1 Classical Entropy

Information theory is the study of storing, transmitting, processing, and general utilization of information. Key to this is a way to quantify what we mean by information in the first place. What does it mean to have a lot of information versus a little? This kind of question may seem obvious. We can simply look at how much memory on a computer something takes to store! However, even then, we can ask: What is the minimum amount of memory (bits) needed to effectively store some information? We can see that there are ways to reduce the amount of memory needed to store a picture or a song through compression. So then is there some intrinsic amount of information contained in, for example, a text document, or a short snippet of sound, or an image of a tree? The answer is yes, there is some intrinsic amount of information contained in each of these, which can be quantified through the Shannon entropy.

1.1 Shannon Entropy [1]

To begin, let’s set up some framework for how we think of information. Intuitively, information is only valuable because it could have been something else. For example, if I told you today was Tuesday, it is only useful to you because I could have told you that it was Wednesday or Friday. If for some reason, I could only ever tell you that it was Tuesday, it would be useless since I would say that regardless of what day it actually is. As such, we can define the information derived from an event as how “surprising” the event is. For example, if I were to drop a pencil and it fell to the ground, this would be completely unsurprising. This result should happen every time with 100% probability. Thus we don’t really learn anything and have gained 0 new information from it. However, if something extremely improbable happened, maybe the dropped pencil starts floating mid-air, we could stand to learn a lot from that phenomenon.

This “surprisingness” of an event can be quantified according to some function of the probability that it occurs called the Shannon self-information:

$$I(x) := -\log_2(p_x)$$  \hspace{1cm} (1)

With $I(x)$ the information from event $x$ and $p_x$ the probability that event $x$ occurs. We can see that this satisfies some of our previous intuitive ideas about information. An event $x$ with $p_x = 1$ has $I(x) = 0$ while an event $y$ with $p_y = 0$ has $I(y) = \infty$. 
This also satisfies an intuition that if we observe two separate, independent events, the total information should be the sum of their individual informations, as we can see:

\[ I(x, y) = - \log_2(p_{x,y}) = - \log_2(p_x p_y) = - \log_2(p_x) + - \log_2(p_y) = I(x) + I(y) \]  

(2)

And it is easy to see that for any \( p_x \in [0,1] \), \( I(x) \geq 0 \) so no event produces negative information, and that the information \( I \) increases monotonically with decreasing probability. We can now look at some “information source” that results in events \( \{i\} \) with probability \( \{p_i\} \), respectively, and define an average information produced by the system, called the Shannon entropy \( H \)

\[ H(A) = - \sum_i p_i \log_2 p_i \]  

(3)

1.2 Efficient Coding Example

To see the usefulness of the Shannon entropy, we can start by looking at a simple system: Let our “information source” be some device that returns outcomes \( \{1, 2, 3, 4\} \) with probabilities \( \{.5, .25, .2, .05\} \). Now let Alice be some person who wants to record the outcomes of this device over some long time, but can only do so in a string of binary bits (1’s and 0’s, no spaces). We can see that for this recording to be effective, every distinct sequence of outcomes must be distinguishable from others in the encoding, so for example, the following scheme:

1 → 0  
2 → 1  
3 → 01  
4 → 10

will not work since the sequence 1,2,3 → 0101 is identical to the sequence 3,3 → 0101 and the sequence 1,4,1 → 0101. We can see some simple way to encode each outcome as some 2 bit string:

1 → 00  
2 → 01  
3 → 10  
4 → 11

and Alice can simply read through the code according to the rule of interpreting the coded sequence 2 bits at a time. However, if we calculate the entropy \( H \) of this device, we see that the average information produced is actually

\[ H = -(0.5 \log_2 0.5 + 0.25 \log_2 0.25 + 0.2 \log_2 0.2 + 0.05 \log_2 0.05) \approx 1.68 \text{ bits of information} \]  

(4)
Indeed, we can see that the following encoding scheme allows Alice to record the outcomes with less than 2 bits on average:

1 → 1
2 → 01
3 → 001
4 → 000

To decode the recorded results, Alice can read either 3 0’s in a row, or stop at the first 1 she sees. If she sees 000, she knows the encoded result was a 4, otherwise, she can count how many 0’s came before the 1 in order to recover results 1, 2, 3. And if we look at, on average, how many bits Alice needs to encode a result her, we find that each outcome requires:

\[ E[\text{bits}] = 0.5 \times 1 + 0.25 \times 2 + 0.2 \times 3 + 0.05 \times 3 = 1.75 \text{ bits of information} \]  

which is less than the simple 2 bit code requires. Indeed, a proven result [1] shows that the entropy \( H \) defines the maximum amount of compression possible for a given information source and that for any information source, there exists some encoding scheme that achieves:

\[ E[\text{bits}] \leq H + 1 \]  

2 Connection to physical Entropy [2]

Those who have encountered statistical mechanics will recognize that the name and form of the Shannon entropy are very similar to the Gibbs entropy:

\[ S = -k_B \sum_i p_i \ln p_i \]  

with a difference of some scale factor. Indeed, the two are related, as should be expected. All information is in some way physical, and can be reproduced in some physical system. Besides that, both are measures of uncertainty in some system. \( S \) measuring the uncertainty in the microstate of some physical system, \( H \) being the uncertainty in the outcome of some information source. But any information itself is, in fact, physical in some way. A clear way to examine this is by looking at the Maxwell’s Demon thought experiment.

