, attained by choosing an optimum input distribution $p(x)$.

Definition 3. The information capacity of a DMC is defined as

$$C \equiv \max_{p(x)} I(X;Y) \tag{6}$$

where the maximization is over all possible input distributions on $\mathcal{X}$.

The above definition of C is intuitive, but does not yet carry any operational meaning. This is why we call it the *information capacity*, as opposed to the *operational capacity* which will be defined further on. We have previously introduced entropy $H(X)$ as an intuitive measure of information, and postulated that it is a good operational measure of information. This postulate was confirmed through the source coding theorem (data compression), where $H(X)$ turned out to coincide with the least number of bits per symbol required to describe an i.i.d. sequence X^n. We postulate that C carries a similar operational significance, which will be confirmed once we discuss the channel coding theorem. For now, we drop "information" and simply refer to C as *capacity*, and we focus on calculating C for some basic channels.

Example 1. (BSC capacity). Here we calculate the capacity of the BSC in Definition 1. We find an upper bound on the mutual information $I(X;Y)$, which holds for any input distribution[2] $p_X(x)$. We then find an input distribution which attains this upper bound. This is carried out as follows:

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H(Y) - \big[p_X(0)H(Y|X=0) + p_X(1)H(Y|X=1)\big] \\ &= H(Y) - H(p) \qquad (7) \\ &\leq 1 - H(p). \qquad (8) \end{aligned}$$

Recall that $H(p) \equiv -(1-p)\log(1-p) - (p)\log(p)$ is the binary entropy function. (7) holds since by fixing $X = x$, Y is binary with a pmf of $p_{Y|X}(y|x)$, given in (5). Hence $H(Y|X=0) = H(Y|X=1) = H(p)$. The inequality in (8) holds since Y is binary, and therefore its entropy $H(Y)$ is bounded above by 1.

Equality in (8) holds if and only if Y is uniform, i.e. $p_Y(0) = p_Y(1) = 0.5$. This is attained by choosing X to be uniform, i.e. $p_X(0) = p_X(1) = 0.5$ (check this!). Therefore, the capacity achieving input distribution (i.e. optimum $p_X(x)$) is uniform and the capacity in bits is given by

$$C = 1 - H(p).$$

Example 2. (BEC capacity). Here we calculate the capacity of the BEC in Definition 2. We use the same approach of deriving an upper bound for $I(X;Y)$ and then finding $p_X(x)$ that attains this upper bound. First, however, we examine the remaining uncertainty in the input X upon observing the output Y. If the output is 0 or 1, then $X = Y$ and there is no uncertainty about the input. Therefore, we have $H(X|Y=0) = H(X|Y=1) = 0$. If the output is e, then no information about the input is given. In this case we have $H(X|Y=\mathrm{e}) = H(X)$, as observing $Y = \mathrm{e}$ tells us nothing about X (verify this!).

With the above observations in mind, an upper bound is obtained as follows:

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(X) - \big[p_Y(0)H(X|Y=0) + p_Y(1)H(X|Y=1) + p_Y(\mathrm{e})H(X|Y=\mathrm{e})\big] \\ &= H(X) - \varepsilon H(X|Y=\mathrm{e}) \\ &= (1-\varepsilon)H(X) \\ &\leq 1 - \varepsilon. \qquad (9) \end{aligned}$$

The inequality in (9) follows from the fact that X is a binary random variable, and hence its entropy is at most 1. Equality holds whenever X is uniform and the capacity achieving input distribution here is also given by $p_X(0) = p_X(1) = 0.5$. The capacity in bits is hence given by

$$C = 1 - \varepsilon.$$

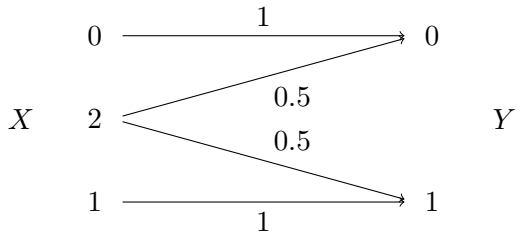
Note that in deriving the capacity of the BEC, we used the expansion $H(X) - H(X|Y)$ of the mutual information as opposed to $H(Y) - H(Y|X)$ used in deriving the capacity of the BSC.

Exercise 1. Derive the capacity of the BEC by expanding $I(X;Y)$ as $H(Y) - H(Y|X)$.

In both above examples, the capacity achieving input distribution is uniform. This is not always the case as we see through the following example.

Example 3. Consider a channel with a ternary input alphabet $\mathcal{X} = \{0,1,2\}$ and a binary output alphabet $\mathcal{Y} = \{0,1\}$. The transition probabilities are shown on the below diagram.

[2] We highlight the variable subscripts in pmfs whenever necessary to avoid confusion.



The capacity of this channel is bounded above by 1 bit. This is seen from $I(X;Y) \leq H(Y) \leq 1$ (why?). This upper bound is attained by choosing an input distribution as follows: $p_X(0) = p_X(1) = 0.5$ and $p_X(2) = 0$. This makes sense as $X = 2$ is a confusable input, and hence we are better off not using it. Note that the capacity achieving distribution here is not uniform on the input alphabet $\mathcal{X}$.

1.2 Transition matrix and symmetric channels

Transition probabilities of a DMC can be arranged into a *transition matrix*, with rows indicating outputs and columns indicating inputs.[3] Without loss of generality, let's assume that the input and output alphabets are given by $\mathcal{X} = \{1, 2, \ldots, k\}$ and $\mathcal{Y} = \{1, 2, \ldots, l\}$, respectively. We use $p_{Y|X}$ to denote the $l \times k$ transition matrix, where the element in y-th row and x-th column is given by $p_{Y|X}(y|x)$.

