Quantum Information Theory

quinten tupker

January 22 2021 - January 29, 2021

Introduction

These notes are based on the course lectured by Professor S Strelchuk in Lent 2020. This was lectured online due to measures taken to counter the spread of Covid-19 in the UK. These are not necessarily an accurate representation of what was lectures, and represent solely my personal notes on the content of the course, combinged with probably, very very many personal notes and digressions... Of course, any corrections/comments would be appreciated.

Information theory is the theory of information storage and transmission. It provides the theoretical limits on what is possible with information technologies in much of our world, and a framework to study many other fields, such as animal communication as well. This is the quantum qersion of that theory.

1 Classical Information Theory

We begin by observing that information is closely related to uncertainty. In particular, what might be able to say it is the opposite of uncertainty, and so then, it is no surprise that we build our theory of information using concepts from probability theory. As such we define

Definition 1. The **surprisal** of random variable X taking values in discrete finite **alphabet** J according to distribution $p(x)$ is

$$\mathcal{I}(x) = -\ln(p(x))$$

Definition 2. The **Shannon entropy** of a random variable X is

$$H(X) = -\sum_{x\in J} p(x)\ln(p(x))$$

(the logarithm is base 2)

This may appear to be a somewhat arbitrary definition, but it has a strong theoretical basis given that

Theorem 1. *The **Shannon Source Coding Theorem** states (informally) that the limit that information can be compressed so that it can be reliably retrieved is the Shannon entropy of the source.*

Here a basic example of a source is a **memoryless source** which is an object producing a sequence of signals, but since it is memoryless, each signal is completely independent from any other, so $\mathbb{P}(u_1, \ldots, u_n) = \mathbb{P}(u_1) \ldots \mathbb{P}(u_n)$. It is also known as an **i.i.d. information source**.

But how do we actually compress information? Conceptually there are two ways

- a **variable length encoding** stores higher probability signals in shorter codes, and lower probability signals in longer codes.
- a **fixed length encoding** stores higher probability signals in unique fixed length codes, and lower probability signals in the same fixed code.

Example 1. *An example of a fixed length code for the numbers* $1, \ldots, 8$ *are their binary representations, but if we also know that* $p(1, \ldots, 8) = 1/2, 1/4, 1/8, 1/16, 1/64, 1/64, 1/64, 1/64$ *then the code* $C(1, \ldots, 8) = 0, 10, 110, 1110, 111100, 111101, 111110, 111111$ *has an average length 2 compared to the fixed length of 3. Furthermore, the Shannon entropy of this source is 2 as well, so this is maximally efficient.*

1.1 Classical Data Compression

Let's start making the definitions necessary to formalise compression.

Definition 3. A **compression map** is a map $C^n : u^{(n)} = (u_1, \ldots, u_n) \mapsto x(x_1, \ldots, x_{nR})$ sending a **message** u to a **code** x.

Definition 4. A **decompression map** D^n sends $D^n : x \in \{0,1\}^{\lceil nR \rceil} \mapsto u'^{(n)}$ with probability $\mathbb{P}(u^{(n)}|x)$.

Definition 5. A **code** of rate R and blocklength n is the triple (C^n, D^n, R).

Here we can compute the **probability of error** as

$$P_{av}^{(n)}(C_n) = \sum_{u^{(n)} \in J^n} \mathbb{P}(u^{(n)}) \mathbb{P}(D^n(C^n(u^{(n)})) \neq u^{(n)}). \tag{1}$$

[End of lecture 1]

Definition 6. A compression/decompression scheme is **reliable** iff $\forall \epsilon > 0 \exists$ sequence of codes C_n such that $\lim_{n\to\infty} P_{av}^{(n)}(C_n) = 0$

(does this require the code to i.i.d?) As such one could define the data compression rate as

$$\inf\{R : \exists C_n = \{C^n, D^n, R\} : \lim_{n\to\infty} P_{av}^{(n)}(C_n) = 0\} \tag{2}$$

1.2 Typical Sequences

It is honestly surprising how easy it is to prove Shannon's source coding theorem, since we find we in fact only need one simple, and honestly rather crude tool to do so. That tool is

Definition 7. a **typical set**, denoted $T_\epsilon^{(n)}$, which is the set of sequences $u = (u_1, \ldots, u_n)$ satisfying

$$2^{-n(H(u)+\epsilon)} \leq \mathbb{P}(u) \leq 2^{-n(H(u)-\epsilon)} \tag{3}$$

Why "typical"? Because if we have a memoryless source generating a sequence of length n, then a "typical" sequence occurs with probability

$$\prod_{u \in J} \mathbb{P}(u)^{n\mathbb{P}(u)} = 2^{-nH(u)} \tag{4}$$

(a typical sequence would have the expected value as the number of occurences of each letter in the alphabet). We consequently find the following theorem (not proven)