2.1 Maxwell’s Demon

Maxwell’s Demon describes a broad group of thought experiments concerning processes that seem to violate the second law of thermodynamics.
The original Maxwell’s Demon thought experiment concerns a chamber of gas separated by a partition in the middle. Somewhere on this partition, is a small demon who controls a trapdoor, which it can open and close with zero work. Suppose the two sides of the partition start at different temperatures. According to the second law of thermodynamics, there should not exist some way to move heat from the colder side to the warmer side without doing some work to the system as a whole. However, consider that temperature only describes a probability distribution of the velocities of each molecule in the gas. We imagine that the tiny demon in charge of our trapdoor is has such good eyesight that it can see both the position and velocity of each molecule of gas on both sides, and has the hand-eye coordination necessary to quickly open and close the trapdoor, so that it only allows the fast-moving molecules from the cold side to pass to the hot side, and only allows the slow-moving molecules from the hot side to pass to the cold side.

By doing this, eventually this process decreases the average velocity of molecules on the cold side and increases the average velocity of molecules on the hot side, decreasing entropy in the system as a whole. This would appear to violate the second law since we have apparently found a cyclic process that results in the decrease of entropy as a whole. However, the resolution of this problem lies in the fact that information entropy is also physical entropy. A simpler system can reveal more about the connection between physical and information entropy.

2.2 Szilard’s Engine [2]

An early analysis of Maxwell’s Demon was offered by Leo Szilard, who designed the following thought experiment. Imagine now, a chamber with precisely 1 molecule of gas inside connected to a heat bath at temperature $T$. At first, a tiny demon inserts a movable partition (without needing to do work) in the center of the chamber. Then it measures which side of the partition the molecule of gas is, and accordingly attaches a load to the partition. Then the demon lets the single molecule expand isothermally, pushing the load and doing work equal to $kT \int_{V/2}^{V} \frac{1}{V} dV = kT \ln 2$. 

$$\Delta S \geq 0 \quad (8)$$
Figure 1: (a) Single molecule in chamber of volume \( V \). (b) partition placed in middle of the chamber, demon observes which side of the partition the molecule is. (c) Accordingly, the demon attaches some opposing load to the partition. (d) The particle is allowed to isothermally expand, doing work on the attached load. Physical realisations of such an engine discussed in source [7]. Figure sourced from [2]

Then when the partition is removed, the engine has successfully extracted \( kT \ln 2 \) of useful work from the heat bath with no corresponding cold bath (as is needed in a Carnot engine) and has reduced the entropy of the heat bath by \( \Delta S = \frac{\Delta Q}{T} = k \ln 2 \) seemingly without a corresponding increase in entropy anywhere else, once again violating the second law of thermodynamics. In Szliard’s original proposal, he argued that the compensating entropy increase comes from the demon’s measurement of the position of the particle. The idea is that any measurement of the position of the single molecule must introduce at least \( k \ln 2 \) entropy to the system somehow. Here we can see a very clear connection between physical entropy and entropy inherent in information. The chamber itself, divided into two equal parts is analogous to the classical bit, and we can see that this \( k \ln 2 \) quantity must be somehow tied to the physical requirements of storing a classical bit.

2.3 Landauer’s Erasure Principle [2]

To further elaborate on the connection between physical systems and information entropy, we can consider the following. Information can be stored in logical states (strings of 1’s and 0’s), but these must be physical in some way. So for some physical system to effectively be able to store information from some information source, each outcome of the information source (logical state) must correspond to exactly one physical state of the physical system. Now consider operations on the information. If the operation is logically reversible (injective), then the memory must be preserved which means that
the physical realisation must also preserve entropy. However, a logically irreversible process (non-injective) must map multiple memory states to the same state, which also implies that multiple distinct physical states are changed into the same final physical state. However, this would mean a decrease in the entropy of the physical state as its degrees of freedom are reduced, implying there must be some entropy increase in the surrounding environment. We can think of erasing memory as a logically irreversible process as it maps any starting memory state to the same “standard state” (eg. a string of all 0’s). This irreversible process must dissipate some heat into the environment and thus increase entropy nearby. In fact, it can be shown that measurement is a logically reversible process as long as the measurement is stored in some initially blank state of memory. Thus the entropy that compensates the lost entropy in the Szilard’s engine is in found by considering the physical memory that the demon uses to record the result of measurement as well as the single molecule chamber. We can see that after extracting the $kT \ln 2$ of work from the chamber, the memory still retains the measurement of molecule position, meaning the complete system has not yet returned to its initial state. To complete the cycle for every part of the system, the memory needs to be erased, which requires $kT \ln 2$ of work, and involves a $k \ln 2$ increase in entropy that compensates the entropy decrease in the heat bath. This $kT \ln 2$ can be shown to be the minimum amount of energy needed to erase a single logical bit of information, known as Landauer’s erasure principle.