Now let's arrange an input distribution $p_X(x)$ into a $k \times 1$ (column) vector denote by p_X. Passing X through the channel with transition matrix $p_{Y|X}$, we obtain an output Y with distribution p_Y given by

$$p_Y = p_{Y|X} p_X \tag{10}$$

in which the y-th entry is given by $p_Y(y) = \sum_{x \in \mathcal{X}} p_X(x) p_{Y|X}(y|x)$. This is expressed more elaborately as

$$\begin{bmatrix} p_Y(1) \\ p_Y(2) \\ \vdots \\ p_Y(l) \end{bmatrix} = \begin{bmatrix} p_{Y|X}(1|1) & p_{Y|X}(1|2) & \cdots & p_{Y|X}(1|k) \\ p_{Y|X}(2|1) & p_{Y|X}(2|2) & \cdots & p_{Y|X}(2|k) \\ \vdots & \vdots & \ddots & \vdots \\ p_{Y|X}(l|1) & p_{Y|X}(l|2) & \cdots & p_{Y|X}(l|k) \end{bmatrix} \begin{bmatrix} p_X(1) \\ p_X(2) \\ \vdots \\ p_X(k) \end{bmatrix}.$$

Each column of $p_{Y|X}$ is a (conditional) pmf, and therefore should sum to 1. Rows of $p_{Y|X}$ on the other hand are not pmfs, and hence each row does not necessarily sum to 1.

Example 4. Transition matrices for the BSC and BEC are respectively given by

$$\begin{bmatrix} 1-p & p \\ p & 1-p \end{bmatrix} \text{ and } \begin{bmatrix} 1-\varepsilon & 0 \\ \varepsilon & \varepsilon \\ 0 & 1-\varepsilon \end{bmatrix}.$$

We now introduce classes of channels based on their channel transition matrices.

Definition 4. A DMC is *symmetric* if the channel transition matrix $p_{Y|X}$ satisfies the following:

- Columns $p_{Y|X}(y|1), p_{Y|X}(y|2), \ldots, p_{Y|X}(y|k)$ are permutations of each other.
- Row $p_{Y|X}(1|x), p_{Y|X}(2|x), \ldots, p_{Y|X}(l|x)$ are permutations of each other.

We can see that the BEC is not symmetric according to the above definition. The BSC, however, is symmetric (and hence the name). The concept of symmetric channels can be generalized as follows.

Definition 5. A DMC is *weakly symmetric* if the channel transition matrix $p_{Y|X}$ satisfies the following:

[3] Note that this is the other way around in Cover & Thomas, where rows indicate inputs and columns indicate outputs. The transition matrix in Cover & Thomas is the transpose of the transition matrix in these notes.

- Columns $p_{Y|X}(y|1), p_{Y|X}(y|2), \ldots, p_{Y|X}(y|k)$ are permutations of each other.
- Sums of all rows are equal, that is $\sum_{x=1}^{k} p_{Y|X}(y|x) = c$ for every $y \in \mathcal{Y}$.

Note that the BEC is weakly symmetric if $\varepsilon = 1/3$, however, it is not weakly symmetric for other values of ε. Moreover, a symmetric channel is weakly symmetric (verify this!). Therefore, any property that holds for weakly symmetric channels also holds for symmetric channels. A nice property of weakly symmetric channels is that their capacity is easily computed as follows.

Theorem 1. *For any weakly symmetric DMC with a transition matrix $p_{Y|X}$, the capacity is given by*

$$C = \log |\mathcal{Y}| - H\big(p_{Y|X}(y|1)\big) \tag{11}$$

where $H\big(p_{Y|X}(y|1)\big)$ is the entropy of the pmf $p_{Y|X}(y|1)$ (i.e. the first column in the transition matrix). Moreover, the capacity achieving input distribution is uniform on the input alphabet $\mathcal{X}$.

Proof. First, note that $H(Y|X=1) = H\big(p_{Y|X}(y|1)\big)$, where the pmf $p_{Y|X}(y|1)$ is also the first column in the transition matrix. It is also true that $H(Y|X=1) = H(Y|X=x)$ for any $x \in \mathcal{X}$, since columns $p_{Y|X}(y|1), p_{Y|X}(y|2), \ldots, p_{Y|X}(y|k)$ of the transition matrix are all permutations of each other in weakly symmetric channels. Therefore, we have

$$H(Y|X) = \sum_{x \in \mathcal{X}} p(x) H(Y|X=x) = H\big(p_{Y|X}(y|1)\big). \tag{12}$$

Now we bound the mutual information $I(X;Y)$ as follows:

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &\leq \log |\mathcal{Y}| - H\big(p_{Y|X}(y|1)\big) \end{aligned}$$

where equality holds if and only if Y is uniform. For weakly symmetric channel, a uniform output Y is obtained from a uniform input X as follows:

$$p_Y(y) = \sum_{x \in \mathcal{X}} p_{Y|X}(y|x) p_X(x) = \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} p_{Y|X}(y|x) = \frac{c}{|\mathcal{X}|}$$

where the last equality follows from the fact that $\sum_{x=1}^{k} p_{Y|X}(y|x) = c$ for every $y \in \mathcal{Y}$, i.e. sums of rows are equal in weakly symmetric channels. It follows that the capacity is equal to $\log |\mathcal{Y}| - H\big(p_{Y|X}(y|1)\big)$, and the capacity achieving input distribution is uniform on $\mathcal{X}$. □

Exercise 2. A generalization of the BSC is the q-ary symmetric channel. This is characterized by input and output alphabets $\mathcal{X} = \mathcal{Y} = \{0, 1, \ldots, q-1\}$, and a transition matrix

$$p_{Y|X} = \begin{bmatrix} 1-p & p/(q-1) & \cdots & p/(q-1) \\ p/(q-1) & 1-p & \cdots & p/(q-1) \\ \vdots & \vdots & \ddots & \vdots \\ p/(q-1) & p/(q-1) & \cdots & 1-p \end{bmatrix}$$

where $p \in [0,1]$ is a fixed parameter. Find the capacity of this channel.

Exercise 3. Consider a channel with an input X, output Y and noise Z, where all variables are q-ary: $\mathcal{X} = \mathcal{Y} = \mathcal{Z} = \{0, 1, \ldots, q-1\}$, for some integer $q > 0$. The output is given by the input plus noise as:

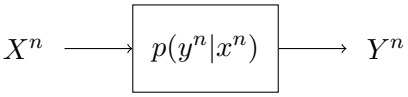
$$Y = X + Z \mod q$$

The noise Z has a distribution $p_Z(z) = p_z$. Write the transition matrix of this channel and calculate its capacity in terms of the noise distribution. How does this channel relate to the one in Exercise 2?

