February 18, 2022

Notes on Information Theory

Aayush Verma[*]

Abstract

In the following notes, we discuss information theory, classical and quantum. We discuss Shanon entropy and its mathematical properties which include mutual information relative information, and Holevo bound. Shanon entropy is discussed using an ensemble. We finally comment on noise and errors in a communication.

[*]Email aayushverma6380@gmail.com

Contents

1 Introduction

Before we discuss the classical information theory, let us have a look at "what information theory is in physics?" or at least in our context. Let us consider that Alice has sent a message to Bob. What parts of the message Bob has received, what is the rate of the communication, relative communication? These are ideas which we need to discuss. The main ideas are [1]

1. What is the ultimate compression of the message?
2. What is the rate of communication?

In this note, we will primarily work on the first idea of information theory and have a good look at the second question which is solved by *channel capacity*. A good treatment of the subject

can be found in [1–3]. Moreover, this subject has been influenced by statistical physics and has influenced the same. It is also used in engineering, we will see one example of such cases.

We write entropy as the measure of uncertainty a random variable x as (see Ch 10 in [3] for detailed discussion)

$$S(X) = \sum -p(x) \log p(x), \tag{1}$$

where $p(x)$ is the mass distribution of the variable. In quantum information theory (or quantum Shannon theory), we use discrete matrices in the place of mass distribution. In Eq.(1), we mostly use logarithms in base 2 and measure entropy in bits. We will discuss this fashion later.

The entropy in Eq.(1) should only become zero when $x \to 0$, as $0 \log 0 \to 0$. Furthermore, $S \geq 0$ as it is in thermodynamics. If one wishes to change the base, one can do

$$S_a(X) = \log_a(b) S_b(X) \tag{2}$$

and this can be easily proven using logarithmic identity.

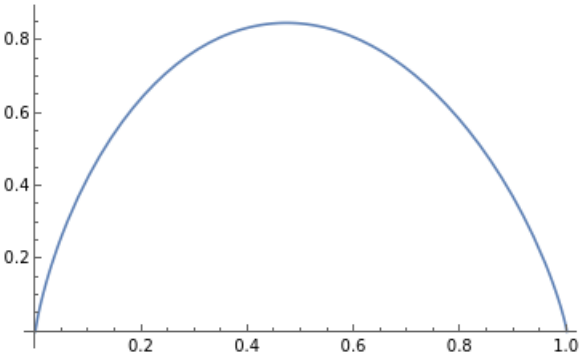

Figure 1: A binary entropy with two possibilities. One can observes that entropy is 1 bit when $p = 1/2$ and $(1 - p) = 1/2$. (We write entropy as $S = -p \log p - (1 - p) \log(1 - p)$)

Relative entropy: It is the measure of the distance between two distributions. (A good example is given in sec 2.3 in [4].) For mass distribution $p(x)$ and $g(x)$, the relative entropy is given by

$$S(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)} \tag{3}$$

We use relative entropy (or Kullback–Leibler divergence) when there are more than one distribution for the variable, in our case it is x. One can also check that $S(p||q) \neq S(q||p)$ and $S(p||q) \geq 0$ with equality only possible if $p = q$.

Mutual information: If we have two variables X and Y with a joint probability mass function $p(x, y)$ and marginal probability mass functions $p(x)$ and $p(y)$. Then the mutual information $(I(X; Y))$ is given by the relative entropy of the joint proability mass function and the product

distribution given by $p(x)p(y)$

$$I(X;Y) = \sum_X \sum_Y p(x,y) \log \frac{p(x,y)}{p(x)p(y)}. \tag{4}$$

An interesting relation between mutual information and entropy is given by

$$I(X;Y) = S(X) - S(X|Y) + S(Y), \tag{5}$$

where one can redefine using chain rule that $S(X|Y) = S(X,Y) - S(X)$, where $S(X|Y)$ is the conditional entropy and $S(X,Y)$ is the joint entropy (see [1]). It is interesting to note that conditioning an entropy reduces the entropy as

$$S(X|Y) \leq S(X). \tag{6}$$

Monotonicity of Relative Entropy: Suppose we have two variables X, Y and two probability distributions $P_{XY}(x,y)$ and $Q_{XY}(x,y)$. This gives us a relative entropy for the joint probability $S(P_{XY}||Q_{XY})$. But if we observe only one variable, say X, then the reduced probability is

$$P_X = \sum_y P_{XY}(x,y) \quad Q_X = \sum_y P_{XY}(x,y) \tag{7}$$

this gives us the confidence in observing X than the initial hypothesis is wrong by $S(P_X||Q_X)$. It is harder to disprove the initial hypothesis if we only observe X [4], so

$$S(P_{XY}||Q_{XY}) \geq S(P_X||Q_X) \tag{8}$$

this is called "monotonicity of relative entropy".

