A TALK ON INFORMATION THEORY

EGOR LARIONOV (20263767)

UNIVERSITY OF WATERLOO

Abstract. The goal of this talk is to introduce the idea of classical and quantum information as it appears in the theory of communication and more generally, information theory as proposed by Claude Shannon in 1948. We address the fundamental problem "of reproducing at one point either exactly or approximately a message selected at another point."[1] We will define and motivate the idea of entropy as a measure of information in order to quantify the ability for communication channels to transmit information reliably. In addition we will extend the definition of entropy from classical states to general quantum states.

This talk assumes an audience with basic linear algebra background. A familiarity with bilinear maps, linear operators and basic analysis may be helpful.

1. Introduction

People have been studying information for decades. Transmission of information through signals over long distances, storage of information on paper, on magnetic strips, and data compression, are all part of this study among other important things. With the invention of quantum computing came a generalization of the mathematical model of information theory, as quantum mechanics introduces new properties to information.

This talk will introduce the mathematical description of communication through classical and quantum channels. In particular we will discuss the capacity of a particular classical channel, and time-permitting quantum capacities of quantum channels. The purpose of this talk is to briefly introduce the theory of communication, and to build a general understanding of information.

1.1. Introduction to information theory.

Information theory was invented by Claude Shannon in his paper from 1948 titled "*A mathematical theory of communication*" [1], where he solved a lot of the problems involved in classical information theory. The main problem in his theory is clearly outlined by the following quote:

> "The fundamental problem is that of reproducing at one point either exactly or approximately a message selected at another point."
>
> (Claude Shannon, 1948)

Real channels are unreliable because the physical noise in the environment disrupts the signal being transmitted (e.g. thermal fluctuations, foreign interacting signals). For example:

$$\begin{array}{c} \text{Tower} \xrightarrow{\text{radio waves}} \text{Receiver} \\ \text{Voice} \xrightarrow{\text{sound waves}} \text{Ear} \\ \text{Modem} \xrightarrow{\text{wire}} \text{Modem} \\ \text{File} \xrightarrow{\text{disk drive}} \text{File} \end{array}$$

Therefore we need to develop methods to transmit information reliably through such channels, by introducing some redundancy and clever methods.

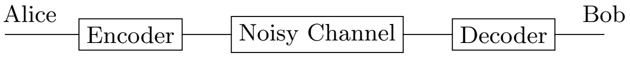


Figure 1.1. Using a noisy channel for reliable communication

Consider the simplest noisy classical channel, the binary symmetric channel (BSC), which transmits a bit from Alice to Bob, where it gets flipped with probability p.

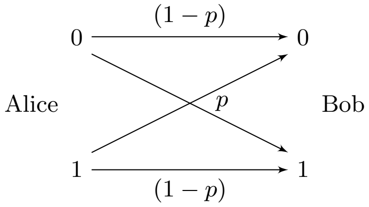


Figure 1.2. Binary symmetric channel.

This is not a realistic channel, just a mathematical model of a channel; realistic channels often have correlation between consecutive uses, or uses of the channel in parallel. It is not enough to send a single bit over this channel to get reliable communication if p happens to be sufficiently high. Say $p = 0.1$, then 10% of the time we will get an incorrect bit of information on the receiving end. We need to develop a system that can provide us with more reliable communication, without modifying the channel. For instance consider a *repetition code*. For every bit of data to be sent, we duplicate it three times and send those three through the channel, so that we determine what the original bit was on the receiving end by taking a majority vote over the transmitted bits. Now if we send one bit with this encoding, we can have the following outcomes:

$$0 \xrightarrow{\text{Encoding}} 000 \xrightarrow[\text{Channel}]{\text{Noisy}} \begin{cases} 000, 001, 010, 100 \to \text{correctable} \\ 111, 110, 101, 011 \to \text{erroneous} \end{cases}$$

$$1 \xrightarrow{\text{Encoding}} 111 \xrightarrow[\text{Channel}]{\text{Noisy}} \begin{cases} 111, 110, 101, 011 \to \text{correctable} \\ 000, 001, 010, 100 \to \text{erroneous} \end{cases}$$

Meaning that the probability of an erroneous outcome will be $p_e = p^3 + 3p^2(1-p) = 3p^2 - 2p^3$. So if $p = 0.1$ as before, then $p_e = 0.028$, which is much less than 0.1, so we have improved the reliability of data transmission for this channel.