3 Von Neumann Entropy

Now that we have seen some basics of information theory and the close connection information has to the physical notion of entropy, we can take a look at some ideas about entropy in quantum information. First we would like to come up with a way to quantify entropy in a quantum system, especially in a way that captures the contribution of purely quantum effects like entanglement. To do this, we can look at the Von Neumann entropy. The Von Neumann entropy is an extension of the classical Gibbs entropy [3]. It is defined for some quantum state given by its density matrix $\rho$:

$$S(\rho) = -\text{Tr}(\rho \log_2 \rho)$$  \hspace{1cm} (9)

The derivation of this definition is lengthy and is not included [3], but it is clear to see that the Von Neumann entropy has very similar properties to the Shannon and classical Gibbs entropy. For example, it is also a non-negative quantity with equality with 0 only for pure states (states with no uncertainty). And for product states $\rho \otimes \sigma$, the Von Neumann entropy adds together as would be expected, with

$$S(\rho \otimes \sigma) = S(\rho) + S(\sigma)$$  \hspace{1cm} (10)

and for a diagonalized density matrix, $\rho_{\text{diag}} = \sum_i p_i |e_i\rangle \langle e_i|$ the Von Neumann entropy is equivalent to the classical form of Shannon entropy:

$$S(\rho_{\text{diag}}) = -\sum_i p_i \log_2 p_i$$  \hspace{1cm} (11)
3.1 Quantum Information in Entanglement

We can first begin to see the usefulness of the Von Neumann entropy when considering entanglement. In class, when we considered some bipartite system with subsystems $A$ and $B$, we saw that the total state $|\psi\rangle$ could be written according to its Schmidt decomposition:

$$|\psi\rangle = \sum_i \lambda_i^2 |i_A\rangle |i_B\rangle$$

(12)

In this case, we defined the Schmidt number, the number of coefficients $\lambda_i$ needed to specify state $i$, and used this as a rough measure of the entanglement of $|\psi\rangle$. If the Schmidt number is greater than 1, then the state is entangled, otherwise it is separable. However, the Schmidt number is not able to distinguish between different degrees of entanglement. Consider the following two states:

$$|\psi_a\rangle = \frac{1}{\sqrt{2}} (|00\rangle + |11\rangle), \quad |\psi_b\rangle = \frac{1}{2} |00\rangle + \frac{\sqrt{3}}{2} |11\rangle$$

(13)

Both have Schmidt number of 2, and it is clear that both are entangled states, but there is some difference between the states not captured by just the Schmidt coefficient. Instead, we can analyze the entanglement in these states through the Von Neumann entropy $S$.

First, we see that because both states are pure states in the product space, the entropy is 0:

$$S(|\psi_a\rangle \langle \psi_a|) = S(|\psi_b\rangle \langle \psi_b|) = 0$$

(14)

This makes sense, as we can simply measure each state in the appropriate basis and measure the same state with probability 1, so in information entropy terms, this should be a 0 entropy state. But instead of looking at the entire bipartite state, it will be useful to look at the entropy of each subsystem:

$$\rho_A = \text{Tr}_B(\rho_{\text{tot}})$$
$$\rho_B = \text{Tr}_A(\rho_{\text{tot}})$$

(15)

which gives the following for the subsystem density matrices:

(a): $\rho_A = \rho_B = \frac{1}{2} (|0\rangle \langle 0| + |1\rangle \langle 1|)$
(b): $\rho_A = \rho_B = \frac{1}{4} |0\rangle \langle 0| + \frac{3}{4} |1\rangle \langle 1|$

(16)

and where the Von Neumann entropy in each subsystem is some non-zero quantity:

(a): $S(\rho_A) = S(\rho_B) = 1$
(b): $S(\rho_A) = S(\rho_B) \approx .811$

(17)

