Master thesis

Landauer's principle – Its classical conception and extension to Quantum Mechanics

Eric Davidsson

August 31, 2018

Abstract

As a foundation we discuss some profound questions, such as: *what is entropy?*, and *how to understand probabilities in physics?* We then look at how *Information Theory* can motivate results in *Thermodynamics*—when considering the *principle of maximum entropy inference*. The purpose is to support a good understanding of *Landauer's principle*—in its inceptive motivation in classical physics, and why it is still a controversial idea that authors continue to disagree about. To remedy the ambiguity, we pursue a universal argument in favour of the principle. The work to extend Landauer's principle to *Quantum Mechanics* is then commenced, and we examine a situation where the principle delivers a seemingly anomalous prediction—before identifying what went wrong. Lastly, we pull at loose threads that will require further work to tie together and treat ourselves with speculations about the *measurement problem*.

Stockholm University, AlbaNova University Center,
Department of Physics, Sweden.

Thesis supervisor:
Prof. Gunnar Björk, Quantum Electronics and Quantum Optics,
Royal Institute of Technology (KTH).

Assisting supervisors:
Ass. prof. Supriya Krishnamurthy, Condensed Matter and Quantum Optics,
Stockholm University.
Ass. prof. Jonas Larson, Condensed Matter and Quantum Optics,
Stockholm University.

Contents

1 Introduction

The primary motivation behind this thesis is to address two problems.

Problem I. As far back as 1961, Rolf Landauer published a seminal paper [1] that until today has continued to generate disagreement and controversy. Landauer's argument concerns physical information processing devices (such as computers etc.). His claim is that there is a link between the logical operations performed on the information and the underlying physics of the system that encodes the information. Specifically, Landauer says that "*logical irreversibility*[1] is associated with *physical irreversibility* and requires a minimal heat generation". Unfortunately he does not provide a general derivation of what is now known as *Landauer's principle*—instead, he argues for his conclusion by demonstrating its validity in certain physical systems. This approach is cultivated to this day, as authors keep publishing papers with conflicting accounts. In this thesis (predominantly in section 5) we attempt to provide an argument in favour of Landauer's ideas, based only on *general physical principles*. The ambition is that such arguments can catalyze consensus and reduce the confusion surrounding *Landauer's principle*. □

Problem II. The question of how to assign ontological interpretations to some results predicted by Quantum Mechanics (and observed in experiments) is a very active ongoing debate. Therefore, it was very exciting to read a paper authored by Adán Cabello, Mile Gu, Otfried Gühne, Jan-Åke Larsson, and Karoline Wiesner in 2016, where they claim to demonstrate experimentally testable differences between two classes of interpretations for Quantum Mechanics [2]. In fact, their claim was ever stronger; one class (to which many cherished interpretations belong) is shown to produce unphysical predictions, casting considerable doubt on their feasibility. The approach taken by Cabello et al.—to produce experimentally falsifiable predictions—is the arguably the most important cornerstone for the whole of Physics, and such claims deserve serious attention from the community. But "extraordinary claims require extraordinary evidence", and in particular, the theoretical argument for such claims need to be carefully scrutinized. Since all interpretations of Quantum Mechanics is *initially* imagined to account for the same mathematical framework, we have good reasons to suspect that any difference between the interpretations should be minimal, contrary to the claim of Cabello et al. In this thesis (section 7) we conduct a careful examination of the argument and identify a crucial problem, which unfortunately invalidates their conclusion. □

Seemingly, these two problems have little to do with each other, but it was in fact the efforts to examine the second problem that required an appreciation of the first. Cabello et al. employs *Landauer's principle* as a step in their argument, and therefore, an understanding of the first problem is of great benefit to the reader wanting to recognize the objections put forth to the second problem.

Nevertheless, the reader is not in any way discouraged to jump ahead to whichever section appears most interesting. Care has been taken to properly

[1] Simply stated, a process is considered *logically irreversible* if we cannot uniquely determine some previous state from the current state.

refer to relevant sections whenever claims are made that is backed up elsewhere, and in that spirit, we will conclude this introduction with a bird's eye view of the content in the main sections.

Also, before we begin, a short remark about *Maxwell's demon* is in order. Even though the most common application of Landauer's principle is to give a detailed account of why any scheme such as Maxwell's demon is impossible, we will not enter into that discussion in this thesis. The literature on this subject is dense, and curious readers will no doubt be able to find appropriate resources. A good place to start can be Bennett's paper from 1987 [3].

1.1 A quick guide to the sections

Sections 2, 3 and 4. Relevant theoretical and conceptual background for the main body of work. □

Section 5. Addresses problem I, with an argument in favour of Landauer's principle based on classical physics. □

Section 6. An extension of Landauer's ideas to a fully quantum-mechanical framework. Not directly connected to problems I and II, but defines a possible path forward for extending Landauer's principle to Quantum Mechanics. □

Section 7. Addresses problem II, with an argument contrary to the conclusion of Cabello et al. □

Section 8. Defines further work, and argues briefly for a third class of interpretations of Quantum Mechanics, not considered by Cabello et al. □

2 Some quantum theory

In order to model statistical distributions, or classically probabilistic notions, in Quantum Mechanics—as we will do in sections 3, 4, and 6—we employ the formalism of the density operator. We will here provide a few necessary definitions and some brief intuitive discussions, all presented as a compact review. Readers familiar with these concepts may prefer to skip ahead to section 3.

2.1 The density operator

In Quantum Mechanics a physical system is associated with a *complex Hilbert space* $\mathfrak{H}$, where a Hilbert space is a *vector space* $(\mathcal{H}, +, \cdot)$ together with a definition of a *sesquilinear*[2] inner product $(\cdot, \cdot)$, and $\mathfrak{H}$ is said to be *complex* since $(\mathcal{H}, +, \cdot)$ is a vector space over the field of complex numbers, $\mathbb{C}$.

$$\mathfrak{H} := \big(\mathcal{H}, +, \cdot, (\cdot, \cdot)\big) \tag{1}$$

In the general case, we would allow $\mathfrak{H}$ to be a *separable*[3] space, but in this thesis all discussions are limited to the finite-dimensional case, and we denote a Hilbert space of n dimensions with $\mathfrak{H}_n$. By conversion, we denote vectors ψ in Hilbert space surrounded by a *ket*.

$$|\psi\rangle \in \mathfrak{H}_n \tag{2}$$

We also introduce a shorthand notation (Dirac notation) for the sesquilinear product, $(\cdot, \cdot)$.

$$\big(\, |\phi\rangle \,, |\varphi\rangle \,\big) \equiv \langle\phi|\varphi\rangle \tag{3}$$

In elementary quantum theory we use *normalized* vectors $|\psi\rangle \in \mathfrak{H}_n$ to represent the states for the physical system, but this construction has one significant limitation. It is not possible to create a statistical or probabilistic mixture of states, since adding vectors from $\mathfrak{H}_n$ will just produce other vectors and not statistical mixtures. This limitation is overcome by modelling physical states with linear *maps*—usually denoted $\hat{\rho}$—which *act* on vectors in the appropriate Hilbert space $\mathfrak{H}_n$. We will here define properties for this map $\hat{\rho}$, and then derive and discuss its properties.

Axiom 2.1 (States in Quantum Mechanics). In Quantum Mechanics, a physical system with a finite number of states is associated with a finite dimensional Hilbert space $\mathfrak{H}_n$. States of the system are modelled by *positive semidefinite*, and *linear*, map $\hat{\rho} : \mathfrak{H}_n \to \mathfrak{H}_n$, for which $\mathfrak{Tr}\big[\hat{\rho}\big] = 1$.

[2] A *sesquilinear* inner product on a complex vector space, is an inner product that is linear in the second argument, and exchanging the vectors introduces a complex conjugation. This implies that it is linear under complex conjugation in the first argument.

$$\big(|\phi\rangle, |\varphi\rangle\big) \equiv \big(|\varphi\rangle, |\phi\rangle\big)^* \quad \wedge \quad \big(|\phi\rangle, b\,|\varphi\rangle\big) \equiv b\big(|\phi\rangle, |\varphi\rangle\big) \quad \Rightarrow \quad \big(a|\phi\rangle, |\varphi\rangle\big) = a^*\big(|\phi\rangle, |\varphi\rangle\big)$$

[3] Loosely speaking, a vector space is *separable* if it allows for an orthogonal and complete basis with a set of *countably infinite* basis vectors. $\{|e_i\rangle \mid i \in \mathbb{N} : \langle e_i|e_j\rangle = \delta_{ij}\}$

We have some remarks to this axiom.

