On the Thermodynamic Meaning of Negative Entropy

Anna Brandenberger[†]

School of Computer Science, McGill University

1. Introduction

Entropy has tens of interpretations throughout many fields such as information theory, probability and thermodynamics. It was introduced in the context of thermodynamics by Carnot and Clausius in thermodynamics, and given a statistical interpretation by Boltzmann and Maxwell. An information-theoretic operational understanding was then given by Claude Shannon in his famous 1948 treatise. Essentially, entropy was offered as the number of bits needed to encode a message when sending it through a noiseless channel. Of course however, information is transmitted through *physical* bits and a link between information theory and thermodynamics is needed. This is Landauer's principle, which states that the erasure of information inevitably requires work, i.e., results in the generation of heat.

In this paper, we examine these operational and physical interpretations of entropy in both the classical and the quantum settings. Some critical differences between classical and quantum entropy induce profound and very curious consequences both operationally and thermodynamically – namely, the possible negativity of conditional entropy in the quantum world.

We begin by introducing the classical view of information theory as formulated by Shannon in section 2.1. We define the classical Shannon entropy and interpret it operationally via Shannon's noiseless coding theorem. Landauer's principle in a classical system is then introduced in section 2.3 and we show how it is wielded to defeat the famous Maxwell's demon paradox.

Moving on to the quantum world, we first give a brief review of important notions of quantum mechanics, then introduce the von Neumann entropy and its operational meaning via the quantum noiseless encoding theorem in section 3.2 – a perfect quantum analog of the classical setting from section 2.2. Then, in section 3.3 we present the operational interpretation of conditional entropy by Horodecki et al. as the amount of qubits that an agent A needs to send to B for B to have the full description of the joint state. Here, we notice a quantum effect, the potential negativity of conditional entropy, and operational meaning is given to it.

Finally, in section 4, we investigate negative conditional entropy further through its thermodynamic interpretation. We examine an observer-dependent Landauer's principle, where the observer has a quantum memory potentially entangled with the system, by exploring the work by del Rio et al.. In this setting, they find that the negative entropy caused by the entanglement of the memory with the system allows work to be extracted from an erasure process!

2. Classical Information Encoded in Systems

2.1. Basics of Classical Information Theory

The information content of a system is, informally, the amount of information one party must transmit to another party – who only has some shared background knowledge such as the language used – for the second party to be able to reconstruct the state of the system. In classical information theory, this information is encoded in sequences of a base unit called the bit, which takes on two possible values, 1 or 0. Usually, the system does not only consist of binary parts; however, the first party can transmit the set of instructions on how to recreate the system, which may be very complicated, as a sequence of k binary choices. We then say that k bits of information are encoded in the system.

[†]anna.brandenberger@mail.mcgill.ca

We thus seek to estimate this amount of information k. Suppose we have a complicated classical system made of a large number N of components, each of which can be in one of n states with probability p_i, $i = 1, \ldots, n$. This is equivalent to a long string of characters, where each character is drawn from a language with n letters occuring with known probabilities. In his seminal 1948 work, Shannon develops a mathematical theory of communication and precisely answers our question of how to quantify the amount of information output by a discrete information source. Let X be a random variable with probability mass function p_i representing one character in our string.

The *information content* of each character can be interpreted as the level of surprise of receiving the outcome i. For a random variable X with distribution p_i, it is also known as *surprisal* or *self-information* [8], and defined to be the function

$$I_X(i) = \log_2 \frac{1}{p_i}. \tag{1}$$

Why this definition? We would like the surprisal of the intersection of two independent events X and Y to be equal to the sum of the individual surprises. The information content is a function of probabilities, and since the probability of the intersection of independent events is multiplicative $\mathbf{P}\{X \cap Y\} = \mathbf{P}\{X\}\mathbf{P}\{Y\}$, we want this function to turn multiplication in its input to sums in its output: $f(xy) = f(x) + f(y)$. The class of functions satisfying this is exactly the logarithm (any base)!

The *Shannon entropy* of X is then defined as

$$\begin{aligned} H(X) &\coloneqq \sum_{i=1}^{n} p_i \log_2 \frac{1}{p_i} \\ &= \mathbf{E}\{I_X(i)\}, \end{aligned} \tag{2}$$

the expected amount of information transmitted per character in a long string of characters.

Shannon proved that this definition of the entropy (2), modulo a constant multiple, is the only one that satisfies three properties we want. It is continuous in the p_i; is monotonically increasing in n in the case where all $p_i \equiv 1/n$; and satisfies additivity (if a choice is broken down into two choices, the original entropy should be the weighted sum of the individual values of entropy for each choice).

Indeed, the form of Shannon's entropy differs from the entropy formula derived by Boltzmann in the context of statistical physics only by a constant multiple [5]. Boltzmann's formula was found from the number of ways an observable macrostate of a thermodynamic system could be obtained from microstates. It is written

$$S = -k \ln 2 \sum_{i=1}^{n} p_i \log_2 p_i, \tag{3}$$

where k is Boltzmann's constant and p_i for $i = 1, \ldots, n$ are the probabilities of each of the n possible microstates of the system. We will discuss the extension of Boltzmann's entropy definition to quantum systems in Section 3.2.

Before moving on to give an operational understanding of entropy in the next section, we will first briefly recall a few important entropic notions that arise when we consider more than one system.

We've defined the entropy of a single random variable in (2). Now we may ask the question of what happens if we have two interdependent systems. Let's consider two random variables X and Y defined respectively on alphabets $\mathcal{X}$ and $\mathcal{Y}$ with joint distribution $p(x, y)$. Then the *joint entropy* of X and Y is simply an application of (2) to the vector-valued random variable (X, Y),

$$\begin{aligned} H(X, Y) &= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log_2 p(x, y) \\ &= -\mathbf{E}\{\log_2 p(X, Y)\}. \end{aligned} \tag{4}$$

Similarly, the *conditional entropy* of the random variable Y given X can be naturally defined as

$$\begin{aligned} H(Y|X) &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log_2 p(y|x) \\ &= -\mathbf{E}\{\log_2 p(Y|X)\}, \end{aligned} \tag{5}$$

and it quantifies the extra amount of information needed to describe the system Y given that the outcome of X is known. Indeed, noting that $H(X,Y)$ bits are required to describe the joint state of the two systems, this understanding of the conditional entropy as extra information is confirmed by the *chain rule*

$$H(X,Y) = H(X) + H(Y|X), \tag{6}$$

which can be shown from first principles. Thus, if the value of Y is completely determined once X is known, the conditional entropy $H(Y|X) = 0$. On the other hand, if X and Y are independent, then knowing X provides no information, and $H(Y|X) = H(Y)$. Another seemingly trivial property is the positivity of entropies, including conditional entropy:

$$H(Y|X) \geq 0. \tag{7}$$

Interestingly, conditional entropy is not necessarily non-negative in the quantum world! This reveals a fundamental difference between quantum and classical systems, and leads to many of the consequences we will discuss later in this paper.

2.2. Sending Information Through a Classical Channel

Now, we can also obtain an operational interpretation of the Shannon entropy, by linking it to fundamental limits on information transfer.

In the setting of a noiseless channel, Shannon showed that there exists an optimal encoding such that a long string of N random variables i.i.d. distributed as X can be compressed into $NH(X)$ bits, where $H(X)$ is the cutoff number of bits necessary to encode each random variable [21]. Any more and there is little risk of information loss, but any less and it is almost certain that information will be lost.

Let the *fidelity* F of an encoding denote the probability that the decoded message is the same as the message sent, i.e., the probability that there is no error in the encoding. The following statement of the theorem and proof are due to Schumacher [19].

