Learner's Space 2025
Information Theory and Coding
Week 2, Handout 2

Suketu Patni

June 20, 2025

Abstract

We have so far mentioned only of Claude E. Shannon that he is the progenitor of information theory and seen a measure of information named after him. However, we haven't yet seen two of his most famous theorems, the source coding theorem (relating to compression) and the noisy channel coding theorem (relating to communication over noisy channels). We will see these in this abstract. The second theorem requires rudimentary knowledge of channels also, so we will see to that as well.

1 Compression

We will consider our primary object of compression to be *files*. These may be plaintext, images, audios, videos, numerical data etc. There are two types of compressors:-

1. **Lossy**: Maps some files to the same encoding. Since we usually require recovery of every file that we compress, this type of compression may lead to failure occasionally. Even so, lossy compressors are pratically useful, since they are guaranteed to compress every file. Also, in image compression applications, lossy compression is considered satisfactory.

2. **Lossless**: Maps all files to different encodings, so the compressed file is always recoverable. A disadvantage is that if it shortens some files, it must necessarily make some others longer. Because this is forced, lossless compressor design often entails minimizing the probability that a file will be lengthened.

2 Bit Contents

Let X be a discrete random variable with finite image $\mathcal{X}$. Here the elements of $\mathcal{X}$ will play the role of the contents of the file, and our objective is to represent them as binary strings

i.e. strings consisting of only two symbols, 0 and 1.

Definition 2.1. *The* ***raw bit content*** *of X is $V_0(X) = \log_2 |\mathcal{X}|$.*

If every value that X can take were to be identified by a string of bits of equal length, $V_0(X)$ is a lower bound on that length (of course, equality being achieved when $|\mathcal{X}|$ is a power of 2). It is trivially shown that a lossless compressor that maps every value of X to a binary string of the same length less than $V_0(X)$ doesn't exist.
If the raw bit content of X is too large, lossless compression may become impractical, and we might want to consider lossy compression. If some outcomes are extremely improbable, we ignore them, saying that they will map to nothing ($\sim$ not be mapped). This will reduce our bit content. The **risk** δ is the probability that an outcome will not be mapped. It is upto our choice.

Definition 2.2. *A δ-**sufficient subset** is a subset S of $\mathcal{X}$ with $P(S) \geq 1 - \delta$.*

Denote by S_δ the smallest δ-sufficient subset.

Definition 2.3. *The δ-**essential bit content** of X is $V_\delta(X) = \log_2 |S_\delta|$.*

S_δ can easily be found by sorting the elements of $\mathcal{X}$ in descending order of probability, and then only considering the first few elements until the total probability is atleast $1 - \delta$.

3 Shannon's Source Coding Theorem

Theorem 3.1. *Let $X_1, X_2, \ldots, X_n$ be n i.i.d random variables all distributed as a discrete random variable X with finite image $\mathcal{X}$. Given $\epsilon > 0$ and $0 < \delta < 1$, there exists a natural number N such that*

$$\left| \frac{1}{n} V_\delta(X_1, X_2, \ldots, X_n) - H(X) \right| < \epsilon$$

$\forall\, n > N$.

For an idea of a proof which heavily uses typical sets, see Mackay's book from page 80.

- The upper bound of $H + \epsilon$ signifies that even if our risk δ is very small, the number of bits per symbol $\frac{1}{n} V_\delta(X_1, X_2, \ldots, X_n)$ required to specify long enough strings need not exceed $H + \epsilon$.

- The lower bound of $H - \epsilon$ signifies that even if our risk δ is very large, the number of bits per symbol required to specify long enough strings cannot fall below $H - \epsilon$ bits.

