CT-111: Introduction to Communication Systems

Notes to Supplement Several Recent Lectures

Yash Vasavada

Revision 2, 31st March 2018

Preface

My dear student in the class of CT111: at the beginning of the semester, I had mentioned to you that my goal is to turn the typical bell-shaped curve[1] of the exam scores into a shape that is highly concentrated[2] on the right side of X-axis.

I, in fact, was not having the grades in my mind as much when I had mentioned this. I instead wanted all of you, without exception, to understand the subject of CT111 well. With the understanding comes clarity, clarity leads to further interest and curiosity and eventually passion about the subject, and that in turn speeds up even more detailed understanding of the subject, thereby completing a positive cycle. Good grades fall out simply as the side-product for those who enter this cycle. My hope was that all of you enter this cycle by the end of the semester.

Now that the semester is drawing to a close, I realize that some students are doing reasonably well in the class, but many are not. After Mid-Term I, we saw that the distribution of grades was indeed bell-shaped. The probability that the probability distribution of Mid-Term II grades will be concentrated on the right side is very small.

I am facing a problem of how to try and help those of you interested in entering the above positive cycle. I realize that many of you are regular in the class, you are making an effort to understand the subject material, but you are not still quite getting it yet.

I have written this note as an attempt to help you in your effort to understand several recent topics that we have been studying in the class. This document is not to be read in a standalone manner. Instead take this document as the companion to the lecture slides. The details that are in the lecture slides are not here, and possibly vice versa. Try to answer the questions at the end of each chapter. We will hold a few seminars to go through this material where you would ask the questions and doubts that you maybe facing.

[1] We have now understood this enough and can replace the informal daily-language phrase by the mathematical terminology; i.e., that this is a Gaussian Distributed Probability Mass Function (PMF) of exam scores.

[2] Possibly even a Dirac Delta function located at "AA"? Alas, this is too much to ask.

Chapter 1

What is Information?

The main business, the "why", of an engineering communication system is to transfer the information. Just as a trucking, shipping or any freight system transfers the goods, the communication system transfers information.

This information transfer can occur from one place to another at the same time, or it can occur from one time to another at the same place.

The first topic to understand well in engineering design of a communication system is what is information.

The word information carries multiple meanings and connotations in the daily life, and a single clear and crisp definition is hard to arrive at. However, from an engineering and mathematical standpoint, a crisp definition is, surprisingly, possible to formulate, and it is as follows:

$$I(x_\ell) = -\log_2(p_\ell) \quad \text{bits} \tag{1.1}$$

The above equation says that information $I(x_\ell)$ carried by some message x_ℓ, which occurs with probability of p_ℓ, is given as $-\log_2(p_\ell)$, and its units are bits. The information is conveyed because the person receiving the information (we will call him or her or it as the receiver) is *uncertain* about the message m_i. This uncertainty at the receiver arises because

1. x_ℓ is one message out of a set of messages:

$$\mathcal{X} = \{x_1, x_2, \ldots, x_\ell, \ldots, x_L\} \tag{1.2}$$

2. the sender or the transmitter can pick any one message out of the above set. The receiver does not know what message the sender is going to send.

Thus, we can say that the transmitter generates the information by his/hers/its act of *selecting* one message out of the set. As seen by the receiver, this selection is probabilistic. The receiver does not know ahead of time that the transmitter is going to pick a particular message and send that. If it knew that with certainty, p_{x_ℓ} would be 1, and the information $I(x_\ell) = -\log_2(p_\ell)$ would be 0 bit. In this case, the transmitter might as well save itself the trouble and not send the message at all.

Following are some points to consider:

1. While the information generated when a particular message x_ℓ occurs is useful to know (its value is given by Eq. 1.1), it is equally useful to know how much information is contained in the entire source itself. Let us say that a particular symbol x_ℓ has a very small probability p_ℓ, and therefore, according to Eq. 1.1, it generates a large amount of information. However, the overall information of the source may still be small if the rest of the symbols occur more frequently. Since they are occurring more frequently, their probabilities are larger and therefore they contain small information. By this logic, we can see that it is not possible to define the overall informaton of the source by Eq. 1.1. One way to do so is to take an average of $I(x_\ell)$ over all values of ℓ. That gives us an overall idea about (or the average value of) the information generated, or contained, in a Discrete Memoryless Source or DMS.

   Recall the Expectation operation that we have studied in Lecture 10. Using that, we write the average information as the entropy $H(X)$ of source X:

$$H(X) = E\left[I(x_\ell)\right] = -\sum_{\ell=1}^{L} p_\ell \log_2(p_\ell) \quad \text{bits} \tag{1.3}$$

   When the DMS is binary, the above is written in a special manner as follows:

$$H(X) = H_b(p) = -p\log_2 p - (1-p)\log_2(1-p) \quad \text{bits} \tag{1.4}$$

2. As mentioned earlier, the information is generated by the act[1] of selecting one message out of a set of L messages. Accordingly, the greater the power that the generator of the information has in making this selection, the greater is its ability to generate more information.