1.1 Shanon Entropy

We discuss Shanon entropy or the classical information theory. Suppose Alice has sent a message to Bob from the alphabets $a_1, \ldots, a_k$, where letter a_i is observed with probability p_i $(i = 1, \ldots, k)$ with the probability distribution P. The total number of messages comprised with the n alphabets, where a_i occurs Np_i times and $n \gg 1$, is given by

$$\frac{n!}{(np_1)! \ldots (np_k)!} \approx \frac{n^n}{\prod_{i=1}^k (np_i)^{np_i}} = 2^{nS_P} \tag{9}$$

where S_P is the Shanon entropy per letter

$$S_P = -\sum_i p_i \log p_i. \tag{10}$$

2 Quantum Information Theory

In Shanon entropy, we had a message of n letter from a list of alphabets and Shanon entropy was the number of bits in the message. If we generalize the situation into a quantum theory, how can we do it?

We can construct another set of alphabets, but this time from ensembles of states, and Alice can create a quantum message to send it to Bob and figure what is the 'Quantum Shanon' entropy of the message (see [2] for elementary discussions).

2.1 Density Matrices

Density matrices, or sometimes called density operators, are hermitian matrices with unit trace. We define density matrix for ψ_j with probability p_j

$$\rho = \sum_j p_j \left|\psi_j\right\rangle \left\langle\psi_j\right| \tag{11}$$

where p_j is non-negative and must add up to one and

$$tr(\rho) = 1. \tag{12}$$

Let us suppose that ψ_A belongs to Hilbert space $\mathcal{H}_A$ and ψ_B belongs to Hilbert space $\mathcal{H}_B$. The product Hilbert space (which is given by tensor product) is $\mathcal{H}_{AB} = \mathcal{H}_A \otimes \mathcal{H}_B$ and we can write ψ_{AB} as the combined vector in $\mathcal{H}_{AB}$ from ψ_A and ψ_B, so $\psi_{AB} = \psi_A \otimes \psi_B$. One interesting thing to realize is that an operator X_A if acts on ψ_{AB} can give information about ψ_A even if we forget ψ_B. Indeed, X_A has corresponding operator $X_A \otimes \mathbb{1}$ in ψ_{AB}.

We must point out that ψ_{AB}, if a generic pure state, is an entangled state in $\mathcal{H}_{AB}$ instead of a product state. (An interesting follow-up would be how to do purification such as ψ_{AB} with bell states.) We can write the pure state in terms of orthonormal vectors

$$\psi_{AB} = \sum_i \sqrt{p_i} v_A^i \otimes v_B^i. \tag{13}$$

We should discuss another concept called 'Fidelity' of quantum states. We define Fidelity $F(\rho, \sigma)$ as the closeness between two states so that one passes the test of another. We can write Fidelity in many ways, and one popular is by trace-norm

$$F(\rho, \sigma) = \left(tr|\sqrt{\rho}\sqrt{\sigma}|\right)^2 \tag{14}$$

and $1 \geq F \geq 0$. If ρ and σ are two pure states, the Fidelity is just inner product between states

$$F(\rho, \sigma) = |\left\langle\psi_\rho|\psi_\sigma\right\rangle|^2. \tag{15}$$

Fidelity is symmetric, i.e. $F(\rho,\sigma) = F(\sigma,\rho)$. Another measure for closeness of states is given by "Trace distance."

2.2 von Neumann Entropy

Once we know that density matrix is important ingredient which replaces the probabilistic mass functions in quantum information theory, we can write the entropy for such ensemble. The entropy we have is von Neumann entropy[1]

$$S(\rho) = -tr(\rho \log \rho). \tag{16}$$

We can write Eq.(16) in terms of Shanon Entropy if we choose a suitable basis

$$\rho \log \rho = \begin{pmatrix} p_1 \log p_1 & & & \\ & p_2 \log p_2 & & \\ & & p_3 \log p_3 & \\ & & & \ddots \end{pmatrix} \tag{17}$$

and with given $tr(\rho \log \rho)$

$$S(\rho) = -\sum_i p_i \log p_i. \tag{18}$$

This is why von Neumann entropy can be regarded as quantum Shanon entropy. Few immediate consequences from Eq.(18) about Eq.(16) are that $0 \log 0 = 0$ and $S(\rho) \geq 0$, as we have for Shanon entropy.