Now we see some difference between the two states. The Bell state in (a) is called a *maximally entangled* state while the state in (b) is quantitatively less entangled in some...
way. For any bipartite state, the Von Neumann entropy of the reduced density matrices defines the entanglement entropy of the system [3,4,5]. In general, the entanglement entropy in each subsystem should be equivalent, as a consequence of Schmidt decomposition. This entanglement entropy can be thought of as the number of qubits entangled with the subsystem. This might be clearer to see in a more general example:

Consider a system with N total spins in some arbitrary state. Now divide the space into some qubits in A, and the rest in B. The entanglement entropy of A and B defines the minimum amount of qubits needed to be entangled with the subsystem to reproduce the information contained in the entire original system of N spins. Indeed, the Von Neumann entropy is such a close analogue to the Shannon entropy that it has been shown [6] that there also exist an efficient coding theorem for quantum communications where entanglement entropy defines the minimum amount of qubits needed to losslessly encode some quantum message.

### 3.2 Applications to Simulation of Quantum Systems [5]

The final topic I will touch on in this paper is a method for efficiently representing quantum systems on a classical computer. The key to this lies in how a large amount of the entropy in a quantum state happens to lie in the entanglement between qubits. Consider the following, the conventional difficulty in representing a quantum state $|\psi\rangle$ of N qubits classically lies in the $2^N$ complex coefficients needed to completely define the quantum state:

$$|\psi\rangle = \sum_{i_1,i_2,...,i_{N}} c_{i_1,i_2,...,i_{N}} |i_1,i_2,...,i_{N}\rangle$$  \hspace{1cm} (18)

However, consider the simple product state with no entanglement:

$$|\psi\rangle = (\alpha_1 |0\rangle + \beta_1 |1\rangle) \otimes (\alpha_2 |0\rangle + \beta_2 |1\rangle) \otimes ... \otimes (\alpha_N |0\rangle + \beta_N |1\rangle)$$  \hspace{1cm} (19)

This product state can be fully described with only $2N$ complex coefficients, far less than the exponential scaling described earlier. Clearly the large amount of extra information that needs to be encoded lies somewhere in the entanglement entropy. However, often times an actual physical quantum system of interest is not some maximally entangled system, in which case, it could be possible to study such systems through simulations on some classical computer. The process of a more efficient representation is detailed below.
3.2.1 Matrix Product Representation [5]

The primary goal of this more efficient representation is to find a way to retain the entanglements present in the system while minimizing resources spent other way. It is constructed iteratively: First take the Schmidt decomposition between the first qubit and the rest of the system. This decomposition will involve some kind of change of basis matrix $|\alpha_1\rangle = \Gamma_{\alpha_1}^{i_1} |i_1\rangle$ and will have some Schmidt coefficients $\lambda_{\alpha_1}^{[1]}$ which will also be kept. Between the two, the correlation between the subsystem and the total system are recorded. We continue on to do this for the first 2 qubits, then the first 3, and so on until we have done this for the first N-1 qubits vs the last qubit. The basis transformations $\Gamma$ and Schmidt coefficients $\lambda$ can then be put together in the following way to recreate the full state:

$$|\psi\rangle = \sum_{\alpha_1} \ldots \sum_{\alpha_{n-1}} \Gamma_{\alpha_1}^{i_1} \lambda_{\alpha_1}^{[1]} \ldots \lambda_{\alpha_{n-1}}^{[n-1]} \Gamma_{\alpha_{n-1}}^{i_n} |i_1, i_2, \ldots, i_n\rangle$$  

(20)

which represents the state $\psi$ according to a matrix product of the change basis matrices taken from the Schmidt decompositions. In general it can be shown that instead of the $2^N$ complex coefficients needed before, the information about the state can now be stored in $2N \chi^2$ [5] complex numbers located inside of the matrices. This $\chi$ quantity is now something that scales according to the entanglement entropy of the system. Intuitively, this can be tied back to the efficient encoding theorem for quantum systems. If the minimum number of qubits needed to reproduce some quantum system is related to the number of entangled qubits, then there’s no need to do what equates to simulating the entire N qubits. Instead, there should be some clever way to get a similar amount of information by simply simulating the “necessary” number of qubits, leading to a great reduction in computation expense. In the worst case of a maximally entangled system, this is not an efficient encoding and the number of coefficients returns to its exponential scaling. But looking at some physical systems, especially in their ground states, they are nowhere near maximally entangled. For example, a 1D chain of spins has an entanglement entropy that scales with the log of its length $l$, which leads to the possibility of efficiently simulating such systems on classical computers.
4 References

    https://arxiv.org/abs/0707.3400v2
    http://www.hartmanhep.net/topics2015/18-entanglement-intro.pdf
    https://pdfs.semanticscholar.org/83e1/658b22ee0fe0e605f2161987ce70f5d806ad.pdf
    a single electron”, https://www.pnas.org/content/111/38/13786