   ▷ It is indeed the case that as L, the number of messages in the set shown in Eq. 1.2, increases, the bigger are the possible values that that $H(X)$ can take.

[1] Read the second paragraph on the first page of Shannon's 1948 paper, which begins with "The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point." Notice the word "selected" in this.

- ▷ In channel coding context, we have seen in the class that the encoder purposefully does not exercise its full selection power. Using N bits, the encoder can actually select one of 2^N messages. However, when the encoder knows that the message, or the channel codeword, is going to be sent over a Binary Symmetric Channel (BSC) with crossover probability of p, it selects one of only $2^{(N\times(1-H_b(p)))}$ predefined codewords. Although the selected codeword has N bits, the selection power exercised is worth $K = N \times (1 - H_b(p))$ bits only. One way to think about this is as follows: for every bit transmitted in this codeword, only a fraction $r = \frac{K}{N}$ carries the information; the remaining fraction $1 - r$ is needed to ensure recovery from the BSC induced errors at the receiver. Channel coding theorem answers the question: what is the largest possible value of r for a given BSC channel with crossover probability p? For BSC(p), this largest value of r is given as $1 - H_b(p)$. This is the theoretical upper limit. Practical channel coding techniques can hope to come close to it, but never exceed it.

3. The information is generated by the process of selection. It is the selections that matter. The exact shape or form of the symbols in the set $\mathcal{X}$ in Eq. 1.2 does not matter, *provided* the sender and the receiver both are in agreement about which particular symbol x_ℓ each selection is going to translate into. This agreement is called a *dictionary*.

   - ▷ When we learn a language, we have to first learn its dictionary. Afterward, we can *decode* the *cognitive* selections that the person that we are talking to (let us say, our friend) is making by listening to him/her, i.e., by *receiving* the "codewords" transmitted by this friend. These codewords are the sentences of the language that we have learnt. These codewords can change, e.g., when our friend switches from speaking in Hindi to speaking in English, but because our mind has stored the dictionaries of both the languages, we can still observe the same cognitive selections. The words and the sentences fade away from our minds once they convey the meaning. We can understand the meaning only when we can observe the cognitive selections being made by our friend. Alternatively, when we tell him/her that "I don't understand what you mean.", it is not that we have not understood the individual words and sentences, but that we have failed to decode the transmitted codeword, i.e., we have failed to observe the mental selection that he/she was making.
   - ▷ This point is going to help us better understand the codebooks required for channel encoding. These codewords in a codebook oftentimes are totally different[2]

[2]However, in many, especially simple type of, channel coding techniques, the message bits of length K

from the message bits. The message (or the information) bits are used to just select one codeword out of this codebook. This mapping of the message bits onto the codewords is unambiguously agreed upon at both the channel encoder and the channel decoder. Similarly, the source encoding and source decoding also rely upon such dictionary-based codebook agreements.

Practice Problems

1. Give an example of a case when the information transfer occurs from one place to another at the same time, and example of a case when information transfer occurs from one time to another at the same location.
2. Why logarithmic function is used in Eq. 1.1?
3. Why is the logarithmic function output is multiplied with -1 in Eq. 1.1?
4. Units mentioned in Eq. 1.1 are bits. What change would be required in this equation to make the units as digits (0 to 9) instead of bits.

block do make up a part of the codeword of length N bits, where $N > K$. However, this is only to simplify the implementation aspects, and this simplification sometimes is obtained at the expense of *weakening* the channel code, i.e., reducing its strength in providing the protection against the noise.

Chapter 2

What is Probability?

2.1 Probability and Information

Information transfer becomes possible only in probabilistic situations, not in deterministic cases. When everything is determined, there is nothing to say/communicate. In the daily life, if someone says some thing already known (we all have friends or relatives that have the habit to repeat the same thing over and over again), the response from the "receiver" is typically "oh, please spare me!"'.

Therefore, to understand the engineering design of a communication system that transfers the information, we need to understand the concept of probability.

This concept was tested in Quiz 3 by the following question: suppose an experiment, or a discrete-valued source of information, has a total of L different outcomes (or symbols or messages). Write the expression for the probability p_ℓ of its ℓ^{th} outcome ($\ell = 1, \ldots, L$).

The answer to this question is given in Lecture 10 material, and it is repeated here. The probability p_ℓ is the limit value of the ratio of the number of occurrences m_ℓ of ℓ^{th} outcome/symbol/message out of a total of $M = \sum_{\ell=1}^{L} m_\ell$ different experimental trials (or a string of M symbols) as $M \to \infty$.

$$p_\ell = \lim_{M \to \infty} \frac{m_\ell}{M} \tag{2.1}$$

The main idea[1] here is that the probability is a limit, and that this limit *exists* and that it is not changing. The DMS is non-deterministic, but its asymptotic behavior (as $N \to \infty$) is *indeed* deterministic provided we know p.