If a bi-partite system in $\mathcal{H}_{AB}$ has a pure state

$$\psi_{AB} = \sum_i \sqrt{p_i} v_A^i \otimes v_B^i, \tag{19}$$

then we can write for system A

$$\rho_A = \sum p_i \left|\psi_A\right\rangle \left\langle\psi_A\right|, \tag{20}$$

and for system B

$$\rho_B = \sum p_i \left|\psi_B\right\rangle \left\langle\psi_B\right|, \tag{21}$$

and so we can see that

$$S(\rho_A) = S(\rho_B). \tag{22}$$

This is an example of an entangled state, the equality of entropy states the entanglement of A and B in AB. For the pure bi-partite state we have a vanishing entropy $S(\rho_{AB}) = 0$.

[1] For explicit discussion on von Neumann entropy see chapter 11 in [3]

3 Quantum Conditional Entropy, Relative Entropy

We talked about conditional and relative entropy in classical case in section 1, we will now talk about quantum cases of conditional and relative entropy for von Neumann entropy. The relative entropy for two density matrices ρ and σ is given by

$$S(\rho||\sigma) = tr(\rho \log \rho) - tr(\sigma \log \sigma). \tag{23}$$

Just as classically, $S(\rho||\sigma) \geq 0$, with equality only possible if $\rho = \sigma$. This can be proved using 'Klein's inequality' [5, 6], which states that for A and B

$$tr(\log A - \log B) \geq tr(A - B) \tag{24}$$

with equality only possible if $A = B$. Quantum relative entropy is finite, it is only infinite when the kernel of σ has a non-trivial intersection with the support of ρ.

Conditional entropy in quantum cases can be written for a bi-partite system AB as

$$S(A|B) = S_{AB} - S_B, \tag{25}$$

where we must say that conditioning an entropy in a quantum sense is nothing like classical conditioning of an entropy. Another interesting fact to notice is that contrary to classical case, quantum conditional entropy can be negative. Since for a pure system the entropy is zero, so $S_{AB} = 0$, and for a not-pure state (or mixed state) the entropy is non-negative, so $S_B > 0$. Hence $S(A|B)$ can be negative.

Just as classically, we can write mutual information between A and B

$$I(A; B) = S_A - S_{AB} + S_B \tag{26}$$

which can be proven as

$$S_A - S_{AB} + S_B = S(\rho_{AB}||\sigma_{AB}) \tag{27}$$

where $\sigma_{AB} = \rho_A \otimes \rho_B$ and $\rho = \rho^{AB}$. And just as classically, fortunately, positivity of relative entropy implies the positivity of mutual information, which is sometimes called 'subadditvity of entropy'. If the mutual information is zero that means there are no correlations between A and B.

We can re-arrange Eq.(25) and write the reversed conditional entropy

$$S(B|A) = S_{AB} - S_A \tag{28}$$

which is negative as $S_{AB} = 0$ for a pure state. We can also make a statement that AB is a pure entangled state only when $S(B|A) < 0$.

4 Concavity of Entropy

5 Quantum Cloning and Teleportation

Quantum teleportation [7] can be achieved in quantum mechanics with some methods, which we will be discussing shortly. It turns out that conditional entropy is essential in knowing about teleportation, such that $S(A|B) \leq 0$ becomes a condition for teleportation.

Suppose we have an entangled state AB which Alice and Bob shares. Let us suppose now that Alice wants to share another qubit information, let us say C, to Bob. Bob takes out B and goes to his town. In this case, Alice has to create a state AC in a basis so no information is lost about C. It can be projected on any of the four bell states

$$\Psi_{AC} = \frac{1}{\sqrt{2}}(|00\rangle \pm |11\rangle)_{AC} \tag{29}$$

$$\Psi_{AC} = \frac{1}{\sqrt{2}}(|01\rangle \pm |10\rangle)_{AC}. \tag{30}$$

One thing Alice can do now is to measure ACB and know AC. Suppose AC is found to be

$$\Psi_{AC} = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)_{AC}, \tag{31}$$

then she can figure out B (from her view) to be

$$\Psi_{B} = \frac{1}{\sqrt{2}}(\alpha\,|0\rangle + \beta\,|1\rangle), \tag{32}$$

and now Alice can tell Bob about B and Bob can create ACB with the knowledge about B given by Alice and B he had. This is how entanglement will work.