Remark I. We also introduce a shorthand *Dirac notation* for taking the inner product between a vector $|\phi\rangle$, and the vector resulting from the map $\hat{\rho}$ acting on another vector $|\varphi\rangle$.

$$\left(|\phi\rangle , \hat{\rho} |\varphi\rangle \right) \equiv \langle \phi | \hat{\rho} \varphi \rangle \tag{4}$$

□

Remark II. A map $\hat{\rho}$ is said to be *positive semidefinite* if the complex scalar $\langle \varphi | \hat{\rho} \varphi \rangle$ is zero or positive for any $|\varphi\rangle \in \mathfrak{H}_n$. This implies that all eigenvalues for $\hat{\rho}$ are either zero or positive real numbers. □

Remark III. The word "operator" is frequently used instead of "map", and hence $\hat{\rho}$ is most often called a *density operator*. For finite Hilbert spaces we can choose some finite basis for $\mathfrak{H}_n$, and express the density operator as a matrix, thus $\hat{\rho}$ is sometimes also referred to as a *density matrix*. From now on we will assume this vocabulary. □

Remark IV. In general terms, the trace of some operator $\hat{A}$ is defined as a sum over an arbitrarily chosen orthonormal and complete basis $\{|\varphi_i\rangle\}$ for $\mathfrak{H}_n$, where we take the inner products with each basis vector, i.e. $\mathfrak{Tr}\left[\hat{A}\right] := \sum_i \langle \varphi_i | \hat{A} \varphi_i \rangle$. The trace is then shown to be invariant under change of basis (section A.4). □

The finite-dimensional case of the spectral theorem (see section A.1) allows us to show that $\hat{\rho}$, as defined by axiom 2.1, becomes a diagonal matrix in some appropriately chosen orthonormal basis $\{|\phi_i\rangle\}$ of $\mathfrak{H}_n$.[4] We denote the positive semidefinite eigenvalues in the eigenbasis $\{|\phi_i\rangle\}$ as $\{P_i\}$, and we can write $\hat{\rho}$ as a diagonal operator in terms of its eigenvectors.

$$\hat{\rho} = \sum_{i=1}^{n} P_i \, |\phi_i\rangle\langle\phi_i| \tag{5}$$

In section A.2 we demonstrate how to conduct the transition from a matrix notation, as in theorem A.1, to the sum over eigenvalues in the above expression, and the notation $|\phi_i\rangle\langle\phi_i|$ is defined there. Note that the requirement $\mathfrak{Tr}\left[\hat{\rho}\right] = 1$ from axiom 2.1 implies that the eigenvalues $\{P_i\}$ are normalized to unity. It is then possible to relate an eigenvalue P_i of some state $|\phi_i\rangle$ to the *probability for that state*.

From a physical point of view, equation (5) is more intuitively helpful than the initial axiom 2.1. Here we can view $\hat{\rho}$ as a classical probability distribution $\{P_i\}$ over a corresponding set of orthogonal pure states $\{|\phi_i\rangle\}$, and with these two sets we can describe the state of any quantum system.

However, even though a set of *orthonormal* pure states and their probabilities are sufficient, they are not necessary. In fact, any set of normalized pure states $\{|\varphi_j\rangle\}$ (not necessarily complete, nor orthogonal), and their corresponding probabilities $\{P_j \,|\, \sum_j P_j = 1\}$ is all we need to construct the state of a quantum system.

[4] We can make the argument brief by noting that since all eigenvalues of $\hat{\rho}$ are non-negative they are clearly real, and thus $\hat{\rho}$ is a *Hermitian operator*, a subset of *normal operators*, for which theorem A.1 applies.

$$\hat{\rho} = \sum_{\{|\varphi_j\rangle\}} P_j \, |\varphi_j\rangle\langle\varphi_j| \tag{6}$$

Note that if we express a density operator in some *arbitrary* basis, the resulting matrix will generally not be diagonal, as it was in equation (5). However, we know that some diagonalizing basis exists, so with some effort, any density operator can be written on its diagonal form.

In the remainder of the thesis, when we define some *density operator* $\hat{\rho}$ as *associated* with a Hilbert space $\mathfrak{H}_n$, we mean this in the sense of axiom 2.1, and we will use equation (5) or (6) to represent the system.

3 Entropy and the second law

The concept of *entropy* has a long history. In this section, we will make a few stops along the way, and discuss some foundational concepts in order to understand *the second law of Thermodynamics*. Note that we will not focus on historical results in all their colourful detail, but instead put emphasis on the more modern statistical formulations—where *probabilities* are used to define *entropy*, which in turn enable us to express the *second law*.

$$\text{Probability theory} \quad \rightarrow \quad \text{Entropy} \quad \rightarrow \quad \text{The second law of Thermodynamics}$$

We will also discuss entropy in Information Theory (see section 3.5), and derive Claude Shannon's famous formula from three credible assumptions. The mathematical properties exposed in this section will also be useful in physical considerations (for instance in section 5.6).

3.1 The importance of closed systems

Entropy arguments for physical systems—such as the one we will carry out in section 5—generally make statements about what must hold in some system *regardless* of the specific dynamics, and in order to make such strong statements, generally, we need to close off our system from outside tampering by evil demons.

Here, we define a *closed (physical) system* to mean one which has no interactions with any outside environment, or one for which the interactions are so weak as to make them negligible in the context of our theoretical model.[5]

3.2 Entropy and second law in Classical Thermodynamics

We shall begin a restricted historical expose of *entropy* with a brief discussion about the most central features in classical Thermodynamics, as developed by Clausius.

Let a *closed* physical system undergo some process, such that the system begins in a macroscopic state A, and ends in the macroscopic state B. The change in entropy, $\Delta S := S_B - S_A$ is defined, as an integral over a continuous ensemble of equilibrium states, where T is the temperature of the system, and $\mathrm{d}Q$ is an infinitesimal transfer of heat *into* the system. [4]

$$\Delta S := \int_A^B \frac{1}{T}\,\mathrm{d}Q \tag{7}$$

There are a few basic consequences we should point out.

Remark I. Only *changes* in entropy are well defined. So if we are asking for the entropy of a system at some particular state—such as, what the value of S_A is—it is only defined up to an arbitrary additive constant. □

[5] This definition can sometimes be referred to as an *isolated system*.

Remark II. This integral requires *temperature* to be well defined at each point. Thus for a closed system in Classical Thermodynamics, we can only define *entropy* for states in equilibrium. □

Remark III. The entropy S is derived and measured from the macroscopic properties of the system, $S(V, T, N)$, and there are no references to *microstates*, or probabilities, such as there will be in more modern statistical theories. (See section 3.3, 3.6, and 3.7.) □

In this framework, *the second law of Thermodynamics* becomes a proposed *macroscopic* principle for *closed systems in equilibrium*, stating that, *any change in entropy cannot be negative*. The principle is observed to hold in experiments.

$$\Delta S \geq 0 \tag{8}$$

3.3 Entropy and second law in Statistical Mechanics

In Boltzmann's kinetic theory of gases, he proposed a new definition of *entropy*—shown to be in agreement with Clausius[6]—that was built on statistical notions of the *microscopic* behaviour of a physical system, rather than its *macroscopic* properties.

Let M represent the macroscopic properties that we can observe a system to have (such as volume, pressure, and temperature). To the physical *macrostate* M, we assign a measure Ω_M, for *the number of accessible microstates* that our system can assume while satisfying M. Boltzmann then defined *entropy* for a macrostate M, as the logarithm of the number of accessible microstates.[7]

$$\boxed{S_B(M) := k \ln \Omega_M} \tag{9}$$

The factor k is *Boltzmann's constant*. It scales the entropy and provides appropriate units such that this notion of entropy agrees with Clausius.

Note that every microstate compatible with M is treated on equal footing (it simply adds 1 to Ω_M). This is equivalent with the assumption that *each microstate is equally probable*.

If in Clausius' Classical Thermodynamics, the second law felt somewhat ad hoc, with the statistical definition of entropy the second law becomes much less mysterious. We can state the second law of Thermodynamics as the following theorem, based on the analysis by Lev Landau and Evgeny Lifshitz, [6] (see pages 28 and 29).

[6] This connection is however not discussed here.

[7] Equation (9) can be derived from course-graining a classical phase space, as shown by Frigg and Werndl [5].

Theorem 3.1 (Second law of Thermodynamics, in Statistical Mechanics). If at some time t the entropy of a closed macroscopic system does not have its maximum value, then the (most probable[8]) development at subsequent times $t + \Delta t$ is such that the entropy will increase, or at least remain constant.

$$S_B(t + \Delta t) - S_B(t) \geq 0 \qquad \forall \ \Delta t \geq 0 \tag{10}$$

At macroscopic equilibrium, the system will assume the macrostate with the greatest possible entropy.

To motivate why the above theorem 3.1 holds, consider two macrostates M and N of some physical system, such that $\Omega_M \gg \Omega_N$. The system will overwhelmingly prefer the macrostate M, not because any microstates compatible with M is somehow different to those compatible with N, but simply because there are so many more of them. A random walk through the space of microstates is almost guaranteed to give us the macroscopic behaviour M, and in most considerations, due to large number of degrees of freedom in any macroscopic system, a very large number will be assigned to Ω for a very small set of closely related macroscopic states, and Ω has negligible values for all other macroscopic states. This situation is often characterized by noting that Ω has a very sharp peak around some macrostate M. Thus decreasing entropy is strictly speaking not prohibited, it is just really, really unlikely to happen in large systems. Typically, we can expect to consider timescales much longer than the age of the universe before a macroscopic system spontaneously does something unexpected.