Theorem 2. $\forall \epsilon, \delta > 0, \exists n$ *such that*

1. $u \in T_\epsilon^{(n)} \implies H(u) - \epsilon \leq \frac{-1}{n} \log(\mathbb{P}(u)) \leq H(u) + \epsilon$
2. $\mathbb{P}(T_\epsilon^{(n)}) > 1 - 3$
3. $|T_\epsilon^{(n)}| \leq 2^{n(H(u)+\epsilon)}$
4. $|T_\epsilon^{(n)} > (1-\delta)2^{2(H(u)-\epsilon)}$

and so

Corollary 1. $\forall \epsilon, \delta > 0, \exists n_0, \forall n > n_0, J^n$ *decomposes into the disjoint* ***atypical and typical sets*** $A_\epsilon^{(n)}, T_\epsilon^{(n)}$ *satisfying*

1. $\mathbb{P}(A_\epsilon^{(n)}) < \delta$
2. $2^{-n(H(u)+\epsilon)} \leq \mathbb{P}(u) \leq 2^{-n(H(n)-\epsilon)}$ *(on the typical set?)*

That allows us to formally state

Theorem 3. ***Shannon's source coding theorem****, which claims that for a i.i.d. source U, if $R > H(U)$, then we can find a reliable compression scheme and if $R < H(U)$ there are no reliable compression schemes.*

Proof. If $R > H(U)$ then pick $\epsilon > 0$ such that $H(U) + \epsilon < R$ and n such that $T_\epsilon^{(n)}$ satisfies our typical set theorem condiions. Then for $\delta > 0$ there are at most $2^{n(H(U)+\epsilon)} < 2^{nR}$ ϵ-typical sequences. Our compression scheme then works according to

1. Split J^n into typical and atypical sequences
2. order typical sequences somehow (say lexicographically), assigning each sequence a binary index.
3. typical sequences are sent to their binary code, prefixed with a 1, leading to a total length of $\lceil nR \rceil + 1$
4. atypical sequences are all sent to the fixed string $00\ldots0$ of length $\lceil nR \rceil + 1$

If $R < H(U)$ there does not exist a reliable compression scheme (see lemma below). □

Lemma 1. *For a collection of strings of length n, $S(n)$, with $|S(n)| \leq 2^{nR}, R < H(U)$. Then, $\forall \delta > 0, \exists n, \sum_{u \in S(n)} \mathbb{P}(u) < \delta$*

Basically, when $R < H(U)$ we will always get some typical sets that are indistinguishable after compression. Assume $S(n)$ is a set of such indistinguishables. This lemma states that these become less significant?

Proof. Again, we split $S(n)$ into its typical and atypical part and then observe

$$\begin{aligned}
\mathbb{P}(S(n)) &= \sum_{u \in S(n)} \mathbb{P}(u) \\
&= \sum_{u \in S(n) \cap T_\epsilon^{(n)}} \mathbb{P}(u) + \sum_{u \in S(n) \cap A_\epsilon^{(n)}} \mathbb{P}(u) \\
&\leq |S(n)| 2^{-n(H(n)-\epsilon)} + \mathbb{P}(A_\epsilon^{(n)}) \\
&\leq 2^{-n(H(n)-R)} + \mathbb{A}_\epsilon^{(\kappa)}
\end{aligned}$$

□

which completes the proof of Shannon's source coding theorem. Finally, we note that using a variable length encoding scheme or something like that does not fundamentally change this result. [End of lecture 2]

Today we discuss the types of relationships one can have between various types of entropy. As such,

Definition 8. the **joint entropy of** X, Y is

$$H(X,Y) = -\sum_{x \in J_X} \sum_{y \in J_Y} p(x,y) \log(p(x,y)) \tag{5}$$

Definition 9. the **conditional entropy** is

$$H(Y|X) = -\sum_{x \in J_X} \sum_{y \in J_Y} p(x,y) \log(p(y|x)) = -\sum_{x \in J_X} p(x) H(Y|X=x) \tag{6}$$

where we note the "chain rule"

$$H(X,Y) = H(Y|X) + H(X) \tag{7}$$

which generalises nicely as

$$H(X,Y,Z) = H(X) + H(Y|X) + H(Z|Y,X) \tag{8}$$

and similar for higher numbers of variables.

Definition 10. We further define p to be **absolutely continuous** wrt q iff $q(x) = 0 \implies p(x) = 0$ or equivalently $\text{supp}(p) \subseteq \text{supp}(q)$ which we can denote as $p << q$.

and

Definition 11. the **relative entropy** or **Kullback-Leibler divergence** of $p << q$ to be

$$D(p||q) = \sum_{x \in J} p(x) \log(p(x)/q(x)) \tag{9}$$

Note that if $q(x) = 1 \forall x \in J$ then $D(p||q) = -H(X)$ so this is stronger than the Shannon entropy.