One interesting thing about this teleportation is that $S(A|B) \leq 0$ is necessary. (In this sense, one can say that conditional entropy $S(A|B)$ is the amount of additional information that Bob having B needs to have from Alice to know about A.) To get an idea of how negative information works [8], remember that $S(AB) \leq S(B)$ for entangled states. We can presume that $S(B)$ is the information Bob has, and it is clearly more than $S(AB)$. So Bob knows more. If Alice sends information about A, which she does using conditional entropy and it can be negative, to Bob, then Bob ends up with $S(AB)$ which is lesser bits than what he knew before. This is negative information.

While we can copy unknown classical information, the simplest example is just to copy a digital file in a computer, we can not *clone* an unknown quantum information. This is simply called *no-cloning theorem*. In fact, the way we did the teleportation is the only accessible way to transfer information otherwise it is not possible to clone any quantum system as we can classically.

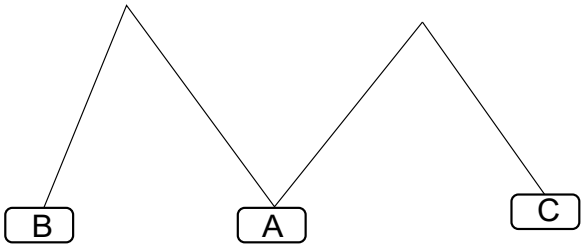

Figure 2: A somewhat rough diagram of how to achieve teleportation between Alice and Bob.

6 Holevo Information and Bound

Recall that the mutual information between Alice and Bob is given by $I(X;Y)$ and it is

$$I(A;B) = S_A - S_{AB} + S_B, \tag{33}$$

which was the maximum information attained by Bob (or accessible to Bob). Interestingly, there is a bound to how much Bob can know through mutual information, this bound is called Holevo bound [9].

Suppose now that Alice prepares a state from an ensemble $\varepsilon = \{p(x)\rho(x)\}$ and send it to Bob. The improvement in the information that is with Bob after he tries to know what Alice sent is given by mutual information. Bob would want to maximize this mutual information, as it would help in a clear understanding of Alice's state. We write the Holevo information χ as

$$\chi(\varepsilon) = S(\rho) - \sum_x p_x S_(\rho_x), \tag{34}$$

which is a non-negative quantity (because mutual information is non-negative). This information is the loss to Bob's knowledge about Alice's state after he knows about the specific ensemble that was used. Holevo information χ also acts as an upper bound to accessible information

$$\chi(\varepsilon) = S(\rho) - \sum_x p_x S_(\rho_x) \geq I(X;Y), \tag{35}$$

and the proof of Holevo found can be found in Chapter 12 in ref. [3]. Holevo information and Holevo bound is very practical in quantum computation and information theory. One can also argue that Alice can only send n qubits of classical information using n qubits using Holevo bound.

For a pure state χ is just the entropy of density operators while for mixed states χ is smaller. Another interesting property is the *monotonicity* of Holevo information. (Though we have not called out the monotonicity of mutual information in Sec. 1, but it does exist.) Monotonicity of Holevo information implies that a change of channel using a super-operator[2] φ

$$\varphi : \varepsilon = \{p(x)\rho(x)\} \to \varepsilon' = \{p(x)\ \varphi\rho(x)\} \tag{36}$$

[2]Super-operator does not have any connections with super-algebra or super-symmetry. It is widely used linear operator (or map) in quantum computing and information theory.

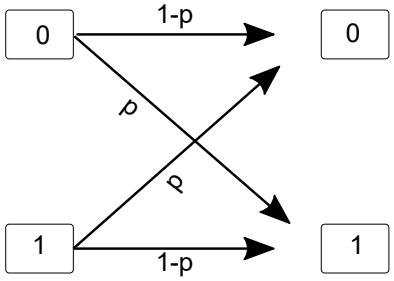

Figure 3: A binary discrete system with possibilities of error. 0 can end up to 1 with probability p and 1 can end up to 0 with probability $1-p$.

and

$$\chi(\varepsilon') \leq \chi(\varepsilon). \tag{37}$$

7 Noise Channels

Finally, we comment on how we use random codes to get error-free communication [10]. To get an idea of how these errors arrive, we can see it in a binary system, Fig. 3. We denote the possibility of errors with λ. Channel capacity $\mathcal{C}$ is defined as the maximum allowed mutual information

$$\mathcal{C} = \max\, I(X;Y) \tag{38}$$

and $\mathcal{C} \geq 0$ since $I(X;Y) \geq 0$. We also define rate R as the number of meaningful bits n' over number of bits sent n

$$R = \frac{n'}{n}, \quad n > n'. \tag{39}$$

Channel coding theorem (or Shanon theorem) states that all rates are achievable below the channel capacity with error probability $\lambda \to 0$. The converse of coding theorem implies that rates above channel capacity do not come with low error probability, see *sec 7.9* in ref. [1]. We can write that for low-error probability, and we have a condition to meet

$$R \leq \mathcal{C} = \max\, I(X;Y). \tag{40}$$

This rate is achievable using random codes, which was introduced by Shanon in [10]. Technically, suppose that we have a code-book of random codes, we then take the average of the probability of error over a random code from the code-book, which symmetrizes the structure of probability and which can be used thereafter.