Note however that the situation changes when we consider small systems, with considerably fewer degrees of freedom. Suddenly, fluctuations into states of lower entropy are more likely, and this will motivate us to talk about the behaviour of entropy, *on average*. In section 3.6 we will take a closer look at small systems.

Another important remark is that—even though we assert that entropy cannot decrease—this says nothing about *how fast* the entropy will increase. Depending on properties of the system, the evolution to a high entropy state will progress at differing rates.

3.4 Probabilities in physical systems

As we are moving closer to modern concepts of entropy, and probabilities will become central for definitions.[9] As we shall later demonstrate in section 3.5, any probability distribution $\{P(x)\}$ over some finite set of mutually exclusive events $\{x\}$ can be associated with an entropy. However, the conceptual interpretation assigned to *entropy* will partially depend on our attitude towards the probabilities in the first place. Therefore, we begin by a brief exposition of probabilities—as they are understood and conceptualized in physics.

[8] For large (macroscopic) systems the probability of transitions into states of lower entropy is so unlikely that they, for all practical purposes, are never observed.

[9] Already in the Boltzmann formulation, we said that each microstate was equally *probable*.

Following closely an argument made by Edwin T. Jaynes [7], broadly speaking we can divide attitudes towards probabilities into two different schools of thought, here referred to as the "objective" and "subjective" attitude.

The *objective school of thought* regards the *probability of an event* as some objective property of that event, or the physical system which generates it. This is predominantly meant in the following sense: Underlying any probability there is either some physical propensity for different events[10], or, some measurement of a frequency ratios that could (at least in principle) be made, and in the limit of an infinite number of measurements, probabilities will be reproduced from the frequencies of experimental outcomes. In the objective attitude, testing whether a probability distribution $\{P(x)\}$ is accurate, is to answer the question: "Does $\{P(x)\}$ correctly represent *the observed* distribution over $\{x\}$?"

We can note that in the physical world it is difficult to come by an infinite set of measurements. Thus, relying solely on frequency ratios implies that we can never be absolutely certain that a probability distribution is accurate, we can only reduce our doubts to be arbitrarily small.

In contrast, the *subjective school of thought* regards probability as an expression of the ignorance that some agent has. In this approach, the *probability of an event* is a formal expression of some degree of belief that an event will occur—when taking all available information into account. In this approach, we are concerned with finding the best possible inference, when there is in fact not enough available information to construct a certain prediction. In the subjective attitude, testing whether a probability distribution $\{P(x)\}$ is accurate is to answer the question: "Does $\{P(x)\}$ correctly represent our *state of knowledge* about the value of $\{x\}$?"

Since any frequency measurement or physical propensity can be incorporated into *our state of knowledge*, any question asked in the objective framework also has meaning in the subjective framework [7]. But there are some ideas one may pursue that only seems to make sense in the context of subjective probabilities. For instance, the approach taken in section 4.2, where we will derive the canonical thermal state in Quantum Mechanics by finding the most appropriate representation given incomplete information.

However, as Jaynes points out, we can expect both the objective and subjective schools of thought to be applicable in physics, and "needless controversy has resulted from attempts to uphold one or the other in all cases" [7].

Regardless of what point of view is found more suitable in some particular situation—the mathematical framework for both attitudes are identical, and thus their differences are mostly related to the kind of questions we might ask.

3.5 Entropy in Information Theory

In 1948, Claude Shannon published a seminal paper "The mathematical theory of communication" [9] that would become the starting point for the field of Information Theory.[11]

A central achievement of Shannon's is his idea to look for a measure of

[10] For instance a symmetry argument can convince us that the *propensity* of a coin-toss to give us heads must be 1/2.

[11] The original title was "*A* mathematical theory of communication.", but in 1949 the paper was published as a book, coauthored with Warren Weaver, with a slightly altered title.

how much "uncertainty" that resides in a probability distribution. Usually, this kind of *uncertainty* is referred to as *Shannon entropy*[12], and it is calculated from the probabilities $\{P(x)\}$ over a finite set of *mutually exclusive* events $\{x\}$. The expression for *Shannon entropy* H is derived as a unique solution to three axioms that we require H to satisfy. The axioms and the derivation is presented in section 3.5.3, and the result is as follows.

$$H\big(\{P(x)\}\big) = -K \sum_{\{x\}} P(x)\, \log_2 P(x) \tag{11}$$

Here, we define $0 \log_2 0 := \lim_{x\to 0}(x \log_2 x) = 0$, and the constant K simply determines the unit in which *entropy* is measured. Note that choosing a different basis for the logarithm is equivalent to changing K by some numerical factor. For most applications in Information Theory we set $K = 1$, and thus measure information in *bits*.[13]

$$\boxed{H\big(\{P(x)\}\big) := -\sum_{\{x\}} P(x)\, \log_2 P(x) \quad \text{(in bits)}} \tag{12}$$

Let us examine an example to illustrate some key points, and demonstrate how equation (12) can be useful.

3.5.1 Example: Application of Shannon entropy

Consider a device that has three light-bulbs in the colours *red*, *green*, and *blue*. When turned on, the device will light up one bulb at a time, to create a random sequence of *events*. I.e. we may get a sequence such as:

$$\{\text{green, blue, red, red, red, blue, red, } \cdots\} \tag{13}$$

However—to make this example a bit more interesting—we assume the device is constructed such that the probability of red is twice the probability of green or blue.

$$P(\text{red}) = \frac{1}{2} \qquad ; \qquad P(\text{green}) = P(\text{blue}) = \frac{1}{4} \tag{14}$$

We then want to record whatever sequence of colours the device is producing, using a binary memory (such as a computer memory). And we ask ourselves: What is the theoretical minimum for the number of *bits* needed to store a sequence of colours produced by the device, on average? Or equivalently: What is the *average* number of bits needed to store a single event (red, green or blue)?

We can think of a *bit* as the answer to a yes/no question (where we can label "yes" with `1`, and "no" with `0`). In this context, the average number of *bits* will translate to the average number of questions we need to ask. Since red is the most likely colour, it will be a good idea to first ask: "Is the bulb red?", and

[12] Sometimes also referred to as *information*, *Shannon information* or simply *entropy*. An appropriate label can be determined from context, and how to make this distinction is discussed later in this section.

[13] We note that *Shannon entropy* has no physical unit in the same manner as say mass, distance, or *Boltzmann entropy* (section 3.3), but we still need to distinguish between the different measures, similarly to how degrees and radians are differentiated.

with probability 1/2 we do not need to ask any more questions. However, if the answer is "no", we have to ask again: "Is the bulb green?" Then, regardless of the answer, we will know the colour of the bulb. Thus we can calculate how many bits (or questions) we need on average, by weighing them by their probabilities.

$$\frac{1}{2}\,1\,\text{bit} + \frac{1}{2}\,2\,\text{bits} = 1.5\,\text{bits} \tag{15}$$

However, not all situations lend themselves to be easily analyzed in terms of yes/no questions. In that case, we can use the general formula of *Shannon entropy*, from equation (12), to find the minimum amount of information we need to store.

$$\begin{aligned} H_{\text{rgb}} &= -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{4}\log_2\frac{1}{4} + \frac{1}{4}\log_2\frac{1}{4}\right) = \\ &= \frac{1}{2}\log_2 2 + \frac{1}{2}\log_2 4 = \frac{1}{2}\,(1+2) = 1.5\,\text{bits} \end{aligned} \tag{16}$$

Given that we can store the information about the colours optimally, we will need (on average) 1.5 bits for each event. Thus for a sequence of, say, 10 000 colours, we can expect to consume about 15 000 bits (again, *on average* since any particular sequence can vary in its memory consumption).

Conceptually, there are a few different ways to approach the result of this calculation. Let us say that we are considering a sequence of 10 000 events. The simplest conceptualization is just in terms of how much physical memory in our binary storage we will need for such a series, and we can loosely speaking say that we have calculated some amount of *knowledge*, or *information*.

However, in physics, there is a different conceptualization which is very useful. Consider the set of *every* possible sequence if 10 000 events, let us call it $X_{10\,000}$. Since the cardinality of this set is very large ($3^{10\,000}$ members), the probability of any specific member being realized is of course very low, but more importantly, the probabilities are not uniformly distributed. Members that have a distribution of red, green, and blue close to that of the probabilities in equation (14) will be more likely than members of $X_{10\,000}$ with a distribution of red, green, and blue that does not reflect these probabilities. So even though we are quite clueless about which member of $X_{10\,000}$ will be realized when we turn on the device, the fact that the distribution is not uniform over $X_{10\,000}$, actually constitutes some knowledge or *information* about the outcome. In this construction, we can view the calculation from equation (16) as an answer to the question: How much *more* information do we have to provide (on average), to reduce the initial probability distribution over $X_{10\,000}$, to a distribution with a certain outcome (where the distribution has a spike at one single member and is zero everywhere else)? The answer is 15 000 bits (on average), and we can loosely speaking say that we have calculated our *uncertainty*, *ignorance*, or *entropy* associated with the initial, non-uniform, probability distribution over $X_{10\,000}$.