1.3 Multiple channel uses and memorylessness

We have seen the "discrete" part (alphabets) and the "channel" part (transition probabilities) of DMCs, but what about the "memoryless" part? This feature kicks in when the channel is used several times in succession, which is the case in most communication scenarios of interest.

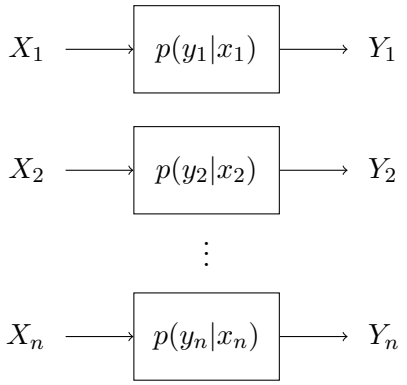
For instance, consider a digital communication system in which the input X is a modulated symbol (e.g. PAM or QAM) sent by a transmitter, and the output Y is a corrupted version of X (see, e.g., Exercise 3). In such scenario, and many others, a transmitter never sends one symbol in isolation, and instead sends a sequence of symbols $X^n = (X_1, X_2, \ldots, X_n)$ in a communication session spanning n uses of the channel, or time instances (e.g. a WiFi packet). This induces a corresponding output sequence $Y^n = (Y_1, Y_2, \ldots, Y_n)$ at the other end of the channel. In general discrete channels (not necessarily memoryless), transition probabilities that map X^n to Y^n take the form $p(y^n|x^n)$.



In DMCs with no feedback (defined below), transition probabilities $p(y^n|x^n)$ factorize as follows:

$$p(y^n|x^n) = \prod_{i=1}^{n} p(y_i|x_i). \tag{13}$$

That is, for any time instance $i \in \{1, \ldots, n\}$, the output Y_i depends only on the input at the time X_i, and is conditionally independent of everything else. This is represented by the following diagram.



The property in (13) greatly simplifies analysis. We now provide formal definitions of memorylessness and no feedback. Note that a sequence $y^{i-1} \equiv (y_1, y_2, \ldots, y_{i-1})$ is empty for $i = 1$.

Definition 6. Consider n uses of a discrete channel. The channel is said to be memoryless if for every use $i \in \{1, 2, \ldots, n\}$ (i.e. time instance), we have

$$p(y_i|x^i, y^{i-1}) = p(y_i|x_i). \tag{14}$$

In words: the probability distribution of the output Y_i is only dependent on the input at the time X_i, and is conditionally independent of previous outputs Y^{i-1}.

Note that the above memorylessness property is equivalent to the following Markov property

$$(X^{i-1}, Y^{i-1}) \to X_i \to Y_i.$$

Definition 7. In a DMC with no feedback, for every $i \in \{1, 2, \ldots, n\}$ we have the following:

$$p(x_i|x^{i-1}, y^{i-1}) = p(x_i|x^{i-1}).$$

In words: X_i may only depend on previous inputs X^{i-1}, and is conditionally independent of previous outputs Y^{i-1}. This represents scenarios with no feedback link between the decoder and the encoder.

Exercise 4. Show that memorylessness (Definition 6) and no feedback (Definition 7) together imply (13).

Lemma 1. *Consider n channel uses of a DMC with no feedback and with capacity C, and let X^n and Y^n be the corresponding input and output sequences. Then for any input sequence distribution $p(x^n)$, we have*

$$\frac{1}{n} I(X^n; Y^n) \leq C. \tag{15}$$

Equality holds when X^n is i.i.d. drawn from the capacity achieving distribution in Definition 3.

Intuition. When $n > 1$, C is interpreted as the capacity in *bits per channel use* (or per channel symbol). Lemma 1 shows that using a DMC several times does not increase the capacity (per channel symbol).

Proof. $I(X^n; Y^n)$ is bounded above as follows:

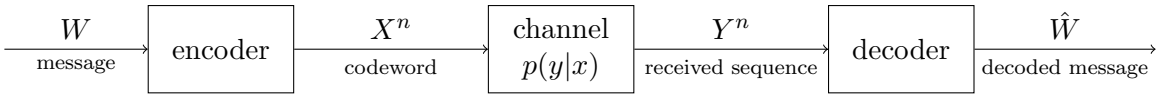
$$\begin{aligned}
I(X^n; Y^n) &= H(Y^n) - H(Y^n|X^n) \\
&= H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y^{i-1}, X^n) \\
&= H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_i) && (16) \\
&\leq \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_i) && (17) \\
&= \sum_{i=1}^{n} I(X_i; Y_i) \\
&\leq nC. && (18)
\end{aligned}$$

The key step is the equality in (16), which holds due to (13). The inequality in (17) follows from the independence bound on entropy, while (18) holds since each term $I(X_i; Y_i)$ is individually bounded above by C. Verifying conditions for equality to hold in (15) is left as an exercise. □

A main takeaway point from this part is that even if a DMC is used multiple times in succession, its behaviour is fully characterized by the single-use triple $(\mathcal{X}, p(y|x), \mathcal{Y})$.

2 Operational Capacity and the Channel Coding Theorem

Here we give an operational definition of capacity, while focusing on DMCs with no feedback. We define a communication protocol that includes a message, an encoder (or transmitter) and a decoder (receiver), as seen in the diagram below. Note that the channel is reduced from $p(y^n|x^n)$ to $p(y|x)$, which fully describes the transition probabilities of a DMC with multiple channel uses.



Definition 8. An (M, n) code for a DMC characterized by $(\mathcal{X}, p(y|x), \mathcal{Y})$ consists of:

- An index set $\mathcal{W} \equiv \{1, 2, \ldots, M\}$.
- An encoding function $x^n : \mathcal{W} \to \mathcal{X}^n$ that maps each index w to a distinct *codeword* $x^n(w)$, which is a sequence in $\mathcal{X}^n$. The set of M codewords $\{x^n(1), x^n(2), \ldots, x^n(M)\}$ is called a *codebook*.
- A decoding function $g : \mathcal{Y}^n \to \mathcal{W}$, which is a deterministic rule that assigns an index $g(y^n)$ from $\mathcal{W}$ to each possible received sequence $y^n \in \mathcal{Y}^n$.