[1]This is not the only way to define the probability. End of the chapter problems describe two other ways.

We may not know an individual outcome of this random information source ahead of time. That is actually good, since this randomness, as mentioned earlier, allows it to be an information generating source. However, there is something that we can say with *certainty*, which is as follows (notice that this is just another way of stating Eq. 2.1). If we observe the outcome of this information source sufficiently long (for large enough M), we are likely to see a total of $\approx p_\ell \times M$ occurrences of symbol m.

Suppose we are given a value of $p_\ell = 0.24$. Can we observe a sequence of length $M = 10000$ symbols in which symbol x_ℓ occurs $m_\ell = 79$ times?

- ▷ If we did observe such a sequence, from Eq. 2.1, $p_\ell \approx 79/10000 = 0.0079$. This, however, contradicts the value of $p_\ell = 0.24$ as the probability of symbol m.
- ▷ Therefore, we can say that we should never (technically, with a probability approaching zero) observe a sequence where symbol m occurs $m_\ell = 79$ times in a string of $M = 10000$ symbols when $p_\ell = 0.24$.

This is the underlying idea behind the concept of the typical (or surviving) set, which we will elaborate upon in the next chapter.

Joint, Conditional and Total Probabilities; Bayes Theorem

The following paragraph and example are taken from the book by Henry Stark and John Woods.

Joint Probability

Suppose we are in a certain city and are interested in three different events:

A is the event that on a particular day, the temperature is greater than 10^o C. We also consider the complementary event of event A, i.e., that the temperature is equal to or below 10^o on a given day, and denote it as $\bar{A}$.

B is the event that on a given day, the rainfall amount exceeds 5 mm. Here, $\bar{B}$ denotes the event that the rainfall is equal to or less than 5 mm.

C is the event that on a given day, both A and B occur, i.e., the temperature exceeds 10^o C and the rain is more than 5 mm. We think of event C as logical-AND of events A and B, i.e., $C \stackrel{\text{def}}{=} A \textit{ and } B$, or for simplicity, $C = AB$.

Since $C = AB$ is an event, we call $p(C) = p(AB)$ as the joint probability of events A and B (this notion can obviously be extended to more than two events). Let m_i denote the number of days that an event i occurs, and we make observations over a total of $M = 1000$ days, and observe that $m_A = 811$, $m_B = 306$ and $m_{AB} = 283$, and $m_{A\bar{B}} = 528$. According to Eq. 2.1 (assuming, for the sake of simplicity, that these values calculated using $M = 1000$ are the same as the asymptotic values when $M \to \infty$),

$$\begin{aligned} p(A) &= \frac{m_A}{M} = 0.811 \\ p(B) &= \frac{m_B}{M} = 0.306 \\ p(AB) &= \frac{m_{AB}}{M} = 0.283 \\ p(A\bar{B}) &= \frac{m_{A\bar{B}}}{M} = 0.528 \end{aligned}$$

Here, $p(AB)$ is called the joint probability of events A and B, i.e., it is the probability that events A and B both happen together.

Conditional Probability

Now consider the ratio m_{AB}/m_A. This ratio (when $M \to \infty$) gives the probability with which event AB occurs when event A has already occurred. Essentially, this ratio considers a hypothetical "world" in which event A always happens (the rainfall is always more than 5 mm no matter which day is picked). In this "world", what is the probability that event B occurs (i.e., temperature exceeds 10^o C)? A moment's reflection should convince you that the answer to this question is given by the ratio m_{AB}/m_A, and it is called the *conditional probability* of event B given event A has occurred.

$$p(B|A) = \frac{m_{AB}}{m_A} = \frac{m_{AB}}{M} \times \frac{M}{m_A} = \frac{p(AB)}{p(A)} \tag{2.2}$$

We can similarly write, the conditional probability of event A *given* event B as follows:

$$p(A|B) = \frac{m_{AB}}{m_B} = \frac{m_{AB}}{M} \times \frac{M}{m_B} = \frac{p(AB)}{p(B)} \tag{2.3}$$

Events and Outcomes

We have considered two different concepts. First one is that of an information source that we call the DMS. The other concept that we have thought about above is that of an event.

However, there is a unity between these two apparently different concepts. Each event of our daily lives can be thought of as providing some information. It may seem a little strange to think of the events as being information sources and vice versa, but if you think about it, you will come to a realization that the events and information sources are analogous.

Total Probability

Is it possible to calculate the total probability of event A when we are given the joint probability $p(AB)$?