We can then argue that our *uncertainty* is just some measure of how much *knowledge* we lack, and likewise, we can generally view *entropy* as an absence of *information*. Since this mathematics of *information*, or *uncertainty*, are identical, we can only turn to the context of the initial question to determine how to think conceptually about the result of such a calculation. For example, when

calculating the entropy as described by Boltzmann, in section 3.3, we are clearly evaluating some quantity of ignorance, since we never expect to actually measure the exact microstate of a large macroscopic system.

3.5.2 Example: The composition law

Before we will walk through Shannon's assumptions in the general case, and derive equation (11), let us look at a concrete example of the so-called "composition law", one property which we will require H to satisfy.

Consider the following probability distribution over mutually exclusive events.

$$\left\{\frac{1}{12}, \frac{1}{4}, \frac{1}{6}, \frac{1}{2}\right\} \tag{17}$$

Say that we combined the first and second event into one event with total probability $1/3$, yielding a new set of probabilities $\{1/3, 1/6, 1/2\}$. We can draw two *tree diagram*, one for each situation.

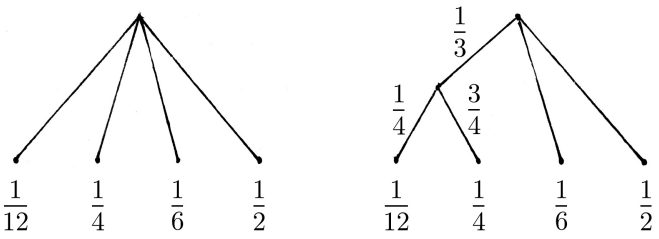


Figure 1: Two *tree diagrams* for identical probability distributions. In the rightmost tree, the *events* (nodes in the diagram) with probabilities $1/12$ and $1/4$ have been composed to create two *successive* events.

Since the final probabilities in both trees are identical, the total amount of uncertainty (or *entropy*) H should be equal in both cases. However, for the rightmost tree of figure 1, the total *uncertainty* should be possible to decompose into our uncertainty from $\{1/3, 1/6, 1/2\}$, with some remaining uncertainty, should the event with probability $1/3$ be realized. The remaining uncertainty should be calculated as if it was any independent probability distribution, however, since it is only realized with probability $1/3$ we put this weight factor in front.

$$H\left(\frac{1}{12}, \frac{1}{4}, \frac{1}{6}, \frac{1}{2}\right) = H\left(\frac{1}{3}, \frac{1}{6}, \frac{1}{2}\right) + \frac{1}{3}\, H\left(\frac{1}{4}, \frac{3}{4}\right) \tag{18}$$

This is a property that we wish any measure H of *uncertainty*, *entropy*, or *information* to have, and we will generalize the idea in the next section as *the composition law* (axiom 3.3).

3.5.3 Deriving Shannon entropy

Shannon's seminal paper from 1948 [9] includes a derivation of the entropy formula. However, when it comes to explanatory details, Shannon's proof is quite

barren. Here, we endeavour with a detailed reconstruction of the derivation, exposing every step and removing any ambiguity so that the reader can follow the derivation with minimum effort. It would generally be appropriate to relegate a proof like this to the appendix, but there are some details here that can create a deeper understanding of entropy. However, the reader in a hurry, or anyone familiar with Shannon's derivation, can skip ahead to section 3.5.4.

Consider a finite set of n *mutually exclusive events*, $\{x\}$. Our intention is to derive the relation for *Shannon entropy*, shown in equation (11), for a probability distribution $\{P(x)\}$ over this set of events.

Shannon begins by defining three axioms that a measure, H, of *uncertainty*, a.k.a. *entropy*, should satisfy.

Axiom 3.1. H is continuous in all probabilities $\{P(x)\}$.

Axiom 3.2. If all $\{P(x)\}$ are equal, i.e. $P(x) = 1/n$, then H should be a strictly increasing function of n.[14]

Axiom 3.3 (The composition law). If a subset of $\{x\}$ is composed to create successive events, the original H splits into two terms. One represents a measure of entropy in the first set of events, and the second represents the entropy in the remaining events, weighted with its probability—as discussed in detail below.

Consider selecting a subset of members from the set $\{P(x)\}$ and compose them into a set C, like it was a probability of a single event.

$$C = \{P(x_C)\} \qquad \text{for some subset} \qquad \{P(x_C)\} \subseteq \{P(x)\} \tag{19}$$

We denote the set of members that were not selected as $\{P(x_R)\}$; they are the *remaining* members. By construction we then have that $C \cup \{P(x_R)\} = \{P(x)\}$, and $C \cap \{P(x_R)\} = \varnothing$. We define the probability of the composed set, $P(C)$, as the sum of its individual probabilities.

$$P(C) := \sum_C P(x) \tag{20}$$

In a new set C', we renormalize the probabilities from C such that C' sums up to 1.[15]

$$C' := \left\{ P(x) \;\middle|\; \forall\, P(y) \in C : P(x) = \frac{P(y)}{P(C)} \right\} \quad \Rightarrow \quad \sum_{C'} P(x) = 1 \tag{21}$$

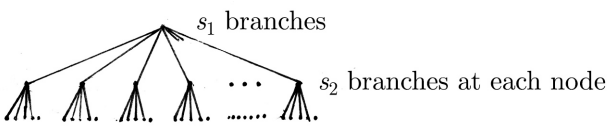
The entropy can then be evaluated from the probability for the composed event, $P(C)$, together with all probabilities for the remaining events $\{P(x_R)\}$, plus the entropy which remains if C should be realized. The latter happens with probability $P(C)$ and is thus weighted accordingly.

$$H\big(\{P(x)\}\big) = H\big(\{P(C)\} \cup \{P(x_R)\}\big) + P(C)\, H(C') \tag{22}$$

[14] With equally likely events, there is more *choice*, or *uncertainty*, when there are more possibilities.

Before immersing into the derivation of Shannon entropy, let us formulate a corollary from *the composition law* (property 3.3) that will be useful in our proof.

Corollary 3.1 (to property 3.3). Let $n = |\{P(x)\}|$ be divisible by s_1, such that $n = s_1 s_2$. Also, let all probabilities be equal $P(x) = 1/n \ \forall \ x$. We can apply the composition law s_1 times on the set $\{P(x)\}$ in order to create a *tree diagram* where the first level branches into s_1 events, each with probability $1/s_1$. On the second level, each event branches s_2 times, each with probability $1/s_2$.



Calculating the entropy for each level separately, according to *the composition law*, will yield the following result.

$$H\left(\frac{1}{n}, \cdots, \frac{1}{n}\right) = H\left(\frac{1}{s_1}, \cdots, \frac{1}{s_1}\right) + H\left(\frac{1}{s_2}, \cdots, \frac{1}{s_2}\right) \tag{23}$$

Proof (Corollary 3.1). We apply the composition law (property 3.3) sequentially, at the set of remaining probabilities $\{P(x_R)\}$. After the first application, we replace the first s_1 arguments in H by one argument and add one weighted term.

$$H = H\left(\frac{1}{s_1}, \frac{1}{n} \cdots, \frac{1}{n}\right) + \frac{1}{s_1} H\left(\frac{1}{s_2}, \cdots, \frac{1}{s_2}\right) \tag{24}$$

The pattern repeats for each application, and after the second application we have two weighted terms.

$$H = H\left(\frac{1}{s_1}, \frac{1}{s_1}, \frac{1}{n} \cdots, \frac{1}{n}\right) + \frac{1}{s_1} H\left(\frac{1}{s_2}, \cdots, \frac{1}{s_2}\right) + \frac{1}{s_1} H\left(\frac{1}{s_2}, \cdots, \frac{1}{s_2}\right) \tag{25}$$

After s_2 applications we have replaced all n arguments of our original H with s_1 arguments (where $s_1 \leq n$ is typically much smaller than n), and we have introduces a series of weighted terms.

$$H = H\left(\frac{1}{s_1}, \cdots, \frac{1}{s_1}\right) + \frac{1}{s_1} H\left(\frac{1}{s_2}, \cdots, \frac{1}{s_2}\right) + \cdots + \frac{1}{s_1} H\left(\frac{1}{s_2}, \cdots, \frac{1}{s_2}\right) \tag{26}$$

Since we have s_1 identical terms with the coefficient $1/s_1$, the expression can be simplified to the right-hand side of equation (23), of corollary 3.1. ■

Note that we can apply *the composition law*, and thus corollary 3.1, not only on the entropy corresponding to the first level in a *tree diagram*. We can also create sequences of events that are larger than two, as illustrated in figure 2.

[15] We have overlooked the pathological case when $P(C) = 0$, where renormalization according to equation (21) becomes undefined. In that case we simply define $H(C') := 0$ in equation (22).