- **"As $n \to \infty$, n i.i.d random variables drawn from the distribution of a random variable X can be compressed into atleast $nH(X)$ bits with negligible information loss, and if they are compressed into less than $nH(X)$ bits, information is almost surely lost."**

4 Channels: Definitions and Examples

Definition 4.1. *A **channel** Γ is a tuple (A_X, A_Y, Q_Γ) where A_X and A_Y are sets of symbols (the **input** and **output alphabet** respectively), and Q_Γ is a **transition probability model**. This means that given random variables X and Y with images A_X and A_Y respectively, Q_Γ specifies all the conditional probabilities $P(Y = y \mid X = x)\, \forall x \in A_X,\, y \in A_Y$.*

Essentially, one must think that we are sending symbols of the input alphabet over the channel, and the receiver is receiving symbols of the output alphabet. In case the channel is noiseless and perfect, the input and output symbols always match. However, all real channels have atleast some amount of noise.

Definition 4.2. *A channel is called **memoryless** when the probability distribution over possible output symbols at time t is determined solely by the input symbol sent at time t (and is independent of any inputs/outputs prior/post time t).*

Definition 4.3. *In the case that both A_X and A_Y are finite, Γ is called a **discrete channel**.*

We will deal with only discrete memoryless channels (DMCs). For such channels, Q_Γ is typically represented as a $|A_X| \times |A_Y|$ matrix with

$$(Q_\Gamma)_{ij} = P(\text{output symbol } = b_j \mid \text{ input symbol } = a_i)$$

Exercise 4.1. *Let $\vec{p_X}$ and $\vec{p_Y}$ be the probability mass functions (represented as vectors) of X and Y respectively for a DMC. Verify that $\vec{p_Y} = \vec{p_X} Q_\Gamma$.*

5 Examples of Channels

We look at some binary DMCs, which refer to DMCs with $|A_X| = 2$. The transition probability matrices are typically specified in terms of a parameter $f \in (0, 1)$.

- **Binary Symmetric Channel (BSC)**: $A_X = \{0, 1\} = A_Y$,

  $$Q_{BSC}(f) = \begin{pmatrix} 1-f & f \\ f & 1-f \end{pmatrix}$$

  Bits are flipped with probability f and preserved with probability $1 - f$. Quite simply, the probability that I receive a 1 if I have sent a 0 (or receive a 0 if I have sent a 1) is f. Ideally we would want f to be as small as possible.

- **Binary Erasure Channel (BEC)**: $A_X = \{0, 1\}$, $A_Y = \{0, 1, ?\}$,

  $$Q_{BEC}(f) = \begin{pmatrix} 1-f & 0 & f \\ 0 & 1-f & f \end{pmatrix}$$

  Bits are never flipped; just preserved or erased with probability f ("?" is the symbol for erasure).

- **Z Channel**: $A_X = \{0, 1\} = A_Y$,

$$Q_Z(f) = \begin{pmatrix} 1 & 0 \\ f & 1-f \end{pmatrix}$$

'0's are never flipped, '1's are flipped with probability f.

6 Capacity of a Channel

The mutual information between the source and the received signal emerges as a natural measure of the information transmitted through the channel. Clearly, this depends on the input distribution $\vec{p_X}$ only, since $\vec{p_Y}$ can be readily calculated given $\vec{p^*}_{X,\Gamma}$ and Q_Γ.

Definition 6.1. *The **capacity** of a DMC Γ is defined as*

$$C(\Gamma) = \max_{\vec{p_X}} I(X;Y)$$

The distribution that achieves this is denoted as $\vec{p_X}^*$.

For binary DMCs, finding the capacity is simple: assume $\vec{p_X} = (p, 1-p)$, express $I(X;Y)$ as a function of p and then maximize it.

1. **BSC:**

$$\begin{aligned} \vec{p_X} &= (p, 1-p) \\ \Rightarrow \vec{p_Y} &= \vec{p_X} Q_{BSC}(f) = (p + f - 2pf, 1 - (p + f - 2pf)) \end{aligned}$$

So $H(Y) = H_2(p + f - 2pf)$. Now we need $H(Y|X)$.