In order to quantify how effective the channel is for communication, consider the *rate of communication*, $R = m/n$, which is the number of bits we would like to send m, divided by the number of times we have used the channel (or the size of the encoding), n. We demonstrated how we can decrease the probability of error by decreasing the rate of communication of our channel using a particular error correcting code. A point (R, p_e) for which there exists a corresponding error correcting code is called *achievable*. Now the question is, what is the best (highest) rate at which we can communicate information reliably (meaning with arbitrarily small error). If we follow the repetition codes as plotted in Figure 1.3, it would seem that we must decrease the rate to zero in order to get reliable communication. However in 1948, Claude Shannon showed that this is not true in general, and the supremum of achievable rates that provide reliable communication (also called the *capacity* of the channel) is NOT zero for this binary symmetric channel, which is surprising because it contradicts our intuition that we can't get something for nothing. For those familiar with mathematical analysis, if we consider the maximum achievable rate of communication given an error probability as a

Date: March 24, 2013.

function $R_{\max}(p_e)$, then the capacity C_X, of the channel X, can be written mathematically as

$$C_X := \limsup_{p_e \to 0} R_{\max}(p_e) = \lim_{\epsilon \to 0} \left(\ \sup\{R_{\max}(p_e) : 0 < p_e < \epsilon\}\ \right).$$

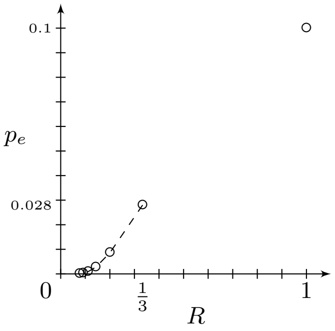


FIGURE 1.3. Probability of error (p_e) versus rate of communication (R) plot. As the probability of error approaches zero, so does the rate, in a repetition coding scheme.

We will examine the reason for this non-zero capacity by first learning about *entropy* – the measure of information.

1.2. **The twelve ball problem.** Consider the following problem. You are given 12 tennis balls and a balance scale. One of the 12 balls weighs differently than the other 11. The problem is to determine which ball it is and whether it weighs more or less than the other 11. The challenge is to do this in as little weighings as possible.

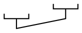
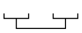
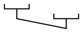
For our first weighing we can choose one of a few possibilities, which will yield different outcome probability distributions. Note that this process is modelled by a discrete random variable with three outcomes, each with its own probability. We claim that the correct measure of information content of an outcome in a random variable is given by $h(k) = -\log_2(p_k)$ where $k \in \{1, 2, 3\}$ indexes one of the outcomes. Therefore in order to choose the best possible first weighing, we should choose the one giving the largest amount of information on average. Thus we define the Shannon entropy to be

$$H(\{p_k\}) \equiv -\sum_k p_k \log(p_k), \tag{1.1}$$

the information content of a discrete random variable. We can interpret entropy in two complementary ways [2]:

- as a measure of our uncertainty *before* we learn the value of the random variable, or
- as a measure of how much information we have gained *after* we learn the value of the random variable.

To convince ourselves of the validity of our choice of information measure, consider all the possiblities of the first weighing:

				$-\sum_k p_k \log(p_k)$
6 and 6	1/2	0	1/2	1.00
5 and 5	5/12	1/6	5/12	1.48
4 and 4	1/3	1/3	1/3	1.58
3 and 3	1/4	1/2	1/4	1.50

We can see that the most uniform distribution yields the largest entropy. This is true in general for any discrete random variable. Similarly we can compute the entropy of subsequent weighings and choose the one with the largest entropy, in order to solve the problem.

The rest of the problem is left as an exercise.

1.3. **Capacity of the BSC.** Recall the example of the binary symmetric channel. In particular, recall that the capacity of a channel, C_{BSC}, is the supremum over rates at which the channel can transmit information with arbitrarily low probability of error. Hence no more than C bits of information can be transmitted per channel use (no matter how small the probability of error is), so we must lose at least $1 - C$ bits of information in the process. That's precisely the entropy of the binary symmetric channel:

$$1 - C_{\text{BSC}} = H_{\text{BSC}}(\{p_k\}).$$

This gives the capacity of the BSC:

$$C_{\text{BSC}} = 1 - H_{\text{BSC}}(\{p_k\}). \tag{1.2}$$

The best interpretation for entropy here is that of a measure of uncertainty before we learn the value of the random variable. Clearly we would like to minimize this uncertainty to be able to transmit information reliably. Alternatively we can think of the entropy as the amount of information we gain about the process, which we also would like to minimize, because ideally we want to expect the channel to act as the identity.

A more rigorous discussion of classical capacity and the proof of (1.2), can be found in Shannon's paper [1].

1.4. **Quantum Mechanics.** So far we have considered only classical information. We will now briefly introduce the motivation for quantum information, and its mathematical description. We consider a simple experiment that demonstrates quantum effects that cannot be explained with classical physics.