Theorem 1 (Noiseless Coding Theorem)**.** *Consider a message source outputting i.i.d. random variables distributed as X, and let $\varepsilon, \delta > 0$.*

i) *If $H(X) + \delta$ bits are available per random variable X, then for sufficiently large N, a sequence of N random variables i.i.d. distributed as X can be encoded with fidelity greater than $1 - \varepsilon$.*

ii) *If $H(X) - \delta$ bits are available per random variable X, then for sufficiently large N, if a sequence of N random variables is encoded into a binary string, the fidelity will be less than ε.*

Proof. The proof uses the weak law of large numbers to relate the Shannon entropy to the number of "likely" sequences of N identical random variables. The weak law states that a long string of i.i.d. random variables has average very close to the mean of each variable with probability approaching 1. Consider a message $\alpha = x_1 x_2 \cdots x_N$ of length N where each $x_i \sim X$ and define the probability of this message to be $\mathbf{P}\{\alpha\} = p(X = x_1)\cdots p(X = x_N)$. The surprisal random variable $y_i = -\log p(X = x_i)$ as seen in (1) has expected value $\mathbf{E}\{y_i\}$ equal to the entropy $H(X)$. Also, its average over the long string can be written

$$\frac{1}{N}\sum_{i=1}^{N} y_i = -\frac{1}{N}\log \mathbf{P}\{\alpha\}.$$

The weak law of large numbers then tells us that most long messages have an average surprisal lying very close to the entropy. A set A of "likely" sequences can thus be defined for arbitrarily small δ and ε,

$$A \coloneqq \left\{\alpha : \left| -\frac{1}{N}\log \mathbf{P}\{\alpha\} - H(X)\right| \leq \delta\right\}, \tag{8}$$

such that the probability that a message lies in this set is greater than $1-\varepsilon$ for messages of a sufficiently large length N. From this condition on the set A, the probability $\mathbf{P}\{\alpha\}$ of each likely message $\alpha \in A$ can be bounded above and below. From there, so can the number ν of likely messages:

$$(1-\varepsilon)2^{N(H(X)-\delta)} \leq \nu \leq 2^{N(H(X)+\delta)}.$$

The bound (i) now directly follows, as we can code each of the likely messages into unique sequences of the $N(H(X)+\delta)$ available bits; other messages occur with probability less than ε and can all be mistakenly coded as one arbitrary fixed sequence.

The proof of (ii) to show the probability of error will be greater than $1-\varepsilon$ requires a bit more work. We are given $\varepsilon > 0$ and $\delta > 0$, and begin by picking a larger N than originally in (8) to have a larger set of likely sequences. We choose it such that (8) holds with ε replaced by a smaller $\varepsilon' = \frac{\varepsilon}{2} > 0$ and δ replaced by a smaller $\delta' = \frac{\delta}{2} > 0$. Now since we consider all possible encodings, we must think about mapping the $2^{N(H(X)-\delta)}$ distinct available binary sequences to a set of messages. The leftover messages then must induce coding errors. The number of correctly coded messages is definitely less than $2^{N(H(X)-\delta)}$ messages, from either the likely or unlikely sets. Using the bound

$$\mathbf{P}\{\alpha\} \leq 2^{-N(H(X)-\delta')}$$

for the probability of a likely message $\alpha \in A$, the probability $\mathbf{P}$ that a message is coded correctly satisfies

$$\mathbf{P} < \varepsilon' + 2^{N(H(X)-\delta)}2^{-N(H(X)-\delta')} = \tfrac{\varepsilon}{2} + 2^{-N\delta/2}.$$

This can be shown to be less than ε for a large enough N, completing the proof of (ii). ■

An alternate formulation of Theorem 1 exists when we focus on prefix codes. These codes require that there is no whole code word in the system that is a prefix of any other code word in the system; they are thus decodable uniquely in one pass, earning them the name of instantaneous codes. In this setting, where a random variable X is mapped via a prefix code to a binary string, Shannon showed that the expected length L of an optimal encoding of X satisfies [8]

$$H(X) \leq L \leq H(X) + 1. \tag{9}$$

This can be proven straightforwardly using Kraft's inequality and the non-negativity of relative entropy [8].

2.3. Landauer's Principle

In the previous section, we have talked about information rather theoretically, as a mathematical theory. However, any information is encoded and transmitted via physical devices and processes. For example, in a very simple model, one bit of information could be encoded in a spin-1/2 particle which is either in the $\uparrow$ or the $\downarrow$ state. It could also be encoded by a one-molecule gas in a box with a middle partition, where the molecule is either on the left or on the right. Thus, since the properties of basic components of information theory are controlled by physical laws, we would expect different information processing capacities when we consider the classical and quantum settings.

In 1961, Rolf Landauer proposed a principle fundamentally linking information theory to thermodynamics [13]. In essence, he posited that in order to do logically irreversible operations on a system, a minimum amount of energy must be expended per bit of information. This is in contrast to reversible operations, which can be performed without wasting any energy (in a quasistatic limit) [20]. It can be stated as follows.

> The erasure of one classical bit of information results in the generation of $kT\ln 2$ joules of heat released to the environment, where k is Boltzmann's constant and T is the temperature of the environment heat sink.

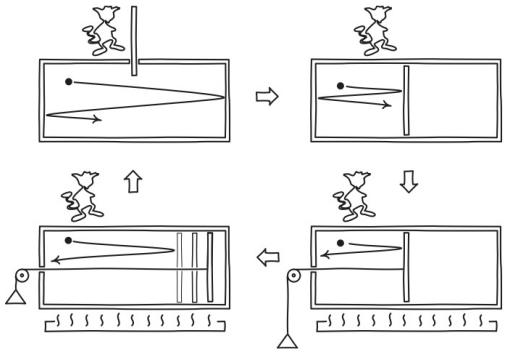

Figure 1. Szilard's engine (sourced [16]).

This erasure principle is widely accepted as a physical law, and was used by Charles Bennett to resolve the paradox of *Maxwell's demon* [1, 2, 3]. This paradox is a classical thought experiment designed by James C. Maxwell to violate the second law of thermodynamics [14]. In 1929, Leo Szilard proposed the Szilard engine, a refinement which shows thermodynamic consequences of having information about a system [22]. A diagram is shown in Figure 1.

In this model, the demon observes a box containing one molecule and extracts work by operating a piston attached to a load. A cycle proceeds as follows: first, the molecule is free to move in the box, and then a partition is added, splitting the box into two halves. The demon measures which side of the box the particle is in, records the result, then adds to the empty side a piston coupled to a load. Finally, the one-molecule gas is put in contact with a heat reservoir at temperature T and expands isothermally to fill the whole volume, doing work on the load. Specifically, recalling that the work done by the gas is $W = T\Delta S$, where $\Delta S = Nk\ln(V_f/V_i)$ is the change in entropy of an N-particle ideal gas expanding from volume V_i to V_f, an amount

$$W = T\Delta S = kT\ln 2 \tag{10}$$

of work is exerted on the load. The device is then back in its initial state! The paradox and violation of the second law of thermodynamics lie in the fact that this seems to be a cycle that completely converts heat into work, which is equivalent to decreasing the total entropy of this isolated system.

Before Bennett's treatment of the paradox, it was generally thought that the demon's measurement of the positions of the molecule must have required some work. Surprisingly, Bennett showed that some measurements actually can be made without energy expenditure [1]. However, he pointed out that, as written above, the process is not a complete cycle! The demon must be considered a part of the system. While the box is returned to its initial state, the demon is not – it has one bit of knowledge recorded, which it did not have at the start of the process. Thus, to return the demon's mind to its initial state, that bit of knowledge must be erased. By Landauer's principle, this erasure costs at least $kT\ln 2$ joules of energy, exactly counterbalancing the amount of work extracted from the gas expansion. Thus, as soon as we consider the information in the demon's mind as physical state, the second law of thermodynamics is saved: in the combined system of the box and the demon, no net work is done.

3. Information Encoded in Quantum Systems

We have discussed the idea that information should be viewed through a physical lens. It is therefore natural to examine how the differences between quantum and classical physics will affect both information theory and thermodynamics.

3.1. Important Notions of Quantum Mechanics

A few important notions from quantum mechanics will be relevant to our discussions. This section contains a brief summary of the definitions of pure, mixed, separable and entangled states, as well as some illustrative

examples [18]. A familiar reader may skim through them to pay more attention to section 3.2, where we introduce the quantum analog of Shannon entropy and its computational interpretation.