The above code is used as follows. We have a message W, which is a random variable uniformly distributed on $\mathcal{W}$. The message is encoded into a codeword $X^n = x^n(W)$, which serves as an input to the DMC. Since W is a random variable, then X^n is a random variable as well. The codeword X^n is transmitted and the channel induces an output sequence Y^n. An estimate $\hat{W}$ of W is obtained using the decoding rule $\hat{W} = g(Y^n)$. An error occurs if $\hat{W} \neq W$, which brings us to the definition of error probability.

Given that $W = w$ and hence the transmitted codeword is $X^n = x^n(w)$, a decoding error occurs if $g(Y^n) \neq w$. The probability of this event is called a *conditional probability of error*, and is defined as[4]

$$\lambda_w^{(n)} \equiv \mathbb{P}\Big\{g(Y^n) \neq w \mid X^n = x^n(w)\Big\} = \sum_{y^n \in \mathcal{Y}^n} p(y^n|x^n(w))\mathbb{1}(g(y^n) \neq w). \tag{19}$$

Some codewords might be more confusable than others, leading to a higher conditional probability of error. Reliable communication is guaranteed by ensuring that the conditional probability of error for the *worst* codeword is small. This worst codeword is the one with a maximal probability of error.

Definition 9. For an (M, n) code with conditional probabilities of error $\lambda_1^{(n)}, \lambda_2^{(n)}, \ldots, \lambda_M^{(n)}$, the *maximal probability of error* $\lambda^{(n)}$ is defined as

$$\lambda^{(n)} \equiv \max_{w \in \mathcal{W}} \lambda_w^{(n)}. \tag{20}$$

Having established a measure of *reliability* for an (M, n) code, we now establish a measure of *efficiency*. Since the message W is uniform, its entropy is given by $\log M$. By communicating the value of W, we are communicating $\log M$ bits of information in n uses of the channel. This defines the rate.

Definition 10. The *rate* of an (M, n) code in bits per channel use is defined as

$$R^{(n)} \equiv \frac{\log M}{n}. \tag{21}$$

An (M, n) code is sometimes called a block code: it is used to communicate $\log M$ bits in a block of n channel uses, where n is sometimes called the *block length*. Encoding messages into blocks of channel symbols is done to improve reliability, as seen through the following examples.

Example 5. Consider a BSC with crossover probability $0 < p < 0.5$ (see Definition 1). We wish to transmit a 1-bit message from $\mathcal{W} = \{0, 1\}$ by using the channel once. Since $n = 1$, we omit n from the notation henceforth. The BSC takes binary inputs, therefore the message is transmitted as it is, i.e.

$$x(0) = 0 \quad \text{and} \quad x(1) = 1.$$

The decoder declares that $x(0)$ has been transmitter transmitted if it receives $y = 0$. Otherwise, it declares that $x(1)$ has been transmitted (this is in fact the optimum decoder). Due to symmetry, we have

$$\lambda = \lambda_1 = \lambda_2 = p.$$

In this transmission scenario with $n = 1$, reliable communication is not possible as the probability of error is bounded away zero. There is always a chance of error and the receiver can do nothing about it.

Example 6. (Repetition code) Consider the same setting in the previous example. Here, however, we use a block code of length n with two codewords given by

$$x^n(0) = \overbrace{000\cdots0}^{n \text{ times}} \quad \text{and} \quad x^n(1) = \overbrace{111\cdots1}^{n \text{ times}}.$$

[4] $\mathbb{1}$(statement) is an indicator function, equal to one if the "statement" holds, and equal to zero otherwise.

Upon receiving y^n, the decoder uses a majority vote (which is optimal here), described as follows:

$$g(y^n) = \begin{cases} 0, & \text{if } y^n \text{ contains more zeros than ones} \\ 1, & \text{if } y^n \text{ contains more ones than zeros.} \end{cases}$$

If n is even and there is a tie, the decoder declares an error. Due to symmetry, we also have $\lambda_1^{(n)} = \lambda_2^{(n)}$, and the probability of error is given by (verify this!)

$$\lambda^{(n)} = \sum_{i=\lceil n/2 \rceil}^{n} \binom{n}{i} p^i (1-p)^{n-i}.$$

It can be shown that $\lim_{n\to\infty} \lambda^{(n)} = 0$ (exercise*). Therefore reliable communication is asymptotically achieved in this scenario, since $\lambda^{(n)}$ can be made negligibly small by taking n to be sufficiently large.

For noisy channels as the BSC (and many others), where reliable communication is not possible for small n, Example 6 shows that reliable communication is made possible by increasing n and using channel codes, e.g. a repetition code. However, the price paid in terms of efficiency in Example 6 is huge: since M is fixed, the rate $R^{(n)}$ goes to zero as n goes to infinity. Before the invention of information theory in 1948, it was widely believed that this penalty is inescapable, and that reliable communication over noisy channels in general is only possible if efficiency is entirely sacrificed.

$$\text{Pre-Shannon belief: } \lim_{n\to\infty} \lambda^{(n)} \to 0 \implies \lim_{n\to\infty} R^{(n)} \to 0.$$

Shannon published his landmark paper in 1948 and showed that this common belief is wrong.

2.1 Operational capacity

Suppose that we wish to communicate with an *efficiency* of R bits per channel use. For any number of channel uses n, we shall be using a $\left(\lceil 2^{nR} \rceil, n\right)$ code, as an index set of size of $M = \lceil 2^{nR} \rceil$ guarantees a rate no less than R. We say that *reliability* at rate R is attained if we can make the maximal error $\lambda^{(n)}$ as small as desired by making n large enough. This defines achievable rates as follows.

Definition 11. A rate R is *achievable* if there exists a sequence of $\left(\lceil 2^{nR} \rceil, n\right)$ for which the maximal probability of error $\lambda^{(n)}$ tends to zero as n tends to infinity, i.e.

$$\lim_{n\to\infty} \lambda^{(n)} = 0.$$

As n grows large, we have $\lceil 2^{nR} \rceil \approx 2^{nR}$. Therefore, we use $\left(2^{nR}, n\right)$ when referring to $\left(\lceil 2^{nR} \rceil, n\right)$ codes henceforth. We are now ready to define the operational capacity.

Definition 12. The operational capacity is the supremum of all achievable rates, i.e.

$$C_{\text{op}} \equiv \sup\left\{R : R \text{ is achievable}\right\}.$$

According to the false pre-Shannon belief, C_{op} is zero for channels as the BSC in Example 6.