The answer to this question is "yes". Suppose the only two observations we have made are those of m_{AB} and $m_{A\bar{B}}$. If we sum these two numbers, we obtain m_A. In the example above, $m_{AB} = 283$ and $m_{A\bar{B}} = 528$, and on summing these two numbers, we obtain 811 which equals m_A.

Reason for this is that event AB corresponds to a hypothetical world in which both A and B occur. Similarly, event $A\bar{B}$ occurs in a world where A occurs and B does not. If we combine these two worlds, we obviously obtain a combined world in which A occurs.

We summarize this as follows:

$$\begin{aligned} p(A) &= p(AB) + p(A\bar{B}) \\ p(\bar{A}) &= p(\bar{A}B) + p(\bar{A}\bar{B}) \\ p(B) &= p(AB) + p(\bar{A}B) \\ p(\bar{B}) &= p(A\bar{B}) + p(\bar{A}B) \end{aligned}$$

If we compute the (total) probability of an event from the joint probabilities in the above manner, it is called the process of marginalization, and the obtained probability is sometimes called the marginal probability.

Bayes Theorem

Bayes Theorem rewrites Eq. 2.3 in terms of Eq. 2.2.

$$p(A|B) = \frac{p(AB)}{p(B)} = \frac{p(B|A)P(A)}{p(B)} \tag{2.4}$$

Refer to Lecture 10 for further material on joint and conditional probabilities and Bayes' Theorem.

Practice Problems

1. There is a concept of probability as intuition. This concept is different from the concept underlying Eq. 2.1. Consider the meaning of the word "probably" or "probability" in the following sentences:

   ▷ What is the probability that there is life on other planets?

   ▷ What is the probability that Sejal will marry Joseph?

   ▷ Saurin probably was driving too fast when he met with an accident.

   In what sense, the probability concept underlying the above sentences is different from the one outlined in Eq. 2.1? What is the main limitation of this way of defining probabilities?

2. There is another way to measure the probability, which is by taking the ratio of the number of ways a specified event can occur in an experiment to the total number of outcomes of this experiment. This is explained by means of an example below:

   - Suppose an ideal fair coin (i.e., the coin for which heads and tails occur with equal chances) is tossed two times. What is the probability of getting at least one tail?

     → Calculation of probability in answering this question can be performed as follows.

       ▷ There are four outcomes of this experiment: HH, HT, TH and TT.

       ▷ The specified event, at least one tail, can occur three ways: HT, TH or TT.

     → Thus, the probability that at least one tail occurs is 3/4.

   - How is this definition of probability different from the definition in Eq. 2.1?
   - What is the main limitation of this definition of probability?

3. Definition of probability in Eq. 2.1 does not suffer from the problems with these alternative definitions. However, there is still one significant limitation of defining probabilities using Eq. 2.1. What is this limitation?

4. We have seen in the above that there are (at least) three different ways of defining the probability. Next, consider the following problem. In a spinning-wheel game, the spinning wheel contains the numbers 1 to 9 and the contestant wins if an *even* number shows. What is the probability of a win? By which of the three methods did you calculate this probability? What is the underlying assumption?

5. Consider events A, B and C defined in this chapter (on page 8) as being information sources. What are the value of L for these three sources (refer to Eq. 1.2 for the definition of L.)

6. A fair die is tossed twice (a die is fair if all outcomes $1, 2, \ldots, 6$ are equally likely). Given that 3 appears on the first toss, what is the probability that the total of the two numbers obtained on the two tosses is 7?

7. There are three machines A, B and C at a semiconductor manufacturing facility that makes integrated circuits (ICs). They manufacture 25, 35 and 40 percent of the total ICs manufactured. For machine A, 5% of the manufactured ICs are defective, for machine B, 4% of chips are defective and 2% of chips made by machine C are defective. A chip is drawn randomly from the combined output of these three machines and it is found to be defective. What is the probability that it was manufactured by machine A? by machine B? by machine C?

8. BTech 2017 class is taking a quiz in which there is a multiple-choice question with m alternatives. The fraction of the students who know the answer to this question is p, and remaining $1 - p$ fraction of the total students will guess. The probability that the student who knows the answer gives the correct answer is 1. The probability that the student who guesses will give the correct answer is $1/m$. Compute the following:

   (a) the (conditional) probability that a student knew the answer to the question *given* that he/she answered it correctly.

   (b) the (total) probability that this question is answered correctly.

Chapter 3

What is the Typical Set?

We concluded Section 2.1 of Chapter 2 by briefly alluding to the concept of typical set. Let us think further about this topic.

Suppose we consider a binary Discrete Memoryless Source (DMS) for which the probability of 0 is 0.99, and probability of occurrence of 1 is $p = 0.01$. Suppose we consider a length $M = 100$ long binary string at the output of this DMS. Since probability of 1 is 0.01, we expect roughly 1% of $M = 100$ bits to be 1, i.e., we expect 1 bit out of 100 bits to be 1 and the rest 99 bits to be zero. Recall this is the definition of probability in Eq. 2.1. However, even without explicitly checking with Eq. 2.1, we intuitively feel this to be correct.