Consider two qubits A and B. The joint state of the two atoms belongs to a four-dimensional Hilbert space $\mathcal{H} = \mathcal{H}_A \otimes \mathcal{H}_B$, where $\mathcal{H}_A$ and $\mathcal{H}_B$ are two-dimensional Hilbert spaces for A and B respectively.

We begin with *pure states*, which are states for which we have complete knowledge of the preparation procedure. They can be represented in Dirac's ket notation as four-dimensional vectors $|\psi_{AB}\rangle \in \mathcal{H}$.

In contrast, some states result from a preparation procedure for which we have some classical uncertainty. For example, we could be producing states in a pure state $|\psi\rangle$ for p_1 of the time, and states in another pure state $|\phi\rangle$ for $p_2 = 1 - p_1$ of the time. Such states cannot be simply represented as four-dimensional vectors. Instead, we represent them as matrices belonging to the space $\mathcal{HM}$ of 4×4 Hermitian matrices with positive eigenvalues and normalized such that $\operatorname{Tr} \rho = 1$. $\mathcal{HM}$ is a subspace of the Hilbert-Schmidt space $\mathcal{HS} = \mathcal{H} \otimes \mathcal{H}^*$ of bounded linear operators on $\mathcal{H}$, where $\mathcal{H}^*$ is the dual space of $\mathcal{H}$. We call this the *density matrix formalism*. We give a motivation for this formalism by considering the expected value of some operator A. In our example above, we should add the classical uncertainty on top of the regular expected values of A over the states $|\phi\rangle$ and $|\psi\rangle$. We should therefore have

$$\langle A \rangle = p_1 \langle \psi | A | \psi \rangle + p_2 \langle \phi | A | \phi \rangle . \tag{11}$$

Adding to this expression the resolution of the identity $\mathbf{I} = \sum_n |n\rangle\langle n|$, where $\{|n\rangle\}$ is a complete set of basis states (here, it can for example be $\{|00\rangle, |01\rangle, |10\rangle, |11\rangle\}$), we can rewrite (11) and simplify to

$$\begin{aligned} \langle A \rangle &= \sum_n \langle n | A (p_1 |\psi\rangle\langle\psi| + p_2 |\phi\rangle\langle\phi|) | n \rangle \\ &= \operatorname{Tr} \rho A. \end{aligned}$$

We denoted the quantity in parentheses the *density matrix*

$$\rho = p_1 |\psi\rangle\langle\psi| + p_2 |\phi\rangle\langle\phi| \tag{12}$$

and noted that the trace of a matrix B can be written $\sum_n \langle n | B | n \rangle$. The density matrix ρ captures all the information that we have about the state, both quantum and classical. Note that a state is pure if and only if its density matrix satisfies $\rho = \rho^2$, thus $\operatorname{Tr} \rho^2 = 1$. Pure states can all be decomposed into a form $\rho = |\psi\rangle\langle\psi|$ for some ket $|\psi\rangle \in \mathcal{H}$; this is illustrated in Table 1.

We can characterize composite systems - joint states of several particles - according to another important (and independent) axis. Going back to pure states to describe this phenomenon, suppose A is in state $|1\rangle_A$ and B in state $|0\rangle_B$. We can write the joint state as the tensor product $|10\rangle = \binom{0}{1} \otimes \binom{1}{0}$. The set of such states $|00\rangle, |01\rangle, |10\rangle$ and $|11\rangle$ forms the canonical basis of $\mathcal{H}$; any general state can be written a superposition of these four states. Note however that not all states in $\mathcal{H}$ can be factorized into a tensor product as we have done for $|10\rangle$! For example, we cannot separate the description of particle A from that of B when they are in the joint state $\frac{1}{\sqrt{2}}(|01\rangle + |10\rangle)$ (see ρ_3 in Table 1). We denote *product*, or *separable states* all joint states that can be factored into tensor products of a state in $\mathcal{H}_A$ and a state in $\mathcal{H}_B$, and denote *entangled states* those that are not separable. These entangled particles cannot be described separately: a measurement of one particle will affect the whole system, leading to many of the counter-intuitive predictions quantum mechanics is famous for.

Finally, suppose we know that particles A and B are in a joint pure state $\rho = |\psi\rangle\langle\psi|$, but we only have access to particle A. The state of our particle A can then be described by the *reduced density matrix* on subsystem A

$$\rho_A \coloneqq \sum_i \langle i |_B (|\psi\rangle\langle\psi|) |i\rangle_B = \operatorname{Tr}_B \rho, \tag{13}$$

the partial trace of ρ over the basis $\{|i\rangle_B\}$ of $\mathcal{H}_B$.

We can now characterize the entanglement of a composite system simply from the reduced density matrix on one of its subsystems. Joint separable states must be written in a $|\psi\rangle\langle\psi|_A \otimes |\phi\rangle\langle\phi|_B$ form and thus have

	Pure	Mixed
Separable	$\rho_1 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$ $= \lvert 0\rangle\langle 0\rvert \otimes \lvert 1\rangle\langle 1\rvert$ $= \lvert 01\rangle\langle 01\rvert$	$\rho_2 = \frac{1}{2}\begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$ $= \frac{1}{2}\big(\lvert 0\rangle\langle 0\rvert \otimes \lvert 1\rangle\langle 1\rvert + \lvert 1\rangle\langle 1\rvert \otimes \lvert 0\rangle\langle 0\rvert\big)$
Entangled	$\rho_3 = \frac{1}{2}\begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$ $= \frac{1}{\sqrt{2}}\big(\lvert 01\rangle + \lvert 10\rangle\big)$ $\frac{1}{\sqrt{2}}\big(\langle 01\rvert + \langle 10\rvert\big)$	$\rho_4 = \frac{1}{4}\begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}$ $= \frac{1}{2}\frac{1}{\sqrt{2}}\big(\lvert 00\rangle + \lvert 11\rangle\big)$ $\frac{1}{\sqrt{2}}\big(\langle 00\rvert + \langle 11\rvert\big) + \frac{1}{2}\rho_3$

Table 1. Illustration of the pure-mixed and separable-entangled axes, with examples of the various types of states. The pure-mixed axis is determined by whether a density matrix can be written as $\rho = \lvert\psi\rangle\langle\psi\rvert$ for some column vector $\lvert\psi\rangle \in \mathcal{H}$; while the separable-entangled axis is determined by whether ρ can be written as a sum of tensor products of the density matrices for the individual particles.

reduced density matrix (on A) $\lvert\psi\rangle\langle\psi\rvert_A$, a pure state in the smaller space of 2×2 density matrices for particle A. Conversely then, a joint pure state is entangled if and only if its reduced state over a subsystem is mixed. For example, the entangled state $\lvert\psi\rangle = \frac{1}{\sqrt{2}}(\lvert 0\rangle_A \lvert 1\rangle_B - \lvert 1\rangle_A \lvert 0\rangle_B)$ has reduced density matrix on A

$$\rho_A = \tfrac{1}{2}(\lvert 0\rangle\langle 0\rvert_A + \lvert 1\rangle\langle 1\rvert_A), \tag{14}$$

a mixed state on $\mathcal{H}_A$. So knowing the reduced state of one of the particles in a joint system gives us information about the entanglement of the system as a whole.

3.2. Quantum Entropy and Qubit Encoding

In the preceding subsection, we've examined mixed states arising from two situations: from an experiment with some classical uncertainty, or when we can only observe a subsystem of an entangled state. Such a system is in one of a set of pure states $\{\lvert\psi_i\rangle\}$ with classical probabilities $\{p_i\}$, similarly to a classical system that can be in a set of states with given probabilities. A measure of uncertainty analogous to the entropy (3) in the classical setting can be defined here [18].