Figure 2: Illustration of *the composition law* applied at several levels in of a *tree diagram*.

Armed with corollary 3.1 and some understanding of *the composition law*, we will state, and then prove, the formula for Shannon entropy.

Theorem 3.2. The only function $H(\{P(x)\})$ that satisfies the axioms 3.1, 3.2, and 3.3, is on the following form, where we define $0 \log_2 0 := 0$.

$$H\big(\{P(x)\}\big) := -K \sum_{\{x\}} P(x) \log_2 P(x) \tag{27}$$

Proof (Theorem 3.2). Consider the special case where all probabilities are equal, $\{P(x)\} = 1/n$, and to keep the notation clear we define a special function f for this case.

$$H\left(\frac{1}{n}, \frac{1}{n}, \cdots, \frac{1}{n}\right) =: f(n) \tag{28}$$

Then—again looking at another special case—consider n being a whole power, $n = s^\sigma$. We can then apply corollary 3.1 σ times, each time at the level below the current. This gives us a *tree diagram* with σ levels, where each node branches into s equally likely possibilities. From equation (23), of corollary 3.1, we find that we should add one term $f(s)$ for each level in the diagram, and thus f can be evaluated in two ways.

$$f(s^\sigma) = \sigma f(s) \tag{29}$$

For some other whole power $n = t^\tau$, we of course get the same relation.

$$f(t^\tau) = \tau f(t) \tag{30}$$

Then, for every freely chosen values of s, t and σ, there exists some τ to fulfill the following inequality.

$$t^\tau \leq s^\sigma < t^{\tau+1} \tag{31}$$

We take the logarithm of this inequality (using $\log_2$ for later convenience). Then we divide by $\sigma \log_2 t$, and subtract τ/σ.

$$\tau \log_2 t \;\leq\; \sigma \log_2 s \;<\; (\tau+1) \log_2 t \qquad \Rightarrow \tag{32}$$

$$\frac{\tau}{\sigma} \;\leq\; \frac{\log_2 s}{\log_2 t} \;<\; \frac{\tau}{\sigma} + \frac{1}{\sigma} \qquad \Rightarrow \tag{33}$$

$$0 \;\leq\; \frac{\log_2 s}{\log_2 t} - \frac{\tau}{\sigma} \;<\; \frac{1}{\sigma} \tag{34}$$

We will come back to this result. Now, from property 3.2, we find that f must be a strictly increasing function, thus we can translate the inequality from equation (31) to an inequality in f, and then we apply equation (29).

$$t^\tau \leq s^\sigma < t^{\tau+1} \qquad \Rightarrow \qquad f(t^\tau) \leq f(s^\sigma) < f(t^{\tau+1}) \qquad \Rightarrow \tag{35}$$

$$\tau f(t) \;\leq\; \sigma f(s) \;<\; (\tau+1) f(t) \tag{36}$$

We divide this by $\sigma f(t)$, and subtract τ/σ.

$$\frac{\tau}{\sigma} \;\leq\; \frac{f(s)}{f(t)} \;<\; \frac{\tau}{\sigma} + \frac{1}{\sigma} \qquad \Rightarrow \tag{37}$$

$$0 \;\leq\; \frac{f(s)}{f(t)} - \frac{\tau}{\sigma} \;<\; \frac{1}{\sigma} \tag{38}$$

We then combine equation (34) and (38) by subtracting the first from the second, such that τ/σ cancel, and the result is bounded by $\pm 1/\sigma$.

$$-\frac{1}{\sigma} \;<\; \frac{f(s)}{f(t)} - \frac{\log_2 s}{\log_2 t} \;<\; \frac{1}{\sigma} \qquad \Rightarrow \tag{39}$$

$$\left| \frac{f(s)}{f(t)} - \frac{\log_2 s}{\log_2 t} \right| \;<\; \frac{1}{\sigma} \tag{40}$$

All the variables s, t, and σ were chosen as free parameters, thus this relation holds for any values. In particular, we must require this to hold for any s and t, as $\sigma \to \infty$.

$$\lim_{\sigma\to\infty} \left| \frac{f(s)}{f(t)} - \frac{\log_2 s}{\log_2 t} \right| = 0 \qquad \forall\, s,t \qquad \Rightarrow \tag{41}$$

$$f(s) = K \log_2 s \tag{42}$$

From property 2, we see that K must be chosen positive in order for f to be strictly increasing.

$$f(s) = K \log_2 s \qquad \text{where} \qquad K > 0 \tag{43}$$

With this result, we retrace our steps using the equalities $f(s^\sigma) = \sigma f(s)$ and $n = s^\sigma$.

$$f(n) = f(s^\sigma) = \sigma f(s) = \sigma K \log_2 s = K \log_2 s^\sigma = K \log_2 n \tag{44}$$

Thus, we have found H for equal probabilities.

$$H\left(\frac{1}{n}, \frac{1}{n}, \cdots, \frac{1}{n}\right) = K \log_2 n \tag{45}$$

However, we are of course looking for the general case, with arbitrary probabilities. Thus we begin anew and consider some finite set of events $\{x\}$, with cardinality n, and whose probabilities are limited to *rational* values $\{P(x) | P(x) \in \mathbb{Q}\}$. This means that there exists some common denominator p such that each probability $\{P(x)\}$ can be written as a fraction of whole numbers.

$$P(x) = \frac{p_x}{p} \qquad \text{where} \qquad p_x \in \mathbb{N}_0 \quad \text{and} \quad p = \sum_{\{x\}} p_x \tag{46}$$

Such probabilities can be created from composing smaller probabilities from a

set of p members with equal probabilities. Thus we apply *the composition law* on $H(1/p, \cdots, 1/p)$, n times.

$$\begin{gathered} H\Big(\frac{1}{p}, \cdots, \frac{1}{p}\Big) = \\ = H\Big(\frac{p_1}{p}, \cdots, \frac{p_n}{p}\Big) + \frac{p_1}{p} H\Big(\frac{1}{p_1}, \cdots, \frac{1}{p_1}\Big) + \cdots + \frac{p_n}{p} H\Big(\frac{1}{p_n}, \cdots, \frac{1}{p_n}\Big) \end{gathered} \tag{47}$$

The term $H(p_1/p, \cdots, p_n/p) = H(\{P(x)\})$ is what we are looking for.

$$\begin{gathered} H\big(\{P(x)\}\big) = \\ = H\Big(\frac{1}{p}, \cdots, \frac{1}{p}\Big) - \frac{p_1}{p} H\Big(\frac{1}{p_1}, \cdots, \frac{1}{p_1}\Big) - \cdots - \frac{p_n}{p} H\Big(\frac{1}{p_n}, \cdots, \frac{1}{p_n}\Big) \end{gathered} \tag{48}$$

Note how the entropy we are looking for is the total entropy over the very large set of p events, minus the entropies that comes from each composition, weighted with its probability.

We multiply the first term by $1 = \sum_x P(x) = \sum_x p_x/p$, and group terms with identical coefficients.

$$\begin{gathered} H\big(\{P(x)\}\big) = \\ = \frac{p_1}{p}\left(H\Big(\frac{1}{p}, \cdots, \frac{1}{p}\Big) - H\Big(\frac{1}{p_1}, \cdots, \frac{1}{p_1}\Big) \right) + \cdots + \\ + \frac{p_n}{p}\left(H\Big(\frac{1}{p}, \cdots, \frac{1}{p}\Big) - H\Big(\frac{1}{p_n}, \cdots, \frac{1}{p_n}\Big) \right) \end{gathered} \tag{49}$$

From equation (45), we have an expression for H over equal probabilities.

$$\begin{gathered} H\big(\{P(x)\}\big) = \\ = \frac{p_1}{p}\left(K \log_2 p - K \log_2 p_1\right) + \cdots + \frac{p_n}{p}\left(K \log_2 p - K \log_2 p_n\right) = \\ = -K\big(\frac{p_1}{p} \log_2 \frac{p_1}{p} + \cdots + \frac{p_n}{p} \log_2 \frac{p_n}{p}\big) \end{gathered} \tag{50}$$

As defined, $p_x/p = P(x)$, and thus we have found the Shannon entropy, from equation (11), for rational probabilities.

$$H\big(\{P(x)\}\big) := -K \sum_{\{x\}} P(x) \log_2 P(x) \tag{51}$$

To show that this expression is true for non-rational probabilities, we only have to note that the rational numbers are dense in $\mathbb{R}$, and axiom 3.1 demands that the function is continuous. Thus it holds for any $P(x) \in \mathbb{R}$. ■

3.5.4 Entropy is additive

An important property of entropy is how it relates to so-called *joint events*.