This measures how different two distributions are in a certain sense, however, it certainly is not a metric (not symmetric, and no triangle inequality).

Definition 12. We similarly define the **mutual information** between X, Y to be

$$I(X:Y) = H(X) + H(Y) - H(X,Y) = H(X) - H(X|Y) \tag{10}$$

Definition 13. and the **conditional mutual information** (CMI)

$$I(X:Y|Z) = H(X|Z) - H(X|Y,Z) \tag{11}$$

Intuition wise, one can think of $H(X,Y)$ as adding entropy, $I(X:Y)$ as taking the intersection between the two, $D(p||q)$ as measuring the difference between one distribution contained in another, and $H(Y|X)$ as removing the overlap of one distribution into another ($H(Y|X) = H(Y) - I(Y:X)$). This also explains the commutativeness of these various operations. I'm not sure about the CMI yet...

From here we can state

Theorem 4. *the data processing inequality for Markov Chain $X \to Y \to Z$:*

$$I(X:Y) \geq I(X:Z). \tag{12}$$

If we imagine X as a perfect source, Y as an observed, noisy signal, and Z as a "cleaned up" version of Y after some data processing, then this states that no matter the data processing used to clean up Y, Z can never contain more information about X than Y.

We finally have the following theorem.

Theorem 5. *1.* $p << q \implies D(p||q) \geq 0$ *with equality iff* $p = q$

2. $H(x) \geq 0$ *with equality iff* X *deterministic*

3. $H(X|Y) \geq 0$ *or equivalently* $H(X,Y) \geq H(Y)$

4. $H(X) \leq \log(|J|)$

5. $H(X,Y) \leq H(X) + H(Y)$ *or equivalently* $H(Y) \geq H(Y|X)$ *with equality iff* $X \perp Y$

6. H *is concave meaning* $H(\lambda p_x + (1-\lambda)p_y) \geq \lambda H(p_x) + (1-\lambda)H(p_Y)$

7. $I(X:Y) \geq 0$ *and equal iff* $X \perp Y$

All of these are exercises on the example sheet. [End of lecture 4]

Classical Information Transmission

Let's build the infrastructure for a bound on reliable information transmission.

Definition 14. A **discrete channel** is a combination of

- discrete alphabets J_X, J_Y
- a set of conditional probabilities $\mathbb{P}(y_1, \ldots, y_n | x_1, \ldots x_m)$

Definition 15. A memoryless channel is one satisfying

$$\mathbb{P}(y_1, \ldots, y_n | x_1, \ldots, x_n) = \prod_{i=1}^{n} \mathbb{P}(y_i | x_i) \tag{13}$$

These are completely characterised by the so-called **channel matrix** $P = \mathbb{P}(y|x)$.

Example 2. *For example we can consider a* ***binary symmetric channel*** *for* $\mathbb{P}(0|0) = \mathbb{P}(1|1) = 1-p$ *where* $p_{err} = 3p^2(1-p) + p^3 = 3p^2 - 2p^3$ *if we use the redundant codingg* $0 \mapsto 000, 1 \mapsto 111$.

Definition 16. Now for messages $m = x_1, \ldots, x_n$ and $m' = y_1, \ldots, y_m$ and $m \in [M] = \{1, \ldots, M\}$ we **encoding** $E_n : [M] \to J_X^n$, **decoder** $D_n : [M] \to J_X^n$, and rate of encoding R such that $M = \lfloor 2^{nR} \rfloor$. The triple $C_n = (E_n, D_n, R)$ is then called an **error correcting code**.

Definition 17. and unlike the average error probabilities consider in information storage, here the error probability is the maximum

$$P_{err}(C_n) = \max_{m \in [M]} \mathbb{P}(D_n(Y^{(n)}) \neq m | X^{(n)} = E_n(m)) \tag{14}$$

Definition 18. and rate R is called **achievable** if there exists an error correcting code such that

$$\lim_{n\to\infty} P_{err}(C_n) = 0 \tag{15}$$

and so quite naturally

Definition 19. the **capacity** of a discrete memoryless channel is $C(N) = \sup\{R : R \text{ is achievable}\}$.

for which

Theorem 6. ***Shannon's noisy channel coding theorem*** *states that*

$$C(N) = \max_{p(x)} I(X : Y) \tag{16}$$

(note that the maximum is only taken over the input, not the output distribution)

Theorem 7. *which has the properties*

- $C(N) \geq 0$
- $C(X) \leq \log|J_X|, \log|J_Y|$

rigorous proof of which is found in the 1991 book *Elements of Information Theory*.

Example 3. *Finally, in the binary symmetric channel from before we find that*

$$I(X : Y) = H(Y) - H(Y|X) = H(Y) - h(p) \implies C(N) = 1 - h(p) \tag{17}$$

[End of lecture 4]