2.2 The channel coding theorem for discrete memoryless channels

Theorem 2. (Channel coding theorem). *For a DMC with capacity C, the following statements are true:*

- *(Achievability) All rates below capacity C are achievable. That is, for every $R < C$, there exists a sequence of $(2^{nR}, n)$ codes with $\lambda^{(n)} \to 0$ as $n \to \infty$.*
- *(Converse) For any sequence of $(2^{nR}, n)$ codes with $\lambda^{(n)} \to 0$ as $n \to \infty$, we must have $R \leq C$.*

It immediately follows from Theorem 2 that

$$C_{\text{op}} = C \equiv \max_{p(x)} I(X;Y). \tag{22}$$

Going back to the BSC in Example 6, Theorem 2 suggestions that we can do much better than repetition coding and achieve strictly positive rates close to $1 - H(p)$ bits per channel use.

3 Achievability via Random Coding

In this section we focus on proving the following statement:

For every $R < C$, there exists a sequence of $(2^{nR}, n)$ codes with $\lambda^{(n)} \to 0$ as $n \to \infty$.

The key to proving achievability is to focus on showing *existence* of *good codes* that achieve all rate up to capacity, instead of pursuing the *explicit construction* of good codes. We start by defining the average probability of error, which plays an important role in the analysis.

Definition 13. For an (M, n) code with conditional probabilities of error $\lambda_1^{(n)}, \lambda_2^{(n)}, \ldots, \lambda_M^{(n)}$, the average probability of error $P_e^{(n)}$ is defined as

$$P_e^{(n)} \equiv \frac{1}{M} \sum_{w \in \mathcal{W}} \lambda_w^{(n)}. \tag{23}$$

Since W is uniform, it follows that the above arithmetic average coincides with the statistical expectation

$$P_e^{(n)} = \sum_{w \in \mathcal{W}} \mathbb{P}\{W = w\} \lambda_w^{(n)} = \mathbb{P}\{\hat{W} \neq W\}.$$

Since $P_e^{(n)} \leq \lambda^{(n)}$, a small average error probability $P_e^{(n)}$ does not necessarily imply a small maximal error probability $\lambda^{(n)}$. Nevertheless, we will show further on that given a sequence of $(2^{nR}, n)$ codes for which $\lim_{n\to\infty} P_e^{(n)} = 0$, we can construct a sequence of $(2^{nR-1}, n)$ codes for which $\lim_{n\to\infty} \lambda^{(n)} = 0$. Since $\lim_{n\to\infty} \left(R - \frac{1}{n}\right) = R$, these codes essentially have the same rate. Hence we focus on proving the following statement: *For every $R < C$, there exists a sequence of $(2^{nR}, n)$ codes with $P_e^{(n)} \to 0$ as $n \to \infty$.*

3.1 Encoding and Joint Typicality Decoding

Fix an input distribution $p_X(x)$ and generate a sequence x^n by drawing from the i.i.d. distribution

$$p(x^n) = \prod_{i=1}^{n} p_X(x_i). \tag{24}$$

This process is repeated independently to generate 2^{nR} codewords $x^n(1), x^n(2), \ldots, x^n(2^{nR})$. These codewords can be arranged into a $2^{nR} \times n$ codebook matrix $\boldsymbol{c}$ given by

$$\boldsymbol{c} = \begin{bmatrix} x^n(1) \\ x^n(2) \\ \vdots \\ x^n(2^{nR}) \end{bmatrix} = \begin{bmatrix} x_1(1) & x_2(1) & \cdots & x_n(1) \\ x_1(2) & x_2(2) & \cdots & x_n(2) \\ \vdots & \vdots & \ddots & \vdots \\ x_1(2^{nR}) & x_2(2^{nR}) & \cdots & x_n(2^{nR}) \end{bmatrix} \tag{25}$$

in which the w-th row is the codeword corresponding to message index w. Once generated, the codebook $\boldsymbol{c}$ remains fixed and it is revealed to both transmitter (encoder) and receiver (decoder).[5] From (24) and (13), the joint input-output distribution is given by

$$p(x^n, y^n) = \prod_{i=1}^{n} p_{Y|X}(y_i|x_i) p_X(x_i) = \prod_{i=1}^{n} p_{X,Y}(x_i, y_i). \tag{26}$$

[5] In order to communicate efficiently, a transmitter and receiver should speak the same language (codebook = language).

For some fixed $\epsilon > 0$, let $\mathcal{A}_\epsilon^{(n)}$ be the set of jointly typical sequence (x^n, y^n) associated with the above joint input-output distribution. Since distributions $p_X(x)$ and $p_{Y|X}(y|x)$ remain fixed, $\mathcal{A}_\epsilon^{(n)}$ is also fixed and hence it can be revealed to the receiver before any communication takes place.

- **Encoding:** A message W is drawn uniformly at random from $\{1, 2, \ldots, 2^{nR}\}$. Given that $W = w$, the transmitter selects the corresponding codeword $x^n(w)$ and sends it over the channel.

- **Channel:** The receiver observes a random output sequence Y^n, drawn from the distribution

$$p(y^n|x^n(w)) = \prod_{i=1}^{n} p_{Y|X}(y_i|x_i(w)). \tag{27}$$

- **Decoding:** The receiver decodes Y^n into an estimated message $\hat{W} = g(Y^n)$ using joint typicality decoding (described below). $g(Y^n)$ takes values on $\{0\} \cup \{1, 2, 3, \ldots, 2^{nR}\}$, where $g(Y^n) = 0$ is the event that the receiver is unable to decode. Suppose that the receiver observes a given sequence $Y^n = y^n$, then the joint typicality decoding rule is given as follows

$$g(y^n) = \begin{cases} \hat{w}, & \text{if } \left(x^n(\hat{w}), y^n\right) \in \mathcal{A}_\epsilon^{(n)} \text{ and } \left(x^n(w'), y^n\right) \notin \mathcal{A}_\epsilon^{(n)} \text{ for all } w' \neq \hat{w} \\ 0, & \text{otherwise.} \end{cases}$$

  That is, for every possible transmitted index $\hat{w} \in \{1, 2, 3, \ldots, 2^{nR}\}$, the decoder checks whether the pair $(x^n(\hat{w}), y^n)$ is in $\mathcal{A}_\epsilon^{(n)}$, i.e. jointly typical. If there is only one such index $\hat{w}$, we have $g(y^n) = \hat{w}$. Otherwise, if there is no such index that gives a typical pair, or if there is more than one index giving a typical pair, then $g(y^n) = 0$ and the receiver declares that it is unable to decode.