Figure 4: A representation of transmission of information with least error. First, the message is encoded and then added to the channel, which must be known by both parties, then the encoded information is decoded by the receiver to gain the message.

7.1 Hamming Code

A very simple error-free coding would be repetition. For instance, if Alice wants to send a message 1, she would simply send 111, a repetition of 3, if Bob does not get 111 but, for example, 101, then an error has occurred. However, such simple code does not work for many problems. Hamming code was introduced for single-bit error correction [11]. Hamming codes work as parity-check codes.

We write numbers in binary form (base 2), 1 is 1, 2 is 10, 3 is 11, 4 is 100, 5 is 101 and so on. 'Parity bits' are those bits that are powers of 2 and contain only one 1 bit in the binary form, for example, $1, 2, 4, 8, 16$. Bits other than parity bits are called 'data bits'. There is two parity - even and odd. Suppose Alice wants to send 1101001; there are four 1-bit which is an even number. So Alice would attach, in this case, 0 at the end to make it 8-bits with parity making it to 11010010[3]. Then Alice computes the modular arithmetic (in mod 2), which, in this case, is 0. Then Alice transmits the 8-bits to Bob and Bob, after receiving it check the modular arithmetic (in mod 2) of 8 bits which is also 0. Bob confirms the parity of decoding to be even. So there was no error in the transmission. Transmission with odd parity is just the same, the reader should verify it.

The choice of parity is just a choice, but sender and receiver must know the chosen parity. Hamming code uses the parity-check concept in it. One might realize that parity-check has a critical failure in determining the error in even parity as if Alice sends 1010 and Bob receives 1100, then Bob is not aware of the error. The bits that were replaced are called corrupted bits. In Hamming code, one has a generator matrix G (for encoding) and a parity-check matrix H. A message by Alice is first encoded with the generator matrix, then a parity bit is added in the way we did, then it is transferred to Bob, and Bob does the parity-check. Another, a more intuitive and simple, way to see Hamming code is through the Venn diagram[4].

References

[1] T. Cover and J. Thomas, *Elements Of Information Theory*. John Wiley Sons, 2006.

[2] J. Preskill, "Lecture Notes on Quantum Information Theory,". `http://theory.caltech.edu/~preskill/ph219/index.html#lecture`.

[3] One can say that it is encoding. After it reaches to Bob, he would decode it

[4] See sec 7.11 in ref. [1]

[3] M. A. Nielsen and I. Chuang, *Quantum computation and quantum information*. Cambridge University Press, 2002.

[4] E. Witten, "A Mini-Introduction To Information Theory," *Riv. Nuovo Cim.* **43** no. 4, (2020) 187–227, arXiv:1805.11965 [hep-th].

[5] M. B. Ruskai, "Inequalities for quantum entropy: A review with conditions for equality," *Journal of Mathematical Physics* **43** no. 9, (Sep, 2002) 4358–4375, arXiv:0205064 [quant-ph].

[6] E. Carlen, "Trace inequalities and quantum entropy: an introductory course," *Entropy and the quantum* **529** (2010) 73–140. `http://www.mathphys.org/AZschool/material/AZ09-carlen.pdf`.

[7] C. H. Bennett, G. Brassard, C. Crépeau, R. Jozsa, A. Peres, and W. K. Wootters, "Teleporting an unknown quantum state via dual classical and Einstein-Podolsky-Rosen channels," *Physical review letters* **70** no. 13, (1993) 1895.

[8] M. Horodecki, J. Oppenheim, and A. Winter, "Quantum State Merging and Negative Information," *Commun. Math. Phys.* **269** no. 1, (2006) 107–136, arXiv:quant-ph/0512247.

[9] A. S. Holevo, "Bounds for the quantity of information transmitted by a quantum communication channel," *Problemy Peredachi Informatsii* **9** no. 3, (1973) 3–11.

[10] C. E. Shannon, "A mathematical theory of communication," *The Bell system technical journal* **27** no. 3, (1948) 379–423.

[11] R. W. Hamming, "Error detecting and error correcting codes," *The Bell system technical journal* **29** no. 2, (1950) 147–160.