We let the *von Neumann entropy* of a system with density matrix ρ be

$$S(\rho) = -\operatorname{Tr}\{\rho \log \rho\}, \tag{15}$$

which can be simplified by writing ρ in its eigenbasis $\{\lvert i\rangle\}$ as $\rho = \sum_i a_i \lvert i\rangle\langle i\rvert$:

$$S(\rho) = -\sum_i a_i \log a_i, \tag{16}$$

a formula that very closely resembles (2) and (3). Note that considering a different basis is perfectly reasonable, as $S(\rho)$ is invariant under unitary transformations of ρ. This must be the case, as physically relevant properties of the system are basis independent.

Let's quickly examine the two limiting cases. For a pure state, $\rho_P = \rho_P^2$ and thus the entropy $S(\rho_P)$ vanishes, as we expect. On the other hand, consider the maximally mixed state $\rho_M = \sum_i \frac{1}{N} \lvert i\rangle\langle i\rvert$ where

$\{|i\rangle\}$ is an orthogonal basis of size N. This state corresponds to the outcome of an experiment generating N possible pure states with equal probability, and has maximal entropy $S(\rho_M) = \log N$.

A final interesting property can be observed when we examine the entropy of an entangled state $|\psi\rangle\langle\psi|_{AB}$. Recall from the discussion at the end of the previous section that the entanglement of a joint system manifests itself in the individual particles having mixed reduced density matrices. These reduced states will then have non-zero entropy. Indeed, returning to the example in (14), the entropy of the entangled (pure) state itself is zero, but its reduced density matrix over A is maximally mixed and has maximal entropy, signaling the existence of entanglement:

$$S(\rho_A) = -\operatorname{Tr}\{\rho_A \log \rho_A\} = \log 2.$$

Thus, in general, a joint system of N particles can be partitioned into two subsystems A and B containing respectively N_A and $N_B = N - N_A$ particles. A notion of *bipartite entanglement entropy* can then be defined to quantify how entangled the system is. For a joint pure state $\rho_{AB} = |\psi\rangle\langle\psi|_{AB}$, it is given by

$$S(\rho_A) = S(\rho_B), \tag{17}$$

where as usual ρ_A and ρ_B are the reduced density matrices over particles A and B. The equality of the entropy of the two subsystems is shown in Appendix A.

Before talking about information erasure in this quantum setting in section 3.4, let us first present Schumacher's quantum analog of Theorem 1. This quantum coding theorem allows us to interpret the quantum entropy of some system as the amount of *quantum* resources required to represent the information about a system. Just as was the case with Shannon's theorem, this operational understanding provides an justification for the definition of the entropy.

Consider a quantum message source M which represents each message i from a message source A as a pure "signal" state $|\psi_i\rangle_M$, occurring with probability p_i. The state of a message thus has density matrix

$$\rho = \sum_i p_i \, |\psi_i\rangle\langle\psi_i|_M \, ,$$

and we can denote the density matrix of each signal state by $\rho_i = |\psi_i\rangle\langle\psi_i|_M$.

The Shannon entropy of the classical message source A is only equal to the von Neumann entropy of ρ if the signal states are all orthogonal! In other cases,

$$S(\rho) < H(A) = -\sum_i p_i \log p_i.$$

There is a discrepancy between the amount of classical information used to create the mixed state ρ, and the accessible information left in ρ.

Rather than relate $S(\rho)$ back to classical information of the message preparation as above, Schumacher interpreted the term in a fundamentally quantum way, in terms of the number of qubits required to encode a large group of N signals from M with high fidelity [19]. A qubit is simply a spin-1/2 system, the quantum analog of a bit.

Theorem 2 (Quantum Noiseless Encoding Theorem)**.** *Let M be a quantum signal source with signal ensemble described by the density matrix ρ, and let $\delta, \varepsilon > 0$.*

i) *If $S(\rho) + \delta$ qubits are available per signal, then for sufficiently large N, a group of N signals can be encoded in the available qubits with fidelity greater than $1 - \varepsilon$.*

ii) *If $S(\rho) - \delta$ qubits are available per signal, then for sufficiently large N, if a group of N signals is encoded in the available qubits, the fidelity will be less than ε.*

This looks incredibly similar to Theorem 1! Before moving on to the proof, the notions of a group of signals and fidelity must be clarified in the quantum context. We must also elaborate on the notion of a limited quantum channel.

Let's first quickly formalize the notion of a group of N signals, as it can be done quite straightforwardly. Consider an extended quantum source M^N consisting of N i.i.d. copies of M. It has signal ensemble

$$\rho^N \coloneqq \rho_1 \otimes \cdots \otimes \rho_N, \tag{18}$$

where each ρ_i is i.i.d. distributed as ρ. The eigenstates of ρ^N will be product states $|n_1, \cdots, n_N\rangle$, where $|n_i\rangle$, $n = 1, \ldots, \dim(\mathcal{H}_M)$ are the eigenstates of the ρ_i.

Now on to channels. Let X denote the quantum channel, to which we must transfer the signal from the source system M. Note that in the quantum setting, copying and transposing a state are two different notions. Copying a pure state has the joint system of M and X evolving as

$$|\psi_i\rangle_M |0\rangle_X \to |\psi_i\rangle_M |\psi\rangle_X ,$$

where $|0\rangle$ is a fixed "null" state. The no-cloning theorem, as shown by Wooters and Zureck [25] among others, states that some signal states cannot be copied.

We therefore cannot copy the signal state from M into the quantum channel X, and must instead use *transposition* to move the state into X. Unlike during cloning, transposition erases the signal from M:

$$|\psi_i\rangle_M |0\rangle_X \to |0\rangle_M |\psi_i\rangle_X .$$

It is a unitary operation U provided that the inner products between signal states are preserved in the coded states in the quantum channel. This is a perfect transposition and imposes the condition that $\dim(\mathcal{H}_X) \geq \dim(\mathcal{H}_M)$ (supposing w.l.o.g. that the signal states span $\mathcal{H}_M$). The channel encoding and decoding can thus be represented as

$$M \xrightarrow{U} X \xrightarrow{U^{-1}} M'$$

where M' is a copy of the message space M and U^{-1} represents a perfect decoding.

We instead want to consider *approximate transpositions*, where the dimension of the quantum channel may be *less* than that of the message space. For such a transposition, let X be composed of a channel subsystem C, and an extra system E which is discarded. The approximate transposition from M to M' through the limited channel C can be represented

$$M \xrightarrow{U} C + E \to C \to C + E' \xrightarrow{U'} M'$$

where U' is some decoding and E' is a copy of the extra system E in a fixed state (such as $|0\rangle\langle 0|_{E'}$).

Now that we have an adequate formulation of a quantum channel where information is potentially lost, we want a measure of how close an output signal[1] ω_i is to its input signal $\rho_i = |\psi_i\rangle\langle\psi_i|$. This can be quantified by a "validation measurement" which indicates whether w_i and ρ_i match. It will succeed (return a match) with probability $\operatorname{Tr} w_i\rho_i$. The *fidelity* F for a general signal ensemble ρ can thus be written as

$$F \coloneqq \sum_i p_i \operatorname{Tr} w_i\rho_i, \tag{19}$$

where w_i can be derived from the diagram above in terms of the unitary operations U and U', a partial trace over E and the fixed state (e.g. $|0\rangle\langle 0|_{E'}$) added right before decoding.

From this definition, we expect that if the channel C is too small, the fidelity should be near zero. Conversely, if C is large enough, we should be able to make the fidelity approach 1. Schumacher formalizes these two intuitions in two lemmas that are then used in the main proof.

Lemma 3. *Let the dimension of the limited channel be $d = \dim(\mathcal{H}_C)$. Suppose that the signal state ρ satisfies that for any projection Γ onto a d-dimensional subspace of $\mathcal{H}_M$,*

$$\operatorname{Tr} \rho\Gamma < \eta \tag{20}$$

for a fixed η. Then the fidelity is bounded above as $F < \eta$.

[1] Note that since we have discarded a subsystem of X, the output signal ω_i may no longer be a pure state.