First suppose that a photon travels from a photon emitter, towards a translucent mirror (called the beam splitter), which reflects the photon 50% of the time, otherwise the photon is not affected. By placing two photon detectors at each end, we will observe that each detector will go off 50% of the time as expected. The easiest way to explain this is to imagine that the beam splitter transmits or reflects the photon with probability 0.5. That is the beam splitter acts as a binary random variable.

Now suppose that, after guiding the two possible paths, of equal length, into another beam splitter with perfect mirrors, we put two photon detectors on either side of the second beam splitter as shown in Figure 1.4.

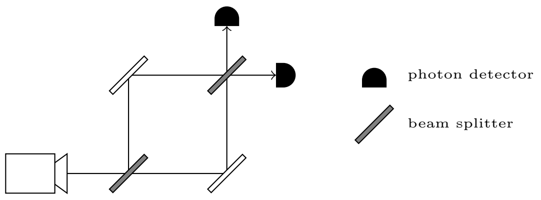


FIGURE 1.4. Beam splitter experiment.

Experiment shows an unexpected result that 100% of the time, the right detector will go off in this setup. Suppose we were to predict what would happen at the detectors by modelling the beam splitter with a random variable as mentioned. Then it is easy to see that both detectors will fire with probability 0.5, which clearly contradicts the experimental result.

This suggests that such a setup cannot be explained classically and we need another mathematical model to explain this phenomenon. Quantum physics models this experiment correctly, with the concepts of *interference* and *superposition*. Consider the two paths that the photon could take given that the second beam splitter is removed, colour coded by green and red in Figure 1.5. Now our system can be in one of two states, which we will label with two vectors in $\mathbb{C}^2$: $\begin{pmatrix}1\\0\end{pmatrix}$ if the photon follows the red path and $\begin{pmatrix}0\\1\end{pmatrix}$ if the photon follows the green path.

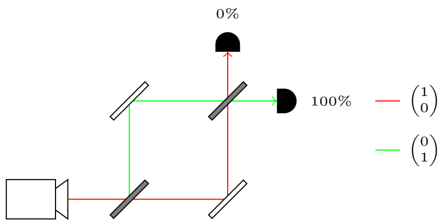


FIGURE 1.5. Beam splitter experiment.

According to quantum mechanics, the beam splitter causes the photon to go into a *superposition* of the two available states. Mathematically a superposition is simply a complex combination of orthonormal state vectors, such as the ones we have given. A general superposition is written as

$$\alpha_0 \begin{pmatrix}1\\0\end{pmatrix} + \alpha_1 \begin{pmatrix}0\\1\end{pmatrix},$$

where $\alpha_0, \alpha_1 \in \mathbb{C}$. When we observe the photon immediately after the first beam splitter, the detector on the red path will fire with probability $|\alpha_0|^2$ and the detector on the green path will fire with probability $|\alpha_1|^2$. Since these are the only outcomes, it must be that $|\alpha_0|^2 + |\alpha_1|^2 = 1$.

As the photon moves through a beam splitter its state is changed according to the following matrix transformation:

$$\frac{1}{\sqrt{2}}\begin{pmatrix}1 & i\\ i & 1\end{pmatrix}$$

Giving the following evolution of the state of our system as the photon passes through the first beam splitter:

$$\frac{1}{\sqrt{2}}\begin{pmatrix}1 & i\\ i & 1\end{pmatrix}\begin{pmatrix}1\\0\end{pmatrix} = \frac{1}{\sqrt{2}}\begin{pmatrix}1\\i\end{pmatrix} = \frac{1}{\sqrt{2}}\begin{pmatrix}1\\0\end{pmatrix} + \frac{i}{\sqrt{2}}\begin{pmatrix}0\\1\end{pmatrix}$$

And the second beam splitter:

$$\frac{1}{\sqrt{2}}\begin{pmatrix}1 & i\\ i & 1\end{pmatrix}\frac{1}{\sqrt{2}}\begin{pmatrix}1\\i\end{pmatrix} = i\begin{pmatrix}0\\1\end{pmatrix}$$

Note that if we observe the system at the end, we find it in the state $\begin{pmatrix}0\\1\end{pmatrix}$ with probability $|i|^2 = 1$, thus correctly explaining the experiment. The reason why we have have that 100% of the time only one of the detectors fire, is because of *interference*. The photon interferes with its alternate path at the second beam splitter in such a way that, it will follow the green path after the second beam splitter 100% of the time. It is important to notice that the path lengths between the two splitters are the same, otherwise the result will be different.

The state of the photon we have just described is called a quantum state, and it can be quantified with a quantum bit: a unit vector in $\mathbb{C}^2$; also known as a *qubit*. A quantum state can be described in terms of multiple qubits, which are written as a tensor (or Kronecker in particular) product of constituent qubits. This product is denoted by $\otimes$. Think of the tensor product as simply a general bilinear operation.