Now suppose the probability p increases to become 0.1. In this case, when we look at $M = 100$ bit sequence out of this binary DMS, we expect to see mostly those binary sequences for which the number of ones is around 10.

Slide 31 of Lecture 12 shows the probabilities of observing $M = 100$ bit long binary sequences that have 0 ones, 1 one, 2 ones, all the way to sequences that have $M = 100$ ones.

This slide 31 is just one element of a *story* that begins on slide 24 and continues upto slide 44. Read these slides and ensure that you understand the main concept that is being conveyed.

Generalizing this, we can say that for arbitrary probability p, when we look at an arbitrary M number of bits, we expect to see *only* those binary sequences for which the number of ones is roughly $M \times p$ out of M. As $M \to \infty$, we can be increasingly certain that the observed binary sequences will be only one of those sequences that have $M \times p$ bits as ones out of a total of M bits.

We will now arrive at the concept of the typical set and the derivation of the entropy formula:

1. Since the source is DMS, we can calculate the probability p_i of observing an M bit

sequence (let us call it i^{th} sequence) with $M \times p$ ones and $M - M \times p$ zeros:

$$p_i = p^{M \times p} \times (1-p)^{M \times (1-p)} \tag{3.1}$$

(a) Note that all sequences with $M \times p$ ones and $M - M \times p$ zeros have the same probability p_i given in Eq. 3.1.

(b) Suppose the total number of such sequences is K. The set of these K sequences is called the typical (or the surviving) (sub)set, since, as mentioned repeatedly earlier, when $M \to \infty$, it is nearly certain that we will observe a binary sequence that is from this set. Alternatively, probability of the typical set tends to be 1 as $M \to \infty$.

(c) Probability of the typical set is simply the sum of the probabilities of its member sequences. From point (a) above, this probability is p_i. Therefore, the probability of the typical set is $K \times p_i$. From point (b) above, $K \times p_i = 1$. Therefore, $K = 1/p_i$. Using simple manipulations, you should be able to prove that $K = 1/p_i$ can also be written as $K = 2^{M \times H(X)}$ (see also slides 43 and 44 of Lecture 12).

2. How many M bit strings there are in which the number of ones is $M \times p$? The answer to this is $\binom{M}{M \times p}$. Therefore, we conclude that $\binom{M}{M \times p} \approx 2^{M \times H(X)}$. Alternatively, if we take $p = k/M$, we can write that $\binom{M}{k} \approx 2^{M \times H_b(p=k/M)}$ ($H_b(p)$ is defined in Eq. 1.4). This was the approximation that you were asked to show in a problem in Part II of Midterm II.

Practice Problems

1. How many $M = 4$ bit strings there are in which the number of ones is $k = 2$? First arrive at the answer by counting. Next, verify your answer and ensure that this is $\binom{N=4}{k=2}$.

2. How many binary bit patterns of length $N = 100$ are there with one 1 and 99 zeros?

3. For a binary DMS with probability p of 1 equal to 0.1, why do we expect to see only those $N = 100$ bit long sequences for which the number of ones is 10?

4. Given $M = 100$ as the length of the binary sequence and $p = 0.1$ as the probability that the DMS generates 1,

   ▷ what is the number of possible different sequences with 30 ones?

- ▷ What is the probability of observing a $M = 100$ long binary sequence that has 30 ones.

5. At a bicycle company, 70% of the newly made cycles pass the quality inspection and 30% fail. What are the chances that 3 of 10 cycles that you're inspecting will fail?

   - ▷ How is this problem similar to the previous one?
   - ▷ Answer to this problem is given on slide 41 of Lecture 10. Understand slides 41 to 46 of Lecture 10.

6. A smuggler, trying to pass himself as a glass-bead importer, attempts to smuggle diamonds by mixing diamond beads among glass beads in the proportion of one diamond beam per 1000 beads. A customs inspector examines a sample of 100 beads. What is the probability that the smuggler will be caught?

   - ▷ Although this question again looks very different, it is also similar to the previous two questions.

7. Using Matlab, generate the plots given on slides 24 to 28 of Lecture 12.

8. In the middle of page 13 (Section 7) of Shannon's 1948 paper, there is a sentence "Suppose in this case we consider a long message of N symbols." (Shannon's notation is that the DMS has a total of n symbols (i.e., it is not necessarily binary since n can be greater than 2), with probabilities $p_1, p_2, \ldots, p_n$). Read the paragraph beginning from this sentence and the mathematical derivation of entropy in the next four equations. Have you understood this derivation?