Consider two sets of events $\{x\}$ and $\{y\}$, where each set contains mutually exclusive and collectively exhaustive events, but *both* an event in $\{x\}$ and in $\{y\}$

will occur. Let $\{P(x,y)\}$ be the *joint probability distribution* over the combined events $\{x\}$ and $\{y\}$. The entropies over the joint probability distribution, and the individual ones, are then defined as follows.

$$H\big(\{P(x,y)\}\big) := -\sum_{\{x\},\{y\}} P(x,y)\,\log_2 P(x,y) \tag{52}$$

$$H\big(\{P(x)\}\big) := -\sum_{\{x\}} P(x)\,\log_2 P(x) \qquad \text{where} \qquad P(x) = \sum_{\{y\}} P(x,y) \tag{53}$$

$$H\big(\{P(y)\}\big) := -\sum_{\{y\}} P(y)\,\log_2 P(y) \qquad \text{where} \qquad P(y) = \sum_{\{x\}} P(x,y) \tag{54}$$

It is then possible to show that the sum of individual entropies is always equal to or larger than the entropy of the joint probability distribution. [1]

$$H\big(\{P(x)\}\big) + H\big(\{P(y)\}\big) \geq H\big(\{P(x,y)\}\big) \tag{55}$$

Here, equality is assumed if and only if the events are independent of each other, i.e. $P(x,y) = P(x)P(y)$. Thus *entropy of independent events is additive.*

3.6 Generalized entropy and the second law

With the framework of evaluating entropy from probability distributions over sets—in section 3.5—we can begin to assign an entropy to a broader class of physical systems than those where Boltzmann entropy is applicable (section 3.3). Say that we for some reason or another cannot assign a *uniform* probability distribution to our set of accessible microstates $\{\mu\}$, but instead need to rely on a variable probability distribution $\{P(\mu)\}$. Additionally, let us assume that the number of microstates is finite—this will be sufficient for our purposes—and clearly, the states are required to be *mutually exclusive*, just as our *events* were when introducing Shannon Entropy. In the literature, this type of entropy is referred to as *Gibbs entropy*, hence we denote it S_G.[16]

$$\boxed{S_G\big(\{P(\mu)\}\big) := -k\sum_{\{\mu\}} P(\mu)\,\ln P(\mu)} \tag{56}$$

Remark I. Clearly, this *Gibbs entropy* looks very similar to *Shannon Entropy* from section 3.5, equation (12)—with the noticeable differences being the basis for the logarithm, and Boltzmann's constant (scaling the quantity and supplying physical units). However, we should emphasize that this probability distribution $\{P(\mu)\}$ has an entirely different origin; relating to a model of a *physical system*, and not some abstract *information*. Nevertheless, since the mathematical formulation is the same, we can prove theorems in either framework, it is just that results will have different interpretations depending on the context in which we operate. □

[16] The name *Gibbs Entropy* often refers to an integral over some *probability density*, ρ, over some continuous *phase space* Γ.

$$S_{G'} := -k\int_{\Gamma} \rho\ln\rho\,\mathrm{d}\Gamma$$

This formula will transform into the sum in equation (56) if we let the continuous phase space Γ become a finite set of states $\{\mu\}$.

Remark II. Comparing this entropy to *Boltzmann entropy* from section 3.3, in Statistical Mechanics, we simply counted all microstates compatible with the *macrostate* of the system (giving each an equal probability). Here, *macrostates* will be replaced by some set of *conditions*, and under these conditions, we calculate a probability distribution $\{P(\mu)\}$ which maximizes entropy, according to the *principle of maximum entropy inference* (see section 4). These two approaches are similar, but not identical. □

We then suppose that the *second law of Thermodynamics* still holds for *Gibbs entropy*. This is of course a stronger claim than the initial motivation done in the framework of regular Statistical Mechanics (section 3.3), and it can be considered controversial. The issue with this assumption has to do with the *Liouville theorem*, which shows that *Gibbs entropy* remains constant under Hamiltonian evolution (valid for both classical and quantum systems). There are a number of proposed solutions to this; the statistical interpretation, coarse-graining, projection, chaos, quantum collapse, expansion of the universe, non-equilibrium initial condition, and canonical typicality [10]. Discussing each of these will however bring us too far off course, so instead, we take this *generalized second law of Thermodynamics* as an axiom.

Axiom 3.4 (Generalized second law of Thermodynamics). Consider a *closed system* with a finite set of mutually exclusive microstates $\{\mu\}$, and their time-dependent probabilities $\{P(\mu, t)\}$. When comparing the system at some time t to any subsequent time $t + \Delta t$, its state will develop such that the Gibbs entropy *on average* will increase, or at least remain constant.

$$S_G(t) := -k \sum_{\{\mu\}} P(\mu, t) \ln P(\mu, t) \tag{57}$$

$$\big\langle S_G(t+\Delta t) - S_G(t) \big\rangle \geq 0 \quad \forall \ \Delta t \geq 0 \tag{58}$$

Remark I. Note that in the case of an equal probability distribution over some subset of $\{\mu\}$, this generalized law is identical to the standard *second law of Thermodynamics* (theorem 3.1 in section 3.3). □

Remark II. When we considered Boltzmann entropy in section 3.3 we assumed that the system was large enough for the number of accessible microstates to have a very sharp peak around some particular macrostate M. Here we relax this condition and consider physical systems of any size. In particular we allow the system to have a small number of degrees of freedom, and consequently, fluctuations into states of lower entropy become much more likely. This means that we can only consider the generalized second law to hold in a statistical limit, i.e. we have to consider averages of entropy. □

Remark III. We will intentionally be somewhat vague about what kind of average for the entropy we are considering. We can take this average to mean either the behaviour of some large ensemble of systems, the averaged behaviour of one system over many cycles, or some notion from a more subjective attitude towards the *probable* behaviour of the system (see section 3.4). Then this generalized second law can be applied in a broader range of situations. □

3.7 Entropy in Quantum Mechanics

Considering the historical account, we should note that entropy in quantum systems—the so-called *von Neumann entropy*—was introduced in 1927 by John von Neumann [11], much earlier than to Shannon's work [9] for instance, which was published in 1948. However, here the topics are organized based on their conceptual relation rather than their historical development.

Von Neumann originally derives his entropy formula from a specific thought experiment using boxes and walls [12]. But here we will simply have a look at the result and compare it to previous discussions, in particular section 3.5.

Let the state of some physical system, $\hat{\rho}$, be associated with the finite Hilbert space $\mathfrak{H}_n$ (as defined in axiom 2.1), then *von Neumann entropy*, S_N, has the following definition.

$$\boxed{S_N(\hat{\rho}) := -\mathfrak{Tr}\left[\hat{\rho}\,\ln\hat{\rho}\right]} \tag{59}$$

This compact equation contains several details that we ought to consider.

Remark I. Considered in the general case, the *natural logarithm* of some operator $\hat{A}$ is defined as the operator $\hat{X}$ that solves $\mathrm{e}^{\hat{X}} = \hat{A}$. Solutions are not guaranteed to exist, or there may be an infinite number of solutions. However, according to axiom 2.1, $\hat{\rho}$ has no negative eigenvalues, and then one can show that there exists one, and only one, solution. (See the appendix, section A.3, for further discussions about the logarithm of operators.) □

Remark II. Again in the general case, the trace of some operator $\hat{A}$ is defined as a sum over an arbitrarily chosen basis $\{|\phi_i\rangle\}$ for $\mathfrak{H}_n$, where the trace is then shown to be invariant under change of basis (see section A.4).

$$\mathfrak{Tr}\left[\hat{A}\right] := \sum_{i=1}^{n} \langle\phi_i|\hat{A}\phi_i\rangle \tag{60}$$

□

Remark III. In the appendix, section A.5, we show that this von Neumann entropy is invariant under unitary transformations of $\hat{\rho}$.

$$S_N(\hat{U}\,\hat{\rho}\,\hat{U}^\dagger) = S_N(\hat{\rho}) \tag{61}$$

□

Remark IV. Von Neumann entropy is defined in terms of the natural logarithm. This means that the unit for von Neumann entropy is different from Shanon entropy by a factor $\ln 2$. Note however that we used the natural logarithm for Boltzmann entropy in equation (9), section 3.3. Ordinarily we use the natural logarithm for entropy in physical systems, and the base 2 logarithm for entropy in some abstract information. □

Often, when we want to evaluate $S_N(\hat{\rho})$ in some practical situation, we rewrite $\hat{\rho}$ on its diagonal form $\hat{\rho} = \sum_i P_i\,|\phi_i\rangle\langle\phi_i|$, and take the trace in the same eigenbasis. Then $S_N(\hat{\rho})$ will reduce to a sum over the eigenvalues.

$$S_N(\hat{\rho}) = -\sum_i P_i \ln P_i \tag{62}$$

This is clearly very similar to the relation for *Gibbs entropy* in equation (56) (except for the scaling factor k), and also *Shannon entropy* in equation (12) (except for the logarithm basis which introduces a conversion factor $\ln 2$). In this light, we can understand *von Neumann entropy* as an extension into Quantum Mechanics, and some of the work we did in the previous sections in order to understand *entropy* (in particular section 3.5 and 3.6) applies here as well.

4 The principle of maximum entropy inference

Here, we will consider ideas of Edwin T. Jaynes from his two influential 1957 papers [7, 8]. In particular, we are interested in the *principle of maximum entropy inference*, a.k.a. the *maximum entropy principle*.