- **Error probability:** Given that $W = w$ is transmitted, the received sequence Y^n is random, and hence the estimate $\hat{W} = g(Y^n)$ is also random. As seen from the decoding rule above, a decoding error occurs if $\left(x^n(w), Y^n\right)$ is not jointly typical or if $\left(x^n(w'), Y^n\right)$ is jointly typical for some (possibly more than one) index w' other than w. The conditional probability of error $\lambda_w^{(n)}$ is given by

$$\lambda_w^{(n)}(\boldsymbol{c}) = \mathbb{P}\left\{\left(x^n(w), Y^n\right) \notin \mathcal{A}_\epsilon^{(n)}, \text{ or } \left(x^n(w'), Y^n\right) \in \mathcal{A}_\epsilon^{(n)} \text{ for some } w' \neq w\right\} \tag{28}$$

  where the dependency on the employed codebook $\boldsymbol{c}$ is highlighted. We know that

$$P_e^{(n)}(\boldsymbol{c}) = \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \lambda_w^{(n)}(\boldsymbol{c}). \tag{29}$$

Analysing $P_e^{(n)}(\boldsymbol{c})$ for a specific $\boldsymbol{c}$ is difficult. Alternatively, we resort to a randomization trick explained as follows. Suppose that we have a collection of codebooks given by $\{\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_K\}$, and let $P_e^{(n)}(\boldsymbol{c}_k)$ be the average error probability of the k-th codebook. We do not know which codebook is best, but we suspect that at least one of them has a "small enough" error probability, no more than $\varepsilon_{\text{desired}}$. We define a *random* codebook $\boldsymbol{C}$, taking values in $\{\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_K\}$ with probabilities $\mathbb{P}\{\boldsymbol{C} = \boldsymbol{c}_1\}, \mathbb{P}\{\boldsymbol{C} = \boldsymbol{c}_2\}, \ldots, \mathbb{P}\{\boldsymbol{C} = \boldsymbol{c}_K\}$. Now suppose that we, somehow, manage to show that the expected probability of error over all codebooks is no more than $\varepsilon_{\text{desired}}$, that is $\sum_{k=1}^{K} \mathbb{P}\{\boldsymbol{C} = \boldsymbol{c}_k\} P_e^{(n)}(\boldsymbol{c}_k) \leq \varepsilon_{\text{desired}}$. This can only happen if at least one codebook in $\{\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_K\}$ has an error probability of no more than $\varepsilon_{\text{desired}}$. This workaround proves the existence of a good code, that achieves a desired performance, without having to specify the code.

3.2 Randomization and existence of good codes

Building on the above logic, we define a $2^{nR} \times n$ *random codebook* matrix $\boldsymbol{C}$, given by

$$\boldsymbol{C} = \begin{bmatrix} X^n(1) \\ X^n(2) \\ \vdots \\ X^n(2^{nR}) \end{bmatrix} = \begin{bmatrix} X_1(1) & X_2(1) & \cdots & X_n(1) \\ X_1(2) & X_2(2) & \cdots & X_n(2) \\ \vdots & \vdots & \ddots & \vdots \\ X_1(2^{nR}) & X_2(2^{nR}) & \cdots & X_n(2^{nR}) \end{bmatrix}$$

where each column is a random codeword, and all codewords are independent and have the same distribution $p(x^n) = \prod_{i=1}^n p_X(x_i)$. Note that $\boldsymbol{c}$ in (25) is one realization of $\boldsymbol{C}$, with a probability

$$\mathbb{P}\{\boldsymbol{C} = \boldsymbol{c}\} = \prod_{w=1}^{2^{nR}} \prod_{i=1}^{n} p_X\big(x_i(w)\big). \tag{30}$$

Given that $\boldsymbol{C} = \boldsymbol{c}$, the error probability of the w-th codeword is equal to $\lambda_w^{(n)}(\boldsymbol{c})$ in (28), and the average over all codewords is given by $P_e^{(n)}(\boldsymbol{c})$ in (29), that is

$$\mathbb{P}\left\{\hat{W} \neq W \mid W = w, \boldsymbol{C} = \boldsymbol{c}\right\} = \lambda_w^{(n)}(\boldsymbol{c}) \quad \text{and} \quad \mathbb{P}\left\{\hat{W} \neq W \mid \boldsymbol{C} = \boldsymbol{c}\right\} = P_e^{(n)}(\boldsymbol{c}).$$

We are interested in the expectation of $P_e^{(n)}(\boldsymbol{C})$ with respect to $\boldsymbol{C}$. This is given by

$$\begin{aligned}
\mathbb{P}\left\{\hat{W} \neq W\right\} &= \sum_{\boldsymbol{c}} \mathbb{P}\left\{\boldsymbol{C} = \boldsymbol{c}\right\} \mathbb{P}\left\{\hat{W} \neq W \mid \boldsymbol{C} = \boldsymbol{c}\right\} \\
&= \sum_{\boldsymbol{c}} \mathbb{P}\left\{\boldsymbol{C} = \boldsymbol{c}\right\} \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \mathbb{P}\left\{\hat{W} \neq W \mid \boldsymbol{C} = \boldsymbol{c}, W = w\right\} \\
&= \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \sum_{\boldsymbol{c}} \mathbb{P}\left\{\boldsymbol{C} = \boldsymbol{c}\right\} \mathbb{P}\left\{\hat{W} \neq W \mid \boldsymbol{C} = \boldsymbol{c}, W = w\right\} \\
&= \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \mathbb{P}\left\{\hat{W} \neq W \mid W = w\right\}
\end{aligned}$$

where $\mathbb{P}\big\{\hat{W} \neq W \mid W = w\big\}$ is the expectation of $\lambda_w^{(n)}(\boldsymbol{C})$ with respect to $\boldsymbol{C}$. Due to the symmetry in $\boldsymbol{C}$, i.e. entries of the matrix are all i.i.d., it turns out that the terms $\mathbb{P}\big\{\hat{W} \neq W \mid W = w\big\}$ are equal for all message indices $w \in \{1, 2, \ldots, 2^{nR}\}$. An alternative way to see this is to suppose that $W = 1$ and proceed to bound $\mathbb{P}\big\{\hat{W} \neq W \mid W = 1\big\}$. We shall see that the result does not depend on the message index.