Chapter 4

Channel Models

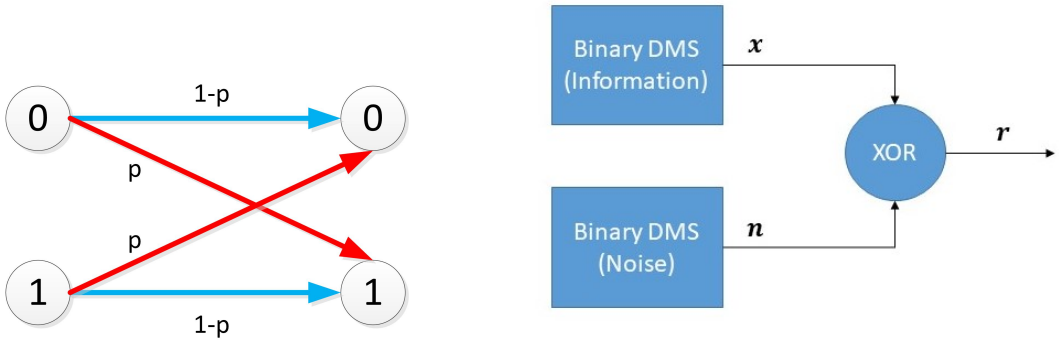
This chapter mentions three different types of channel models: the BSC(p) that we have studied in the class, and two other types of channels.

4.1 BSC(p)

In Lecture 13, we have learnt about a special type of communication channel, called Binary Symmetric Channel or BSC, whose input and output are both bits.

This BSC corrupts, or flips, the bits sent by the transmitter with a probability of p. This is called the cross-over probability of BSC, and the BSC name is often associated with p, i.e., as BSC(p). Given a value of p, there is only one type of BSC. However, since p can take infinite number of values from 0 to 1, the BSC is actually a family of channels, and BSC(p) is a specific member of this family.

Using the exclusive-OR or modulo-2 notation shown on slide 16 of Lecture 13, we can think of BSC(p) as being a binary DMS, not too different from the binary DMS information source that provides the input to this BSC. In fact, the situation is shown in Fig. 4.1b, where both the information source and the BSC are shown as being the random binary number generators. The difference between the two is that the uncertainty inherent in the former is a *good* type of uncertainty. It allows the information generation and transfer. In contrast, the uncertainty or the randomness of the latter is of a *bad* type; it hampers the process of information transfer.



(a) State transition model (b) An alternate model.

Figure 4.1: Model of BSC(p).

4.2 Binary Erasure Channel

The BSC(p) channel flips one bit out of $1/p$ bits in random. The receiver gets bits, but it cannot fully trust what it sees. When it receives 0, the transmitter could have sent 0 (probability of this is $1-p$) or it could also have sent 1 (probability of this is p). When p is small (e.g., $p = 0.01$), and suppose the receiver knows this value of p. In this case, the receiver has a good confidence in the bits it is seeing. The worst-case is when $p = 0.5$. In this case, the receiver cannot trust at all the bits it is getting at the output of the BSC. The bits at the output of the BSC stop conveying any information to the receiver.

The Binary Erasure Channel or BEC is a different type of channel that does not flip the bits, but instead it erases the bits. Those bits that are not erased are known at the receiver with full certainty. However, if a bit is erased, the receiver does not get any information at all. The output of the BEC is not binary, but instead it is ternary. This is shown in Fig. 4.2a. For BEC(e), the bits get erased with an erasure probability of e. However, with a probability of $1-e$ the bit is not erased and it is received without any error.

4.3 Non-Binary Gaussian Noise Channel

Real-world channels are not binary, or ternary, etc. The noise is analog-valued and it typically gets added to the transmitted signal. The block diagram of the resultant channel is shown in Fig. 4.2b.

(a) Binary Erasure Channel with erasure probability of e, i.e., BEC(e).

(b) Gaussian Distributed Noise Channel.

Figure 4.2: Two channel models.

Practice Problems

1. Let X denote the random variable at the input to the channel, and Y denote the RV at the channel output. Let q_0 and q_1 denote the probabilities that X is 0 and 1, respectively. For the BSC(p) and BEC(e), calculate the conditional probabilities $p(Y|X)$, and $p(X|Y)$.

2. Using your formulation of $P(X|Y)$, state the rule for decision at the receiver regarding whether the transmitted bit X is 0 or 1 given observed value Y at the channel output when (i) $q_0 = q_1 = 0.5$, and (ii) $q_0 = 0.9$, and $q_1 = 0.1$.

   ▷ In Lecture 10, we have solved this problem for non-binary Gaussian noise channel (in Section 4.3) using the Bayes' Theorem for a case when the input X takes two values $+A$ and $-A$. Apply the same methodology.