Jaynes' original justification of the principle can be blamed for not being very formal. Therefore we attempt to reinforce the idea with some additional rigour (section 4.1).

We will then apply this principle in a concrete example, to demonstrate its usefulness; deriving the canonical thermal state in Quantum Mechanics (section 4.2), later to be used in section 6.

Jaynes' argument begins with a reference to a tentative principle conceived by Laplace. It is a useful rule—which applies to a limited situation—for assigning probabilities when frequency measurements are not available, called the *principle of insufficient reason*.

Suppose we have some discrete and finite set of events $\{x\}$, such that they are *mutually exclusive* and *collectively exhaustive*. Then, suppose that all events are *indistinguishable*, except for how we label them. The *principle of insufficient reason* states that the events should each be assigned an equal probability, $P(x) = 1/n \ \forall \ x$ (where n is the number of events). For example, consider a die where all six sides are physically symmetric except for the label we have assigned to each side. It is then appropriate to consider this a fair die, with probability 1/6 for any outcome. The *principle of maximum entropy inference* can then be seen as a generalization of Laplace's thinking.

Suppose again that we have some system whose *state* can be *characterized* by a probability distribution over a set of mutually exclusive states $\{x_1, x_2, \cdots, x_n\}$ $\equiv \{x_i\}$. Additionally, we are given some function f of the members in $\{x_i\}$ to which we know the statistical expectation value $\langle f(x_i)\rangle$.

$$\big\langle f(x_i)\big\rangle = \sum_{i=1}^{n} P(x_i)\, f(x_i) \tag{63}$$

Clearly, this restricts the allowed probability distributions $\{P(x_i)\}$, but even including the normalization condition $\sum_i P(x_i) = 1$ we are still lacking another $(n-2)$ independent conditions to be able to determine *all* probabilities $\{P(x_i)\}$. We can however justify that there exists one probability distribution that is the most appropriate given $\langle f(x_i)\rangle$, by borrowing ideas from Information Theory.

We can consider the *Shannon entropy* (section 3.5) as a measure of the amount of "uncertainty" there is in a probability distribution. Then we argue that for each condition that is put on $\{P(x_i)\}$ the entropy (or *uncertainty*) of the probability distribution should decrease. To guarantee that behaviour we ought to select the allowed probability distribution that maximizes the entropy after some condition is imposed.

Jaynes writes: "$[\cdots]$ in making inferences on the basis of partial information we must use that probability distribution which has maximum entropy subject to whatever is known. This is the only unbiased assignment that we can make;

to use any other would amount to an arbitrary assumption of information which by hypothesis we do not have."

To actually find the state that maximizes entropy under some constraints, we will use the *method of Lagrange multipliers*—as we demonstrate in the next section 4.2.

We should point out that the *principle of maximum entropy inference* is based on ideas that make more sense in the *subjective school of thought* (see section 3.4), where we consider probabilities as in some form derived from our state of knowledge. And thus Jaynes' argument receives some opposition for not being sufficiently well grounded. In the following section 4.1 we present a possible strategy for defining a mathematical notion of *biased assumptions*, and create a tentative link between degrees of belief and experimental statistics.

4.1 Pursuit of a formal argument

It may be possible to produce a more formal argument to justify the *principle of maximum entropy inference*. Note however that this argument is not completely conclusive, and further examination will be necessary; in particular, the next step would be to produce some concrete examples of the following discussion.

Consider a system whose state is modelled by a probability distribution over a finite set of mutually exclusive states $\{x_i\}$. Assume that we have some constraint, C, we know to be true for the system. For instance, we could imagine that we have computed the average energy, measured the temperature, or obtained some other expectation value as described in equation (63).

Given a single constraint C, we have a range of compatible probability distributions. Consider that instead of selecting the probability distribution $\{P(x_i)\}$ that maximizes the entropy, we select some probability distribution $\{P'(x_i)\}$ that does not maximize the entropy. We may then be able to argue that there exists some secondary constraint C' under which $\{P'(x_i)\}$ is the probability distribution with maximum entropy, while fulfilling *both* C and C'. Therefore $\{P'(x_i)\}$ cannot be an acceptable choice, since we do not in fact know the second constraint C' to be true. We say that C' represents a *biased assumption*.

Note that even though our true constraint C is generally in some sense *measurable*, the character of the *biased assumption* C' can indeed be very artificial, and for all practical purposes not measurable. Nevertheless, C' still represents some knowledge about the system that we in fact cannot attach any truth-value to. We are then able to argue that the only choice which remains is $\{P(x_i)\}$, the probability distribution that maximizes entropy.

We can then use this construction to relate the distribution $\{P(x_i)\}$ to the real-world statistics that is observed in experiments. If a system is modelled using the *principle of maximum entropy inference* and the constraint C does reproduce the expected statistics, we can conclude that our model is in fact the most accurate we can produce with the available information. If, however, the statistics are somehow *more* restricted than predicted by our model, we can conclude that there exists some further constraint on our system that we are unaware of. And in the opposite case, if the statistics are somehow *less* restricted, our initial constraint C must have been too restrictive.

4.2 Deriving the thermal quantum state

Here we will employ *the principle of maximum entropy inference* to motivate and derive the following *canonical thermal state* from Quantum Mechanics.

$$\hat{\rho} = \frac{\mathrm{e}^{-\beta \hat{H}}}{\mathfrak{Tr}\big[\,\mathrm{e}^{-\beta \hat{H}}\big]} \tag{64}$$

We define a *thermal state* for a quantum system with Hamiltonian $\hat{H}$, as one where we have no information except about the average energy $\langle E \rangle$ of the state.

$$\langle E \rangle = \mathfrak{Tr}\big[\hat{\rho}\hat{H}\,\big] \tag{65}$$

Note that an average energy does not by itself imply some *unique* state.[17] Thus to be able to single out one particular choice of $\hat{\rho}$, we include the additional condition from the *principle of maximum entropy inference*—that $\hat{\rho}$ maximizes the *von Neumann entropy*, $S_N(\hat{\rho})$.

$$S_N(\hat{\rho}) := -\mathfrak{Tr}\big[\hat{\rho}\ln\hat{\rho}\big] \tag{66}$$

Here we have clearly generalized our initial discussion in terms of classical probabilities and *Shannon entropy*, to the quantum case using *von Neumann entropy* (as Jaynes also does, see [8]), but the preceding conceptual discussion remains valid.

Thus, our objective is to determine the state $\hat{\rho}$, that maximizes von Neumann entropy, subject to the constraints $\mathfrak{Tr}[\hat{\rho}] = 1$ and $\mathfrak{Tr}[\hat{H}\hat{\rho}] = \langle E \rangle$. To this end, we introduce two Lagrange multipliers, $\lambda_1 \neq 0$, $\lambda_2 \neq 0$, and we construct the function $f(\hat{\rho}, \lambda_1, \lambda_2)$.

$$f(\hat{\rho}, \lambda_1, \lambda_2) = -\,\mathfrak{Tr}\big[\hat{\rho}\ln\hat{\rho}\big] + \lambda_1\Big(1 - \mathfrak{Tr}\big[\hat{\rho}\big]\Big) + \lambda_2\Big(\langle E\rangle - \mathfrak{Tr}\big[\hat{\rho}\hat{H}\,\big]\Big) \tag{67}$$

We begin by using the spectral theorem to rewrite our state $\hat{\rho}$, and the energy operator $\hat{H}$ on their diagonal form (see section 2), using some diagonalizing basis $\{|\psi_i\rangle\}$, and $\{|\phi_i\rangle\}$, with their eigenvalues $\{p_i\}$ and $\{E_i\}$, respectively.

$$\hat{\rho} = \sum_i p_i\,|\phi_i\rangle\langle\phi_i| \tag{68}$$

$$\hat{H} = \sum_i E_i\,|\psi_i\rangle\langle\psi_i| \tag{69}$$

Then $\mathfrak{Tr}[\hat{\rho}\ln\hat{\rho}]$ becomes $\sum_i p_i \ln p_i$, the trace over $\hat{\rho}$ is simply $\sum_i p_i$, and finding the expression for $\mathfrak{Tr}[\hat{\rho}\hat{H}]$ requires few steps.[18]

$$\begin{gathered} f(\hat{\rho}, \lambda_1, \lambda_2) = \\ = -\sum_i p_i \ln p_i + \lambda_1\left(1 - \sum_i p_i\right) + \lambda_2\left(\langle E\rangle - \sum_{i,j} p_i\,E_j\,|\langle\psi_i|\phi_j\rangle|^2\right) \end{gathered} \tag{70}$$

[17] There is plenty of freedom in our choice of $\hat{\rho}$. For instance, we can consider states corresponding to different probability distribution between the eigenvalues of the Hamiltonian $\hat{H}$, $\{P(E_i)\}$, and we can introduce entanglement in our state, without changing the distribution between eigenvalues $\{P(E_i)\}$.

[18] Left to the reader as an exercise.