This result follows relatively simply by noting that each w_i in (19) is supported on a d-dimensional subspace of $\mathcal{H}_M$; therefore, its eigenvalues are projectors satisfying (20). In addition, it can be shown that no generality in Lemma 3 is lost if we suppose that Γ is a projection onto a subspace of $\mathcal{H}_M$ spanned by d eigenstates of ρ. Indeed, $\operatorname{Tr}\rho\Gamma < \eta$ if and only if the sum of any d eigenvalues of ρ is less than η. The proof of the next lemma uses this remark. It asserts the existence of a scheme with fidelity close to 1 given a sufficiently large channel.

Lemma 4. *Let the dimension of the channel be $d = \dim(\mathcal{H}_C)$. Suppose there exists a projection Γ onto a d-dimensional subspace of $\mathcal{H}_M$ such that*

$$\operatorname{Tr}\rho\Gamma > 1 - \eta \tag{21}$$

for a fixed η. Then there exists a transposition scheme with fidelity $F > 1 - \eta$.

The proof is by explicitly constructing a transposition scheme.

Armed with these lemmas, the quantum noiseless coding theorem can be proved!

Proof sketch of Theorem 2. We first note that the von Neumann entropy of ρ is equal to the Shannon entropy of the distribution of eigenvalues P_n of ρ, $n = 1, \ldots, \dim(\mathcal{H}_M)$. Similarly, the signal ensemble ρ^N has eigenvectors equal to the product states $|n_1, \ldots, n_N\rangle$ of eigenvectors of ρ, and the corresponding eigenvalues are also products

$$P_{n_1,\ldots,n_N} = P_{n_1} \cdots P_{n_N}.$$

Thus, when only considering eigenvalues, this quantum signal is analogous to a classical message of length N that has alphabet $\{1, \ldots, \dim \mathcal{H}_M)\}$ and probability distribution P_n, with Shannon entropy equal to $S(\rho)$. From the proof of Theorem 1, we can consider two orthogonal "likely" and "unlikely" subspaces of $\mathcal{H}_{M^N}$. The "likely" subspace Λ is spanned by the eigenvectors whose classical analog belong to the set Λ in (8).

Thus, for the positive bound (i), just as in Shannon's original theorem, we can faithfully transpose the signals in the likely set Λ to the available qubit sequences. Indeed, from the classical (8), we can pick N large enough such that the probability of unlikely eigenvectors is made less than $\varepsilon/2$ while Λ has less than $2^{N(S(\rho)+\delta)}$ eigenvectors. Then any projection Γ into a d-dimensional subspace of Λ satisfies (21), and the result follows from Lemma 4.

Similarly, the impossibility bound (ii) also follows from the classical analog which tells us that for large enough N, no sum of the $2^{N(S(\rho)-\delta)}$ eigenvectors will have sum greater than ε. Applying Lemma 3 then yields the result.

There is no more need to worry about the technicalities of the new definition of fidelity – they were dealt with in the two lemmas! ■

This theorem is a perfect analog of Shannon's noiseless encoding theorem in the setting of a quantum channel! It provides an interpretation for the von Neumann entropy, and implies that we can naturally measure quantum information in number of qubits.

3.3. Conditional Quantum Entropy and Transferring Partial Information

We have discussed computational interpretations of the Shannon entropy $S(\rho)$ of a state ρ. Another related quantity that also has a very interesting computational meaning is the conditional entropy.

Recall the conditional Shannon entropy $H(X|Y)$ defined in (5) which satisfied the chain rule (6). It can be understood as the amount of classical information that one would need to learn the state of X when knowing some background information Y (which is potentially correlated with X).

In a quantum setting, the above scenario can be represented by a two-party game as well: Alice and Bob each have a state in some unknown joint quantum state ρ_{AB}, and respectively have reduced density operators ρ_A and ρ_B. The analogous question is then how much additional quantum information Alice needs to send Bob such that Bob has the full ρ_{AB}. Horodecki et al. investigate the minimum amount of quantum communication to do this, in units of qubits, when allowing unlimited classical communication [11]. They find that this is exactly the quantum conditional entropy

$$S(A|B) \coloneqq S(AB) - S(B), \tag{22}$$

where $S(AB)$ and $S(B)$ are the von Neumann entropies of ρ_{AB} and ρ_B respectively.

The curious thing about this conditional quantum entropy is that unlike in the classical setting, $S(A|B)$ can be negative! Already Schrödinger had noted that entangled states can possess a strange feature where we can know more about a whole system than about subsystems. For example, consider again the example we used in sections 3.1 and 3.2, the entangled state $|\psi\rangle = \frac{1}{\sqrt{2}}(|0\rangle_A |1\rangle_B - |1\rangle_A |0\rangle_B)$. Its reduced density matrix on subsystem A (14) is maximally mixed , leading to a negative conditional entropy

$$S(B|A) = S(AB) - S(A) = -\log 2.$$

Indeed, Cerf and Adami showed that the conditional entropy is non-negative for separable states [6]. Conversely, negative conditional von Neumann entropy $S(A|B)$ implies entanglement of the bipartite joint system [6, 7]!

The operational understanding of conditional entropy presented by Horodecki et al. also gives an interpretation of negative conditional entropy. In this two-party game, if the conditional entropy is negative, rather than needing to communicate a number of bits, Alice and Bob can actually gain $-S(A|B)$ maximally entangled states, which can be used for future communication.

Instead of elaborating more on this operational meaning of conditional entropy, we will move to its thermodynamic interpretation. We will find that negative conditional entropy also has an effect on Landauer's principle, and through it, on physical quantities such as work.

3.4. Erasure of a Quantum System

We have discussed the operational interpretations of both the von Neumann entropy and the conditional entropy – noting the curious fact that in the quantum world, conditional entropy can be negative. In the following section, we will move to discuss Landauer's principle in a fully quantum setting. However, before doing so, we must answer the question of how to perform erasure of classical information encoded in a quantum state.

For a classical state represented by a molecule in a box with a partition, we could erase the information in the system by using a piston to push the molecule to the left half of the box, irrespective of where it started. When carried out reversibly and isothermally, this procedure generates $kT \ln 2$ of heat, by the same analysis as done to obtain (10).

For a quantum system, we define *erasure* of a system as setting it to a predetermined state, e.g. the pure state $|0\rangle$. Consider a system in a mixed state $\rho = \sum_i p_i |e_i\rangle\langle e_i|$, where $|e_i\rangle$ are the eigenvectors of ρ.

One method of erasure could be to measure the system in the eigenbasis; this measurement gives outcome i with probability p_i. The system will then collapse to an eigenstate $|e_i\rangle$, which we can rotate unitarily to a fixed standard state $|e_0\rangle$. We must then erase the classical record of the measurement outcome, expending $kT \ln 2$ per bit according to Landauer's principle. In total, an amount

$$kT \ln 2 H(\{p_i\}) = kT \ln 2 S(\rho) \tag{23}$$

of heat is generated. If we had measured in any other basis, the Shannon entropy of the classical record would have been larger than $S(\rho)$. The right hand side of (23) is therefore a lower bound for the energy required for erasure using this procedure, in accordance with Landauer's principle.

A more elegant erasure procedure is by *thermal randomization*, introduced by Elihu Lubkin [15, 17]. With this method, the quantum system simply needs to be put into contact with a heat bath at temperature T. When thermal equilibrium is reached, the system will be in a state

$$\omega = \frac{e^{-\beta H}}{Z} = \frac{1}{Z} \sum_i e^{-\beta E_i} |e_i\rangle\langle e_i| , \tag{24}$$

where $\beta^{-1} = kT$, H is the Hamiltonian of the system with eigenstates $|e_i\rangle$ and eigenvalues E_i, and Z is the partition function $Z = \mathrm{Tr}\{e^{-\beta H}\}$. From (24), we can note that the probability $p(|e_i\rangle)$ that w is in the a

pure state $|e_i\rangle$ is exponentially decreasing as the energy E_i increases: the probability follows a Boltzmann distribution

$$p(|e_i\rangle) = \frac{e^{-\beta E_i}}{Z}. \tag{25}$$

Thus, we can make w arbitrarily close to the fixed pure ground state $|e_0\rangle$ by letting the system have an energy spectrum with a large level spacing: $E_1 \gg E_0$. This was exactly the goal of the erasure process!