Now we define the event that the pair $\big(X^n(w), Y^n\big)$ is jointly typical as follows

$$\mathcal{E}_w \equiv \left\{\big(X^n(w), Y^n\big) \in \mathcal{A}_\epsilon^{(n)}\right\}, \text{ for any } w \in \{1, 2, \ldots, 2^{nR}\}.$$

Recall from joint typicality decoding that given $W = 1$, a decoding error occurs if $\big(X^n(1), Y^n\big) \notin \mathcal{A}_\epsilon^{(n)}$ or $\big(X^n(w'), Y^n\big) \in \mathcal{A}_\epsilon^{(n)}$ for any $w' \neq 1$. Therefore, given $W = 1$, the probability of error is

$$\mathbb{P}\big\{\hat{W} \neq W \mid W = 1\big\} = \mathbb{P}\big\{\mathcal{E}_1^{\mathrm{c}} \cup \mathcal{E}_2 \cup \mathcal{E}_3 \cup \cdots \cup \mathcal{E}_{2^{nR}}\big\}$$

where $\mathcal{E}_1^{\mathrm{c}}$ is the complement $\mathcal{E}_1$, i.e. the event that $\big(X^n(1), Y^n\big)$ is not in $\mathcal{A}_\epsilon^{(n)}$. From the union bound (i.e. $\mathbb{P}\{\mathcal{E}_i \cup \mathcal{E}_j\} \leq \mathbb{P}\{\mathcal{E}_i\} + \mathbb{P}\{\mathcal{E}_j\}$), we obtain an upper bound for $\mathbb{P}\big\{\hat{W} \neq W \mid W = 1\big\}$ as

$$\mathbb{P}\big\{\hat{W} \neq W \mid W = 1\big\} \leq \mathbb{P}\big\{\mathcal{E}_1^{\mathrm{c}}\big\} + \sum_{w=2}^{2^{nR}} \mathbb{P}\big\{\mathcal{E}_w\big\}. \tag{31}$$

Next, we bound the probabilities of the events in (31).

- Given that $W = 1$, the pair $\big(X^n(1), Y^n\big)$ is jointly distributed according to $p(x^n, y^n)$ in (26), since the output Y^n is induced by $X^n(1)$. From the joint AEP, we have $\mathbb{P}\big\{(X^n(1), Y^n) \in \mathcal{A}_\epsilon^{(n)}\big\} \geq 1 - \epsilon$ for sufficiently large n (see (1) at the beginning of the chapter). It immediately follows that

$$\mathbb{P}\big\{\mathcal{E}_1^{\mathrm{c}}\big\} = 1 - \mathbb{P}\big\{\mathcal{E}_1\big\} \leq \epsilon, \text{ for large enough } n. \tag{32}$$

- Given that $W = 1$, then $X^n(w)$ and Y^n are independent for any $w \neq 1$. This holds since random codewords are all independent. Therefore, the pair $(X^n(w), Y^n)$ is distributed according to $p(x^n)p(y^n)$ for all $w \neq 1$, i.e. it has the same marginal distributions as $(X^n(1), Y^n)$. From the joint AEP (see (2) at the beginning of the chapter), it immediately follows that

$$\mathbb{P}\big\{\mathcal{E}_w\big\} \leq 2^{-n(I(X;Y)-3\epsilon)}, \text{ for any } w \neq 1. \tag{33}$$

By plugging (32) and (33) into (31), and taking n to be large enough for (32) to hold, we have

$$\begin{aligned}
\mathbb{P}\big\{\hat{W} \neq W \mid W = 1\big\} &\leq \mathbb{P}\big\{\mathcal{E}_1^{\mathrm{c}}\big\} + \sum_{w=2}^{2^{nR}} \mathbb{P}\big\{\mathcal{E}_w\big\} \\
&\leq \epsilon + \sum_{w=2}^{2^{nR}} 2^{-n(I(X;Y)-3\epsilon)} \\
&= \epsilon + (2^{nR} - 1)2^{-n(I(X;Y)-3\epsilon)} \\
&\leq \epsilon + 2^{-n(I(X;Y)-R-3\epsilon)}.
\end{aligned} \tag{34}$$

The bound in (34) does not depend on the transmitted message index, hence it immediately follows that

$$\mathbb{P}\big\{\hat{W} \neq W\big\} \equiv \sum_{\boldsymbol{c}} \mathbb{P}\left\{\boldsymbol{C} = \boldsymbol{c}\right\} P_e^{(n)}(\boldsymbol{c}) \leq \epsilon + 2^{-n(I(X;Y)-R-3\epsilon)}. \tag{35}$$

Since $\mathbb{P}\big\{\hat{W} \neq W\big\}$ is the expectation of $P_e^{(n)}(\boldsymbol{C})$ with respect to the random codebook $\boldsymbol{C}$, there exists at least one realization $\boldsymbol{C} = \boldsymbol{c}^\star$, which is a deterministic codebook, with an error probability $P_e^{(n)}$ satisfying

$$P_e^{(n)} \leq \epsilon + 2^{-n(I(X;Y)-R-3\epsilon)}. \tag{36}$$

Given that $R < I(X;Y) - 3\epsilon$, the term $2^{-n(I(X;Y)-R-3\epsilon)}$ tends to zero as n grows large. Therefore, there exists a sufficiently large n_ϵ for which we have

$$R < I(X;Y) - 3\epsilon \text{ and } n \geq n_\epsilon \implies P_e^{(n)} \leq 2\epsilon. \tag{37}$$

Hence by selecting ϵ to be sufficiently small, we can make $P_e^{(n)}$ as small as desired for any $R < I(X;Y)$. The proof is almost complete, and we only need a couple of improvements.

- **Optimizing the input distribution:** The statement in (37) is strengthened by selecting the input distribution $p_X(x)$ in (30) to be a capacity achieving distribution, for which $I(X;Y) = C$. It follows that we can make $P_e^{(n)}$ as small as desired (hence approaching zero) for any $R < C$.