$$\begin{aligned} H(Y|X) &= pH(Y|X=0) + (1-p)H(Y|X=1) \\ &= H_2(f) \end{aligned}$$

since $H(Y|x=0) = H(Y|x=1) = H_2(f)$. We get $I(X;Y) = H(Y) - H(Y|X) = H_2(p + f - 2pf) - H_2(f)$, say. Then, putting

$$\frac{\partial I(X;Y)}{\partial p} = 0$$

we get $(1-2f)H_2'(p + f - 2pf) = 0$. If $f = 0.5$, the bits are being flipped randomly, and hence absolutely no information is being sent. So assume $f \neq 0.5$. Since $H_2(x)$ maximizes at $x = 0.5$, the maximum of $I(X;Y)$ occurs at $p + f - 2pf = 0.5$, or $(2f-1)(p-0.5) = 0$ i.e. at $p = 0.5$.
This means that the **capacity of a BSC** is $H_2(0.5) - H_2(f) = 1 - H_2(f)$, with $\vec{p^*}_{X,BSC} = (0.5, 0.5)$.

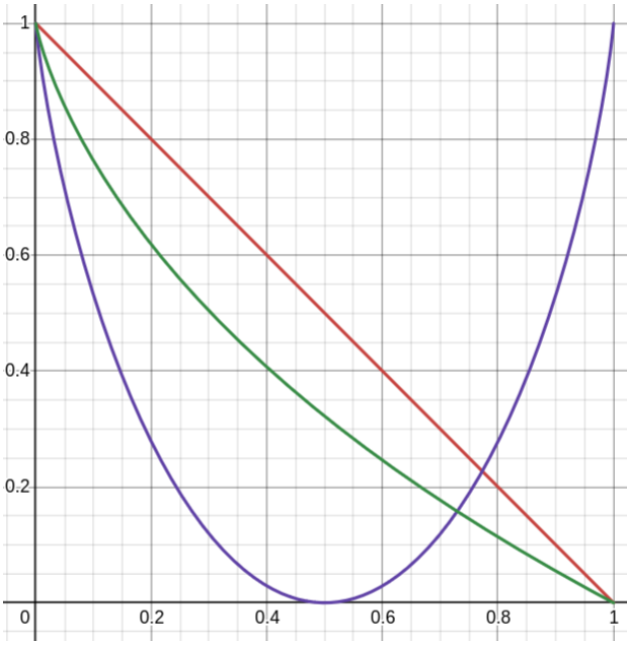

Figure 1: The capacities for three binary DMCs, plotted as a function of the noise level f; red is BEC, purple is BSC, and green is Z.

2. **BEC:**

$$\begin{aligned} p\vec{_Y} &= (p(1-f), (1-p)(1-f), f) \\ \text{and } H(Y|X) &= pH_2(f) + (1-p)H_2(f) = H_2(f) \\ \Rightarrow I(X;Y) &= H(p(1-f), (1-p)(1-f), f) - H_2(f) \\ &= H(1-f, f) + (1-f)H(p, 1-p) + fH(1) - H_2(f) \quad \text{(using decomposability)} \\ &= (1-f)H_2(p) \end{aligned}$$

Clearly maximum at $p = 0.5$. So **the capacity of a BEC is simply** $1 - f$, with $\vec{p^*}_{X,BEC} = (0.5, 0.5)$.

3. **Z:**

Exercise 6.1. *Show that the capacity of a Z channel is*

$$H_2(p^* + f - p^* f) - (1 - p^*)H_2(f)$$

and $\vec{p^*}_{X,Z} = (p^*, 1 - p^*)$*, where*

$$p^* = \frac{1}{1-f}\left(\frac{1}{1 + 2^{\left(-\frac{H_2(f)}{1-f}\right)}} - f\right)$$

7 The Noisy Channel Coding Theorem

A precise definition would require definitions of codes and their rates as well, which I ought not to cover since Nirav is going to be doing so in depth soon. I will settle for giving you some intuition for the theorem. It states that reliable (meaning that we can make the probability of "block error" in "decoding" arbitrarily small) communication is possible across a channel Γ at any rate less than its capacity $C(\Gamma)$ (as defined above) and that it is impossible at rates more than it. This tight upper bound becomes $\frac{C(\Gamma)}{1-H_2(p_b)}$ if a probability of "bit error" upto p_b is possible. Of course, you are encouraged to lookup what "bit errors" and "block errors" are.