3. Let us say that a rate 1/3 repetition code is used at the input to the BSC(p). Thus, each bit is repeated three times, and the transmitter sends two codewords X, either 000 or 111. Due to the effect of BSC(p), the receiver can see as Y any one of eight 3 bit patterns. Use Bayes' Theorem to determine $P(X|Y)$ for both values $X = 0$ and $X = 1$ of the transmitted bit (for the case when $q_0 = q_1 = 0.5$) and all eight values of Y. Using this calculation of conditional probability, state the decision rule for this rate 1/3 repetition code.

Chapter 5

Source and Channel Coding

In Chapter 1, we considered that the information is generated due to the selection that the transmitter makes. The greater the *freedom* the transmitter has in making the selections, the greater is the information generation capability of the transmitter.

Let us think about this further. Suppose that in some hypothetical communication setup, the transmitter and the receiver have a-priori (i.e., ahead of time) reached the following agreement: any time the sender makes use of the communication link, it will be sending to the receiver one of only $L = 4$ messages.

In this case, the actual messages can be any. For example, these four messages can be four movies, or four photos, or two photos and two Harry Potter novels, etc. The source coding scheme says that only two bits are needed in this case because the selection is limited to be from a set whose size is $L = 2$.

One might ask how would it be possible to communicate a movie, which occupies billions of bits if we download it on our laptop, using only two bits. The answer is that, because of the a-priori agreement, the sender is not sending the movie over the communication link. It is sending only its *selection* about which movie, one out of four, it has selected. The actual, gigabyte-sized, four movies are stored at both the transmitter side and the receiver side, in their *codebook dictionaries*, ahead of time. The dictionary, i.e., the movies in this scenario, are not communicated; instead the communication problem is restricted (because of the a-prior agreement) to conveying only which of the four movies from the codebook the transmitter has selected.

Above example illustrates the central aspect of the engineering problem of information transfer, i.e., the engineering problem of communication. This example tells us that the bits are needed to separate one selection from another; bit sequence $\{0, 0\}$ in this case is distinct from the other three $\{0, 1\}$, $\{1, 0\}$, and $\{1, 1\}$, and the receiver is communicated exactly

which of the four messages was selected because of this distinct-ness.

It from Bit

Actual movies that one can download today as a digital file are gigabit-sized because the selections required for making a movie are far more many than just $L = 4$. If a movie (or a novel, an image, etc.) is 10^9 bit long, it is one message from a set whose size L is 2^{10^9}. This is truly an astronomical number. We saw in Lecture 12 that 2^{10^2} is a gigantic number (much greater than the age of the universe in seconds), and 2^{10^3} is even more fantastically big number (much greater than the number of electrons in the entire universe). Therefore, we cannot comprehend how big a number 2^{10^9} is. The question is: are there really those many (2^{10^9}) ways of making a movie?. If the answer is "yes", the data compression is not possible. All 10^9 bits are needed to uniquely distinguish one movie from the rest of 10^9 *minus* 1 possible other movies. If the answer is "no", the data compression is possible since some of 10^9 bits are not required to do the job for which they are recruited, i.e., to separate one movie from the rest.

An alternate way to think about this is as follows. Suppose the world, as our eyes see it, was not rich in the color; instead it was only black and white (e.g., either because our eyes did not have capability to distinguish colors, or because there were actually no colors in the world). Now when the world (the transmitter) sends us (the receiver, i.e., our eyes and our cognitive visual process) the movie (the daily scenes that we see), the transmitter will have much reduced power in making the selections (since no selections are required to choose one color from many others), and accordingly, its information generation capability will be reduced. Movies of this hypothetical world might require only 10^6 bits instead of 10^9 bits. This loss of information generation capability will be perceived at the receiver as the world being less interesting and less rich (in the color).

Taking this argument, some scientists have said that the actual world that we perceive is made up of not things, but instead it is made of information. The object that our brain perceives, the table or the rock we feel when our hand touches this object, is an "it" that is generated inside our brain because of the "bit" that the brain receives. The motto is "It" from "Bit"!

Minimum Hamming Distance

Let us return to the example of transferring one message or symbol out of a set of size $L = 4$.

This requires a two-bit message. However, the minimum Hamming distance d_{min} between this set of four two-bit patterns is 1. This implies that a transmitted message can *turn into* another message when one bit is received in error. The transmitter intended to indicate to the receiver its selection of a movie A by sending a bit pattern $\{0, 0\}$. However, due to one bit of error, let us say in the second bit, the receiver received a bit pattern $\{0, 1\}$. Therefore, the receiver wrongly believes that the transmitter has selected movie B.

The Single Parity Check (SPC) code remedies this situation by adding a parity bit. By adding the parity bit, the transmitted bit pattern is now three-bit long and d_{min} has increased to become 2 bits. When $d_{min} \geq 2$, there is a cushion that protects the transmitted codeword. This cushion is to be imagined as a *sphere* (or an oval) surrounding the transmitted codeword that is sitting at the center of the sphere. When all the possible transmitted codewords have these protecting spheres (or ovals) surrounding them, and when all these spheres do not intersect or touch each other, the error correction and detection both become possible.