Note that we can build a system which will generate different classical bits with various probabilities. This is *not* the same as generating qubits in superposition, and we have illustrated this with the the beam splitter experiment. So naturally, we have an analogue of probability distributions of bits in quantum computing. A quantum state can be described by a probability distribution of various qubits $\{(p_k, \phi_k)\}_k$, where $p_k \in [0,1]$ and $\phi_k \in \mathbb{C}^2$. Such a state can be precisely described by what's called a density matrix $\rho = \sum_k p_k \phi_k \phi_k^*$, where ϕ_k^* is the adjoint (or complex conjugate transpose) of the column vector ϕ_k. For instance if we have a distribution

$$\left\{\left(\frac{1}{3}, \begin{pmatrix}1\\0\end{pmatrix}\right), \left(\frac{2}{3}, \frac{1}{\sqrt{2}}\begin{pmatrix}1\\0\end{pmatrix} + \frac{i}{\sqrt{2}}\begin{pmatrix}0\\1\end{pmatrix}\right)\right\},$$

then its corresponding density matrix will be

$$\rho = \frac{1}{3}\begin{pmatrix}1\\0\end{pmatrix}\begin{pmatrix}1\\0\end{pmatrix}^* + \frac{2}{3}\left[\tfrac{1}{\sqrt{2}}\begin{pmatrix}1\\0\end{pmatrix} + \tfrac{i}{\sqrt{2}}\begin{pmatrix}0\\1\end{pmatrix}\right]\left[\tfrac{1}{\sqrt{2}}\begin{pmatrix}1\\0\end{pmatrix}^* - \tfrac{i}{\sqrt{2}}\begin{pmatrix}0\\1\end{pmatrix}^*\right]$$
$$= \frac{1}{3}\begin{pmatrix}2 & 1\\ 1 & 1\end{pmatrix}.$$

We can clearly see that ρ is a positive semidefinite operator with unit trace. This is true for any density matrix.

1.5. **Types of Linear Operators.** We will use $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$ to denote Hilbert spaces of the form $\mathbb{C}^n$ where n is a positive integer. We denote the set of linear operators of the form $A : \mathcal{X} \to \mathcal{Y}$ by $\mathrm{L}(\mathcal{X}, \mathcal{Y})$, and $\mathrm{L}(\mathcal{X})$ for short, when $\mathcal{Y}$ is the same as $\mathcal{X}$. Now consider the following definitions [3].

Definition 1.1. Let $A \in \mathrm{L}(\mathcal{X})$, then it is

$$\textit{normal} \text{ if } \qquad AA^\dagger = A^\dagger A,$$
$$\textit{Hermitian} \text{ if } \qquad A = A^\dagger.$$

Definition 1.2. A linear operator A is called *positive semidefinite* if it's Hermitian and has only non-negative eigenvalues.

Definition 1.3. A *density operator* is a positive semidefinite operator with unit trace. $\mathrm{D}(\mathcal{X})$ denotes the set of all such operators.

Remark 1.4. Notice that $\mathrm{D}(\mathcal{X})$ are precisely the quantum states described before. The Spectral Theorem implies that a linear operator is positive semidefinite and has unit trace if and only if it can be written as a convex combination of projectors (with unit trace).

1.6. **Quantum Information.**

Definition 1.5. The classical (or Shannon) entropy of a probability distribution $\{p_x\}$, is defined as

$$H(\{p_x\}) := -\sum_x p_x \log p_x$$

Recall that entropy measures the amount information gained from learning the value of a random variable given a probability distribution. So consider a classical probability distribution of possible states some piece of n-bit binary memory (for instance) can be in: $\{(p_x, x)\}$ where $x \in \{0,1\}^n$, for example

$$\left\{\left(\frac{1}{2}, 01\right), \left(\frac{1}{3}, 10\right), \left(\frac{1}{6}, 11\right)\right\},$$

then the entropy in this piece of memory is

$$H(\{p_x\}) = -\tfrac{1}{2}\log\tfrac{1}{2} - \tfrac{1}{3}\log\tfrac{1}{3} - \tfrac{1}{6}\log\tfrac{1}{6} \cong 0.439.$$

Of course, if we know the value in the memory exactly, then entropy will be 0, since $\log(1) = 0$.

Now imagine that our memory stores qubits...

REFERENCES

[1] C. Shannon. *A mathematical theory of communication.* ACM SIGMOBILE Mobile Computing and Communications Review, 5(1), pp. 3–55 (2001).

[2] M. A. Nielsen and I. L. Chuang. *Quantum Computation and Quantum Information: 10th Anniversary Edition.* Cambridge University Press (2010). ISBN 9781107002173.

[3] J. Watrous. *Theory of quantum information* (2011). Lecture notes from the course on quantum information hosted at the Institute for Quantum Computing.