We can compute the net amount of entropy generated by this thermal randomization erasure process, by separately finding the change in entropy of both the heat bath and the system [15]. It is found to be

$$\begin{aligned} \Delta S_{net} &= -k \operatorname{Tr}\{\rho \log \omega\} \\ &\geq kS(\rho), \end{aligned} \tag{26}$$

confirming Landauer's principle. We can notice that the erasure uses the least amount of energy when the equilibrium state of the bath ω is the same as the system ρ that we are trying to erase.

Qubits are a two-level system where we can w.l.o.g. take the ground state energy of $|0\rangle$ to be 0 and the excited energy of $|1\rangle$ to be E_1. In this setting, the thermal randomization procedure then consists in coupling our system to a heat bath at temperature T, and then slowly raising the energy of $|1\rangle$. In this case we can derive (26) more explicitly. Indeed, the probability $p(|1\rangle)$ that the excited state is occupied is

$$p(|1\rangle) = \frac{e^{-\beta E_1}}{1 + e^{-\beta E_1}}.$$

Then, the first law of thermodynamics gives us that the total work done on the system W is equal to the energy cost ΔE. Altering the energy of a state $|\psi\rangle$ from E_0 to $E_0 + dE$ has an average energy cost of $p(|\psi\rangle)dE$, where $p(|\psi\rangle)$ is the probability that the system is in that state. Then, when the excited energy E_1 tends to infinity, we have in the quasistatic limit,

$$\begin{aligned} W = \Delta E &= \int_0^\infty p(|1\rangle)dE \\ &= kT \ln 2, \end{aligned} \tag{27}$$

exactly in accordance with Landauer's principle!

This erasure procedure will be a key part of the paper by del Rio et al. that we will discuss in the following section [9]. Their work considers Landauer's principle where the observer has a quantum memory, giving a thermodynamic meaning to negative entropy.

4. Landauer's Principle and Negative Quantum Entropy

In their paper "The thermodynamic meaning of negative entropy", del Rio et al. investigate what happens to Landauer's principle if the observer uses a quantum rather than a classical memory. That is, if rather than using classical bits to store their information about the system, the observer were to use a set of qubits entangled with the system as memory.

Unlike in a classical setting, in the case of a quantum system, different observers may have different knowledge about the system. So in fact, Landauer's principle should be rewritten in terms of the minimum cost of erasure of a system X for an observer O denoted $W(X|O)$: the amount of work that O needs to do erase X. For an observer with a classical memory, we can denote the observer by O_C. Landauer's principle can thus be written as

$$W(X|O_C) = S(X|O_C)kT \ln 2, \tag{28}$$

where $S(\cdot)$ denotes the von Neumann entropy[2], k is Boltzmann's constant and T the temperature of the environment as before. To tie this into the classical thermodynamics view of the entropy as a property of

[2]The authors actually note that most of the statements are valid for any reasonable entropy definition. In the proofs, they use smooth min- and max- entropies, which reduce to the von Neumann entropy for suitable distributions.

the system, we can consider a *standard observer* who only has access to some macroscopic parameters such as the energy of the system and is maximally ignorant otherwise.

The authors del Rio et al. consider the fully quantum setting in which both the system and the memory consist of qubits. Now there is a slight caveat in that accessing the memory O could change its contents – this would be a problem if we changed parts of the memory containing information about other systems, since it could cause problems for computation. Thus, the authors proceed as in most quantum information theory papers and impose a condition on the erasure process where the information about other systems is preserved. Letting the system of interest be X and the other systems mentioned in the memory be a reference system R, this amounts to requiring that the joint system ρ_{OR} be preserved.

To begin the discussion, del Rio et al. consider an n-qubit system X. They introduce three agents, Alice, Bob and Quasimodo, the first two having classical memories and the third having a quantum memory Q maximally entangled with the system X. Suppose that Alice prepares the system in some pure state (her classical memory A must therefore be large enough to store a full description). Then, all three observers have different conditional entropies:

i) Alice knows the state of the system, thus has $S(X|A) = 0$;

ii) Bob considers each of the 2^n pure states equally likely, and thus has maximal entropy $S(X|B) = n$;

iii) Quasimodo has $S(X|Q) = S(XQ) - S(Q)$, where ρ_{XQ} is pure and ρ_Q is fully mixed. Then $S(XQ) = 0$ and $S(Q) = n$, and the entropy is negative: $S(X|Q) = -n$.

The work cost for each of these observers to erase their system – as we defined previously, to take each particle to a fixed state – will be different. Let us consider an $n = 1$ qubit system for simplicity.

i) Alice knows the state, and can simply apply a unitary operation to rotate it to $|0\rangle$, a reversible operation which can be done at no cost:
$$W(X|A) = 0.$$

ii) Bob has no information about the state, and considers it the fully mixed state $\rho = \frac{1}{2}(|0\rangle\langle 0| + |1\rangle\langle 1|)$. He can apply a thermal randomization erasure procedure as described in section 3.4 to end up with $|0\rangle$, The total cost of this operation was derived in(27) to be
$$W(X|B) = kT \ln 2.$$

iii) Quasimodo can be modelled as having a two-qubit memory $Q = Q_1 \otimes Q_2$. The first qubit is maximally entangled with the system X, in a state $|Q_1 X\rangle$ and the second is maximally entangled with the reference system R, in a state $|Q_2 R\rangle$. Quasimodo must preserve the reduced state
$$\rho_{QR} = \mathrm{Tr}_X(\rho_{QXR}) = \tfrac{1}{2}\mathbb{1}_{Q_1} \otimes |Q_2 R\rangle\langle Q_2 R|\,, \tag{29}$$
where $\mathbb{1}_{Q_1}$ is the identity matrix of the dimension of Q_1: $\mathbb{1}_{Q_1} = |0\rangle\langle 0| + |1\rangle\langle 1|$. Quasimodo will actually be able to *extract* work from this system. We describe the scheme below.

We can notice that Quasimodo's situation is essentially opposite to Bob's: the two-qubit joint system $|Q_1 X\rangle$ is pure rather than fully mixed. So the strategy for extracting work is to run Bob's thermal randomization erasure process but backwards.

Procedure 1 (Work extraction)**.** For an ℓ-qubit system in a pure state, del Rio et al. define the following procedure that extracts exactly $\ell kT \ln 2$ work.

a) initially, the pure state is at some energy E_0, and only one energy level is occupied. We denote this state as $|\psi_0\rangle$, and let the system have a basis $\{|\psi_i\rangle\}$, $i = 0, 1, \ldots, 2^\ell - 1$. The energy of the other (unoccupied) levels $i = 1, \ldots, 2^\ell - 1$ can be raised to a high value E_1 at no cost.

b) Then, the system is coupled to a heat bath and the energy of the empty states is slowly decreased. Just as during the thermal randomization erasure, the energy levels are populated according to a Boltzmann distribution, and the probability that the system is in any of the excited states is given by

$$p(|\psi_1\rangle) = \frac{e^{-\beta E_1}}{(2^\ell - 1)e^{-\beta E_1} + e^{-\beta E_0}},$$

where $\beta^{-1} = kT$. The total probability that the system is in an excited state is then $(2^\ell - 1)$ times this value. Lowering their energy E_1 down to zero results in an energy *gain* equal that can be computed just as in (27). In the quasistatic limit as $E_1 \to \infty$, an amount

$$\int_0^\infty \frac{1}{1 + e^{\beta(E_1 - E_0)}/(2^\ell - 1)} = \frac{\ln(2^\ell)}{\beta} = \ell kT \ln 2$$

of energy is gained, which can be stored in a battery.

c) The final state of the ℓ-qubit system is fully mixed, just like the starting state of Bob's erasure process. Indeed, notice that the thermal randomization procedure that Bob uses is simply this work extraction process run backwards).