Sphere Packing

To create non-overlapping Hamming "spheres" (or "ovals") of radius one bit around all of N-bit transmitted codewords, the channel coding scheme needs to *ensure* that all of a total of $\binom{N}{1} = N$ different binary strings of length N which differ from a chosen codeword in one bit do not correspond to a valid codeword as defined in the codebook dictionary.

Extending the above logic, we can say that to create non-overlapping Hamming spheres of radius t_c bits around all of N-bit transmitted codewords, the channel coding scheme needs to *ensure* that all of a total of $V_{N,t_c} = \sum_{k=1}^{t_c} \binom{N}{k}$ different binary strings of length N which differ from a chosen codeword in one bit, two bits, etc. up to k bits do not correspond to a valid codeword as defined in the codebook dictionary.

The actual sphere available to us has a radius of N bits and it has 2^N different points inside it. The goal of channel coding scheme is to *pack* as many smaller spheres as possible of radius t_c bits that are non-overlapping where there are V_{N,t_c} points in each smaller sphere, in addition to a distinct valid codeword that is at the center. As increasing number of spheres are packed within the outer sphere of size 2^N, the Hamming distance between the distinct valid codeword reduces (because the outer sphere becomes crowded). If we want an error *correction* capability of t_c bits, we have to stop packing more spheres when the minimum value d_{min} of the Hamming distance (i.e., the Hamming distance d between the immediate neighbors) becomes equal to $2\,t_c+1$. If we still try to squeeze in more number of spheres, d_{min}

will reduce below $2\,t_c+1$ bits, and the error correction capability will not be $t_c = \left\lfloor \frac{d_{min}-1}{2} \right\rfloor$ bits anymore.

Channel Coding Theorem

Channel Coding Theorem considers the asymptotic case of the above sphere packing. Following is a step-wise logic leading to the statement of the theorem:

1. For a BSC, as $N \to \infty$, the required size of the spheres becomes $2^{N \times H_b(p)}$.

   → When the size of the individual smaller spheres (which are packed within the larger sphere of size 2^N) equals $2^{N \times H_b(p)}$, the BSC induced error pattern $\mathbf{n}$ is ensured to be *trapped* within this smaller sphere. This is because $\mathbf{n}$ is, with probability nearing 1, one pattern from the typical set of this size.

2. This allows packing a total of $\frac{2^N}{2^{N \times H_b(p)}} = 2^{N \times (1-H_b(p))}$ non-overlapping smaller spheres in the larger sphere of size is 2^N.

3. Number of spheres that are packed determines the size of the codebook dictionary. Center of each sphere is a valid codeword, and it is included in the dictionary.

   → Actual (practical) channel encoding scheme cannot afford to store the valid codewords in a dictionary. This is because for practical values of N and $H_b(p)$, the number $2^{N \times (1-H_b(p))}$ can become astronomical.

4. Number of bits required to *select* one item (or message, symbol, codeword, etc.) from this dictionary is the number of information bits K that can be reliably transferred over the BSC (with a guarantee of recovery from BSC induced errors).

5. This number K, which equals the logarithm base two of the number of packed spheres, is, therefore, given as $N \times (1 - H_b(p))$.

6. *Channel Coding Theorem (for BSC).* Above shows that the ratio $r = \frac{K}{N}$, i.e., the channel code rate, can at the most equal $1 - H_b(p)$ but it cannot exceed it.

Practice Problems

1. Calculate all possible Hamming distances between four un-encoded messages of length two bits, and show that the minimum Hamming distance d_{min} between this set of four two-bit patterns is 1.

2. Calculate all possible SPC codewords obtained by encoding each two-bit message using a single parity check. Repeat the above exercise and show that the minimum Hamming distance d_{min} for this encoded SPC messages is 2 bits.

3. You have been given 16 codewords of $\{7, 4\}$ Hamming Code in Matlab in Lab 8. Calculate all possible Hamming distances between these codewords and show that d_{min} is 3 bits.

4. Prove the following statement: if a channel coding scheme aims to create non-overlapping Hamming spheres of radius t_c bits around an N-bit transmitted codewords, this scheme needs to *ensure* that total of $V_{N,t_c} = \sum_{k=1}^{t_c} \binom{N}{k}$ different binary strings do not correspond to a valid codeword in the codebook dictionary.

5. What are the error correction capability and the error detection capability (specify both as number of bits) of the coding scheme described above?

6. What are the error correction capability and the error detection capability of the coding scheme described above?

7. State the mathematical relationship between code-rate r of the above coding scheme and V_{N,t_c} (recall that $r = k/n$, where n is the length of the codewords in bits, k is the number of information bits that are encoded).