- **From average to maximal error probability:** Consider a codebook $\boldsymbol{c}^\star$ with rate R and for which the average error probability $P_e^{(n)}$ satisfies $P_e^{(n)} \leq 2\epsilon$. This exists by (37). We construct a new codebook $\tilde{\boldsymbol{c}}^\star$ with $\frac{2^{nR}}{2} = 2^{nR-1}$ codewords by throwing away the worst half of $\boldsymbol{c}^\star$, i.e. the 2^{nR-1} codewords with the highest error probabilities (assume that 2^{nR} is even). Next, we show that $\tilde{\boldsymbol{c}}^\star$ has a maximal error probability $\lambda^{(n)}$ which is no more than 4ϵ.

Codewords in the original codebook $\boldsymbol{c}^\star$ are indexed by $\mathcal{W} = \{1, 2, \ldots, 2^{nR}\}$. Partition $\mathcal{W}$ into equal-sized sets $\mathcal{W}_{\text{best}}$ and $\mathcal{W}_{\text{worst}}$, where $\mathcal{W}_{\text{best}}$ has the indices of the best half of codewords, while $\mathcal{W}_{\text{worst}}$ has the indices of the worst half. Let $\tilde{\lambda}$ be the error probability of the worst codeword in the better half. We have

$$\tilde{\lambda} \equiv \max_{w \in \mathcal{W}_{\text{best}}} \lambda_w^{(n)} \leq \min_{w \in \mathcal{W}_{\text{worst}}} \lambda_w^{(n)}. \tag{38}$$

We bound $\tilde{\lambda}$ as follows:

$$\begin{aligned} 2\epsilon &\geq P_e^{(n)} \\ &= \frac{1}{2^{nR}} \sum_{w \in \mathcal{W}} \lambda_w^{(n)} \\ &\geq \frac{1}{2^{nR}} \sum_{w \in \mathcal{W}_{\text{worst}}} \lambda_w^{(n)} \\ &\geq \frac{1}{2^{nR}} \cdot \frac{2^{nR}}{2} \min_{w \in \mathcal{W}_{\text{worst}}} \lambda_w^{(n)} \\ &\geq \frac{\tilde{\lambda}}{2}. \end{aligned}$$

From the above, we have $\lambda_w^{(n)} \leq \tilde{\lambda} \leq 4\epsilon$ for every $w \in \mathcal{W}_{\text{best}}$, which immediately implies that the new code has a maximal probability of error no more than 4ϵ. This new code has 2^{nR-1} codewords, and therefore its rate R' is given by $R' = R - \frac{1}{n}$. Note that having $R < C - 3\epsilon$, as in (37), is equivalent to having $R' < C - \frac{1}{n} - 3\epsilon$. For large n, R' and R becomes almost equal.

Combining the two points above, the statement in (37) becomes

$$R' < C - \frac{1}{n} - 3\epsilon \text{ and } n \geq n_\epsilon \implies \lambda^{(n)} \leq 4\epsilon.$$

Hence by selecting a small enough ϵ and a large enough n, we can make the maximal error probability $\lambda^{(n)}$ as small as desired for any $R' < C$. This completes the proof of achievability.

4 Converse to the Channel Coding Theorem

In this section we focus on proving the following statement in Theorem 2:

For any sequence of $(2^{nR}, n)$ codes with $\lambda^{(n)} \to 0$ as $n \to \infty$, we must have $R \leq C$.

The proof relies on the following key components.

- Having the maximal probability of error $\lambda^{(n)}$ go to zero as n grows large (reliability) implies that the average probability of error $P_e^{(n)}$ must also go to zero. Moreover, $P_e^{(n)} > 0$ immediately implies $\lambda^{(n)} > 0$ (unreliability). Therefore, we work with the average error $P_e^{(n)}$.

- A lower bound on the average probability of error $P_e^{(n)}$ is obtained using Fano's inequality:

$$H(W|\hat{W}) \leq 1 + P_e^{(n)} nR. \tag{39}$$

This is obtained from (4), given at the beginning of this chapter, while noting that $|\mathcal{W}| = 2^{nR}$.

- $W \to X^n \to Y^n \to \hat{W}$ form a Markov chain, since W is mapped (i.e. encoded) into X^n, which in turn is mapped (probabilistically) into Y^n, which is finally mapped (i.e. decoded) into $\hat{W}$. From the data processing inequality (see (3) at the beginning of this chapter), it follows that

$$I(W; \hat{W}) \leq I(X^n; Y^n). \tag{40}$$

Recall that W is uniform, and hence $H(W) = nR$. Combining this with the above points, we obtain

$$\begin{aligned} nR &= H(W) \\ &= I(W;\hat{W}) + H(W|\hat{W}) \\ \text{(Fano)} \quad &\leq I(W;\hat{W}) + 1 + P_e^{(n)} nR \\ \text{(Data Process.)} \quad &\leq I(X^n;Y^n) + 1 + P_e^{(n)} nR \\ &\leq nC + 1 + P_e^{(n)} nR. \end{aligned}$$

In the above, the last inequality follows from Lemma 1. The above inequality is rearranged as follows

$$R(1 - P_e^{(n)}) \leq C + \frac{1}{n}. \tag{41}$$

Recall that as $n \to \infty$, $\lambda^{(n)}$ and $P_e^{(n)}$ both to zero. Therefore, taking the limit $n \to \infty$ in (41) implies that $R \leq C$. Note also that the inequality in (41) may also be written as

$$P_e^{(n)} \geq 1 - \frac{C}{R} - \frac{1}{nR}. \tag{42}$$

This shows that if $R > C$, then $\lim_{n\to\infty} P_e^{(n)} > 0$ and the error remains bounded above zero. This implies that $P_e^{(n)} > 0$ for all n: if $P_e^{(n)} = 0$ for some small n, then we can construct a code with $P_e^{(n)} = 0$ for large n by concatenating smaller codes, hence contradicting $\lim_{n\to\infty} P_e^{(n)} > 0$. Finally, it immediately follows that whenever $R > C$, we will have $\lambda^{(n)} \geq P_e^{(n)} > 0$, and hence reliable communication is not possible.