When Quasimodo applies this work extraction procedure to his two-qubit pure state $|Q_1 X\rangle$, he gains $2kT \ln 2$ of work, and ends up with a fully mixed separable state $\rho_{Q_1 X} = \frac{1}{4}\mathbb{1}_{Q_1 X}$. This implies that both the reduced states of his memory ρ_{Q_1} and of the system ρ_X are fully mixed. The fact that $\rho_{Q_1} = \frac{1}{2}\mathbb{1}_{Q_1}$ is indeed necessary for the joint state of the memory and the reference (29) to be preserved, as we required. Finally, the fully mixed ρ_{Q_1} can be erased using Bob's procedure at a cost of $kT \ln 2$. The total work cost of Quasimodo's erasure is then negative:

$$W(X|Q) = -kT \ln 2,$$

a work gain! The final state of the full system is

$$\rho_{QXR} = |0\rangle\langle 0|_X \otimes \tfrac{1}{2}\mathbb{1}_{Q_1} \otimes |Q_2 R\rangle\langle Q_2 R|\,,$$

and all entanglement is lost.

In their work "The Thermodynamic Meaning of Negative Entropy", del Rio et al. extend this example to settings where the memory is not fully entangled with the system – work can be extracted from whatever entanglement there is. Their main theorem is proven in the setting of a single erasure (single-shot). The statement is probabilistic, since in general, the work required to erase a system is a random variable.

Theorem 5 (Main result)**.** *There exists a process to erase a system X conditioned on a memory O and acting at temperature T, whose work cost satisfies*

$$W(X|O) \leq [H_{\max}^{\varepsilon}(X|O) + \Delta] kT \ln 2 \tag{30}$$

except with probability less than $\delta = \sqrt{2^{-\Delta/2} + 12\varepsilon}$ for all $\delta, \varepsilon > 0$.

The quantity $H_{\max}^{\varepsilon}(\cdot)$ denotes the ε-smooth max-entropy, a single-shot generalization of the von Neumann entropy defined as

$$H_{\max}^{\varepsilon}(X|O) \coloneqq \inf_{\rho'_{XO}} \sup_{\sigma_O} \log_2 F(\rho'_{XO}, \mathbb{1}_{\mathbb{1}_X \otimes \sigma_O}),$$

where F denotes the fidelity, the supremum ranges over all density matrices ρ_O on O, and the infimum is taken over all density matrices ρ'_{XO} that are ε-close to ρ_{XO} in the purified distance. This notion can be operationally understood as the maximum fidelity of ρ_{XO} with a product state that is completely mixed on X [12]. It is often used in information theory, where it and some dual definitions are used to characterize information processing tasks. It satisfies several important properties, like the limits for pure and fully mixed states as $\varepsilon \to 0$. The technical details of this definition will not be elaborated on here. We simply note that

in the thermodynamic limit, where we increase the size of the system to "average away" fluctuations, this ε-smooth max entropy converges to the von Neumann entropy $S(\cdot)$. Indeed, we can define the *work cost rate* $\overline{w}$ of an erasure process as

$$\overline{w}(X|O) = \lim_{N\to\infty} \tfrac{1}{N} W(X^{\otimes N}|O^{\otimes N}), \tag{31}$$

the average work cost of the process in the limit where we take N i.i.d. copies of the system X and the memory O. Using a similar limiting argument as (8) in the proof of Shannon's noiseless encoding theorem, it can be shown that this entropy converges to the von Neumann entropy [9]. We then have:

Corollary 6 (Thermodynamic limit of main result)**.** *There exists a process to erase a system X conditioned on a memory O and acting at temperature T, whose work cost rate satisfies*

$$\overline{w}(X|O) \leq S(X|O)kT\ln 2. \tag{32}$$

Theorem 5 implies that if we have a quantum memory entangled with X, i.e., with conditional entropy $S(X|O) < 0$, we can erase the system with *negative* work cost, actually gaining work by doing so! This works even if the memory qubits are not maximally entangled with the system like Quasimodo's were in our simple example.

Another corollary of the theorem is thus that, for the same n-qubit system, an amount greater or equal to

$$[n - H^{\varepsilon}_{\max}(X|O) - \Delta]kT\ln 2 \tag{33}$$

of work can be extracted via some process (of course, except with probability less than $\delta = \sqrt{2^{-\Delta/2} + 12\varepsilon}$ for all $\delta, \varepsilon > 0$).

We will now present a proof sketch of the main theorem. In addition to building intuition, the example that we presented with Alice, Bob and Quasimodo already gave us a few important tools such as Procedure 1.

Proof sketch of Theorem 5. The result is proven by giving an explicit process satisfying the bound (30). X is assumed to be an n-qubit system. The erasure parallels Quasimodo's erasure procedure illustrated earlier, but here there are more details to deal with since we do not know how the memory is entangled with the system. It proceeds in three main steps:

i) The system X is manipulated to compress the correlations between the memory and X into a pure state of a subsystem of $X \otimes O$ that resembles Quasimodo's starting state. The subsystem has approximately $n - H^{\varepsilon}_{\max}(X|O)$ qubits, and the joint pure state is maximally entangled between two subsystems of $X \otimes O$.

ii) Using that joint pure state, we can extract roughly $[n - H^{\varepsilon}_{\max}(X|O)]kT\ln 2$ of work and leaving the state now fully mixed. *Note that work extraction can end at this step, explaining the extra $nkT\ln 2$ of work in* (33) *relative to the bound in* (30)*.*

iii) Finally, we erase the now separable system X, performing work $nkT\ln 2$.

Out of these three steps, we already know how to deal with (ii) and (iii). The work extraction from a joint pure state can be done using the work extraction procedure 1, with number of qubits $\ell = n - H^{\varepsilon}_{\max}(X|O)$, completing (ii). The erasure has also been described previously as the thermal randomization procedure, which is simply Procedure 1 run backwards.

Before proceeding to (i), we first define a *purifying system* Γ, such that the joint state $\rho_{XO\Gamma}$ is pure (recall how mixed states arise as subsystems of larger entangled systems as discussed at the end of section 3.1).

For the information compression step (i), del Rio et al. prove that it is possible to create via a local unitary transformation on X an ℓ-qubit state of a subsystem of $X \otimes O$ with

$$\ell \geq n - H^{\varepsilon}_{\max}(X|O) + 2\log\big(\delta^2 - 12\varepsilon\big),$$

that is δ-close to a pure state. The distance between states ρ, σ is measured by the trace distance $\frac{1}{2}\|\rho - \sigma\|_1$. Their proof proceeds in two steps, using two lemmas which we will state and justify below.

a) First, a subsystem $X_1 \subseteq X$ of $\ell/2$ qubits can be (up to a probability determined by δ) decoupled from Γ, in a fully mixed state.

   A notion of a system A being δ'-*decoupled* from another system B is introduced. This occurs if their joint state ρ_{AB} is δ'-close to a product state $\frac{\mathbb{1}_A}{|A|} \otimes \rho_B$. This statement is then proven using decoupling results by F. Dupuis and others [10], who give a bound on the average over all unitary operators on the system X of the distance between the desired product state and a state obtained after applying that unitary.

b) Then, since the global state is pure, we must be able to find a subsystem P of $X \otimes O$ of the same dimension $\ell/2$ that purifies the state of X_1. The joint pure state $X_1 \otimes P$ has ℓ qubits and is $\sqrt{2\delta}$-close to a fully entangled state.

   Supposing that X_1 and Γ are fully decoupled, this is quite straightforward, as we can find systems A_1 and A_2 that purify ρ_{X_1} and ρ_Γ. Then $A_1 \otimes A_2$ purifies $\rho_{X_1} \otimes \rho_\Gamma$. The claim for $\delta > 0$ then follows from Uhlmann's theorem [23] regarding the fidelity between two states, and some computations.

This completes the proof! Essentially, the correlations between the system and the memory hinted at by the negative conditional entropy $S(X|O)$ were compressed into a fully entangled subsystem. We could extract work from this entanglement via Procedure 1. The resulting fully mixed state of the system was then thermally erased. ■

Note that for an observer with a classical memory, the conditional entropy $S(X|O)$ will never be negative, so the amount of work required to erase a system is always greater or equal to zero. It can only be less than Landauer's classical limit if the observer has some classically encoded information about the system, like Alice did in our earlier discussion. The classical formulation of Landauer's principle implies "the work cost of erasing an unknown bit". It thus assumes a classical observer who does not have any prior knowledge about the system, i.e., whose joint state ρ_{XO} is fully mixed.

It is important to note that all of the original entanglement is lost at the end of the work extraction process. The process can therefore not be repeated, preventing Maxwell's demon from being able to take advantage of the scheme to produce a perpetual motion machine.

5. Conclusion

We started this work by introducing entropy in both the classical and quantum settings. We interpreted the notion operationally as the amount of bits (or qubits) needed to encode a random variable with no information loss. A connection between information theory and physics was then presented in the form of Landauer's principle, which postulates that information erasure implies heat generation. We then noted that a key quantum effect, quantum entanglement, is linked to the possible negativity of conditional von Neumann entropies. Conditional entropy was discussed first from an operational perspective, then in the context of erasure of a quantum system. The paper by del Rio et al. showed an interesting "violation" of Landauer's principle if the observer has a quantum memory entangled with the system. In such a setting, they found that work could actually be extracted through erasure!

Going forward, it would be valuable to further learn about the operational interpretations of other quantum informational quantities, in the style of Schumacher's work [19]. For example, the notion of channel capacity arises when one considers the quantum analog of the noisy channel theorem.

On the side of ties between information theory and physics, many other insightful papers have been written regarding these topics. In the paper by del Rio et al., information theoretical tools were used to obtain a physical result. The opposite direction is also possible: thermodynamical statements can be translated to information theory. For instance, Landauer's principle can be applied to understanding Shannon's noisy channel theorem [17]. Vladko Vedral also used it in his paper "Landauer's erasure, error correction and entanglement" [24] to analyse both classical and quantum error correction from a thermodynamic point of view.

Finally, to gain further intuition about quantum negative entropy, we could investigate the paper by M. Berta, M. Christandl, R. Colbeck, J. M. Renes and R. Renner, "The uncertainty principle in the presence of quantum memory" [4]. The authors show a violation of Heisenberg's uncertainty principle if quantum information about a system is available, and note that this violation is also linked to the negativity of conditional entropy.

References

[1] Charles H. Bennett. The thermodynamics of computation—a review. *International Journal of Theoretical Physics*, 21(12):905–940, 1982.

[2] Charles H. Bennett. Notes on the history of reversible computation. *ibm Journal of Research and Development*, 32(1):16–23, 1988.

[3] Charles H. Bennett. Notes on Landauer's principle, reversible computation, and Maxwell's Demon. *Studies In History and Philosophy of Science Part B: Studies In History and Philosophy of Modern Physics*, 34(3):501–510, 2003.

[4] Mario Berta, Matthias Christandl, Roger Colbeck, Joseph M. Renes, and Renato Renner. The uncertainty principle in the presence of quantum memory. *Nature Physics*, 6(9):659–662, Jul 2010. ISSN 1745-2481. doi: 10.1038/nphys1734.

[5] Ludwig Boltzmann. *Über die mechanische Bedeutung des zweiten Hauptsatzes der Wärmetheorie (vorgelegt in der Sitzung am 8. Februar 1866).* Staatsdruckerei, 1866.

[6] Nicholas J. Cerf and Christoph Adami. Negative entropy and information in quantum mechanics. *Physical Review Letters*, 79(26):5194–5197, Dec 1997. ISSN 1079-7114. doi: 10.1103/physrevlett.79.5194.

[7] Nicholas J. Cerf, Christoph Adami, and Robert M. Gingrich. Quantum conditional operator and a criterion for separability. *arXiv preprint quant-ph/9710001*, 1997.

[8] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory.* Wiley-Interscience, USA, 1991. ISBN 0471062596.

[9] Lídia del Rio, Johan Åberg, Renato Renner, Oscar Dahlsten, and Vlatko Vedral. The thermodynamic meaning of negative entropy. *Nature*, 474(7349):61–63, Jun 2011. ISSN 1476-4687. doi: 10.1038/nature10123.

[10] Frédéric Dupuis. The decoupling approach to quantum information theory, 2010.

[11] Michał Horodecki, Jonathan Oppenheim, and Andreas Winter. Partial quantum information. *Nature*, 436(7051):673–676, Aug 2005. ISSN 1476-4687. doi: 10.1038/nature03909.

[12] Robert Konig, Renato Renner, and Christian Schaffner. The operational meaning of min- and max-entropy. *IEEE Transactions on Information Theory*, 55(9):4337–4347, Sep 2009. ISSN 0018-9448. doi: 10.1109/tit.2009.2025545. URL `http://dx.doi.org/10.1109/TIT.2009.2025545`.

[13] Rolf Landauer. Irreversibility and heat generation in the computing process. *IBM journal of research and development*, 5(3):183–191, 1961.

[14] Harvey S. Leff and Andrew F. Rex. *Maxwell's demon: entropy, information, computing.* Princeton University Press, 2014.

[15] Elihu Lubkin. Keeping the entropy of measurement: Szilard revisited. *International journal of theoretical physics*, 26(6):523–535, 1987.

[16] John D. Nordon. `https://www.pitt.edu/~jdnorton/Goodies/Idealization/index_old.html`, 2010.

[17] Martin B. Plenio and Vincenzo Vitelli. The physics of forgetting: Landauer's erasure principle and information theory. *Contemporary Physics*, 42(1):25–60, 2001.

[18] John Preskill. Lecture notes for Physics 229: Quantum information and computation. *California Institute of Technology*, 16, 1998.

[19] Benjamin Schumacher. Quantum coding. *Phys. Rev. A*, 51:2738–2747, Apr 1995. doi: 10.1103/PhysRevA.51.2738.

[20] Francis W. Sears. *Thermodynamics, kinetic theory, and statistical thermodynamics*. Addison-Wesley, 1975.

[21] Claude E. Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27 (3):379–423, 1948.

[22] Leo Szilard. Über die Entropieverminderung in einem thermodynamischen System bei Eingriffen intelligenter Wesen. *Zeitschrift für Physik*, 53(11-12):840–856, 1929.

[23] Armin Uhlmann. The "transition probability" in the state space of A*-algebra. *Reports on Mathematical Physics*, 9(2):273–279, 1976.

[24] Vlatko Vedral. Landauer's erasure, error correction and entanglement. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 456(1996):969–984, Apr 2000. ISSN 1471-2946. doi: 10.1098/rspa.2000.0545.

[25] William K. Wootters and Wojciech H. Zurek. A single quantum cannot be cloned. *Nature*, 299(5886): 802–803, 1982.

A. Proofs

Lemma 7 (Entanglement entropy)**.** *Partition a joint system of N particles can be partitioned into two subsystems A and B containing respectively N_A and $N_B = N - N_A$ particles. For a joint pure state $\rho_{AB} = |\psi\rangle\langle\psi|_{AB}$, the entropies of the two subsystems ρ_A and ρ_B are equal:*

$$S(\rho_A) = S(\rho_B). \tag{34}$$

Proof. Note that the joint pure state $|\psi\rangle_{AB}$ belongs to the Hilbert space $\mathcal{H}_{AB} = \mathcal{H}_A \otimes \mathcal{H}_B$. Suppose w.l.o.g. that $N_A \leq N_B$. Our result follows simply from its Schmidt decomposition:

$$|\psi\rangle_{AB} = \sum_{i=1}^{N_A} \lambda_i \, |a_i\rangle_A \otimes |b_i\rangle_B \,,$$

where λ_i are non-negative real numbers, and $|a_i\rangle_A$ and $|b_i\rangle_B$ are orthonormal states in A and B respectively. Then, the partial traces over both subsystems A and B can be taken to obtain

$$\rho_A = \sum_{i=1}^{N_A} \lambda_i^2 \, |a_i\rangle\langle a_i|_A$$
$$\rho_B = \sum_{i=1}^{N_A} \lambda_i^2 \, |b_i\rangle\langle b_i|_B \,,$$

which clearly have the same entropy. ■