We shall then take the derivatives of this expression with respect to λ_1, λ_2, and every degree of freedom in $\hat{\rho}$, then set all derivatives to zero. Since the term that includes $|\langle\psi_i|\phi_j\rangle|^2$ is the only one affected by infinitesimal *rotations* of the state-basis $\{|\psi_i\rangle\}$, if the derivative shall equal zero, the matrix elements $\langle\psi_i|\phi_j\rangle$ must be indifferent (in the linear, first-order sense) to such a rotation. This is only possible if the bases $\{|\psi_i\rangle\}$, and $\{|\phi_i\rangle\}$ coincide (up to an irrelevant phase for each basis vector). Thus $|\langle\psi_i|\phi_j\rangle|^2 = \delta_{ij}$, and we can conclude that $\hat{\rho}$ is a diagonal matrix in the eigenbasis of the Hamiltonian, $\{|\phi_i\rangle\}$.

$$\begin{aligned} &f(p_1, \cdots, p_n, \lambda_1, \lambda_2) = \\ &= -\sum_i p_i \ln p_i + \lambda_1 \Big(1 - \sum_i p_i\Big) + \lambda_2 \Big(\langle E\rangle - \sum_i p_i E_i\Big) \end{aligned} \tag{71}$$

The remaining degrees of freedom of f are the ones with respect to the probability eigenvalues $\{p_i\}$, and of course λ_1 and λ_2.

$$\frac{\partial f}{\partial p_i} = -\big(\ln p_i + 1\big) - \lambda_1 - \lambda_2 E_i = 0 \qquad \Rightarrow \tag{72}$$

$$\ln p_i = -1 - \lambda_1 - \lambda_2 E_i \qquad \Rightarrow \tag{73}$$

$$p_i = \frac{\mathrm{e}^{-\lambda_2 E_i}}{\mathrm{e}^{1+\lambda_1}} \quad ; \quad \forall\, i \tag{74}$$

Together with the trivial derivatives with respect to λ_1 and λ_2, we gather all results in a system of equations.

$$\begin{cases} p_i = \dfrac{\mathrm{e}^{-\lambda_2 E_i}}{\mathrm{e}^{1+\lambda_1}} \quad ; \quad \forall\, i & (75) \\ \displaystyle\sum_i p_i = 1 & (76) \\ \displaystyle\sum_i p_i E_i = \langle E\rangle & (77) \end{cases}$$

Since we know that the basis $\{|\phi_i\rangle\}$ diagonalizes both $\hat{\rho}$ and $\hat{H}$ we can rewrite these equations in terms of $\hat{\rho}$ and $\hat{H}$.

$$\begin{cases} \hat{\rho} = \dfrac{\mathrm{e}^{-\lambda_2 \hat{H}}}{\mathrm{e}^{1+\lambda_1}} & (78) \\ \mathfrak{Tr}\big[\hat{\rho}\big] = 1 & (79) \\ \mathfrak{Tr}\big[\hat{\rho}\hat{H}\big] = \langle E\rangle & (80) \end{cases}$$

In equation (78) we have two unknowns, λ_1 and λ_2, that are uniquely specified by the two conditions in equation (79) and (80). To find an expression for the denominator $\mathrm{e}^{1+\lambda_1}$, we take the trace of equation (78) and set it equal to one.

$$1 = \frac{1}{\mathrm{e}^{1+\lambda_1}} \mathfrak{Tr}\big[\,\mathrm{e}^{-\lambda_2 \hat{H}}\big] \qquad \Rightarrow \qquad \frac{1}{\mathrm{e}^{1+\lambda_1}} = \frac{1}{\mathfrak{Tr}\big[\,\mathrm{e}^{-\lambda_2 \hat{H}}\big]} =: \frac{1}{Z} \tag{81}$$

Thus, the denominator $\mathrm{e}^{1+\lambda_1} \in (0, \infty)$ in equation (78) can be seen as a normalization of the numerator. We name this normalization *the partition function*,

and denote it as $Z := \mathfrak{Tr}\big[\,\mathrm{e}^{-\lambda_2 \hat{H}}\big]$.

$$\begin{cases} \hat{\rho} &= \dfrac{\mathrm{e}^{-\lambda_2 \hat{H}}}{\mathfrak{Tr}\big[\,\mathrm{e}^{-\lambda_2 \hat{H}}\big]} & (82) \\ \mathfrak{Tr}\big[\hat{\rho}\hat{H}\big] = \langle E \rangle & & (83) \end{cases}$$

Then λ_2 is a constant that is fixed by $\langle E \rangle$, and a dimensional analysis gives that it has the physical dimensions of the standard reciprocal temperature $\beta(\langle E \rangle) = 1/kT(\langle E \rangle)$. Thus we can call $\lambda_2 \equiv \beta$, and we arrive at our thermal state for $\hat{\rho}$.

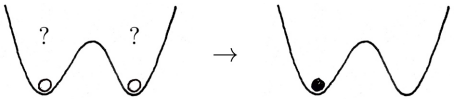
$$\hat{\rho} = \frac{\mathrm{e}^{-\beta \hat{H}}}{\mathfrak{Tr}\big[\,\mathrm{e}^{-\beta \hat{H}}\big]} \tag{84}$$

We conclude that the thermal state is a diagonal matrix in the eigenbasis of the Hamiltonian, where the probability eigenvalues $\{p_i\}$ correspond to the classical canonical distribution (when treated as a function of their energies, $p_i(E_i)$).

Finally, we should mention one technicality. The method of Lagrangian multipliers can only find *candidate points* for extremum values. In this case, the $\hat{\rho}$ we found really is a *maximum* of the von Neumann entropy, however a rigorous proof of this will be arduous, and we shall here refrain from that.

5 Landauer's principle in Classical Physics

In 1961, Rolf Landauer published the paper "Irreversibility and Heat Generation in the Computing Process" [1]. It opens with a discussion about a particle at rest in a double well potential. Imagine that your task is to apply forces to the particle, such as to move the particle to the left potential minima. However, you are not privy to the information about whether the particle resides to the left or right to begin with.



Under classical physics, this is an impossible task with forces alone[19], and your only option is to let some kinetic energy dissipate as heat to a surrounding reservoir, making this process *physically irreversible*. Landauer then presents a number of other examples and argues that this is a general principle. That "logical irreversibility[20] is associated with physical irreversibility and requires a minimal heat generation"—*Landauer's principle*.

But it is problematic to provide specific examples to prove a universal claim. In this section, we will look closer at what Landauer argued, and construct a general argument to demonstrate the result that he predicted, but in a more universal setting.

5.1 Introduction

Even though Landauer published his seminal paper in 1961, until today papers are continuously being published on *Landauer's principle*. Despite this, the presently available literature can be quite opaque, and occasionally conflicting to the point that some authors appear to be in direct opposition to each other—see for instance Maroney [13] vs. Ladyman et al. [14], Bennett's discussion [15], or Sagawa [16], who claims that "the logically irreversible erasure can be performed in a thermodynamically reversible manner in the quasi-static limit", seemingly contradictory to Landauer's original conclusion [1].

There are surely several explanations for the ambiguous situation. We can note that there exists no consensual mathematical framework, and many authors choose to discuss the ideas from the point of view of some particular physical system using boxes, pistons, point particles, constrained Hamiltonians, constrained probability distribution, etc., before claiming that the result should apply in general. See for instance the following papers, [1,14–19]. In this context, it may be difficult to pinpoint where the disagreement occurs, when conflicting attitudes appears.

[19]This is a consequence of Liouville's theorem for Hamiltonian Mechanics, stating that for any physical system obeying Hamiltonian Mechanics, the divergence of the velocity field in its phase is zero, if only conservative forces are allowed. The same argument is put forth by Landauer [1], but expressed in somewhat different words.

[20]Simply stated, the described process is considered *logically irreversible* since we cannot determine the initial state from the final state.

Here we intend to construct an argument in favour of Landauer's original conclusion, without making any assumptions about the physical systems involved. This type of argument is believed to be useful for exposing some causes of disagreements and bringing the literature closer to consensus. We also intend for this argument to be suitable as an introductory remark for the reader unfamiliar with *Landauer's principle*. In order to tailor the approach to this end, the following two notions will be guiding the discussion.

Notion I (A general physical system). In order to keep the discussion relatable to the majority of other publications we will, as mentioned, consider a physical system on general terms, and then conduct an argument contingent on *the second law of thermodynamics*. □

Notion II (Following the original set-up). There is a tradeoff between how general we will be in our initial assumptions, and how much effort is required to follow the argument. Since literature already exists with the purpose to generalize Landauer's principle beyond its original scope—here, we will essentially stick to the original set-up. We shall point out to the reader when we make some needless limitations and cite references to the appropriate generalized analysis. Thus, the treatment in this section is meant as a minimal kernel, to which further arguments (such as the one in section 6) can be attached. □

Also, we will base the argument on *semi-classical* physics in the following sense. We assume classical physics—just like Landauer originally did—but with the additional assumption that it is possible to find a *finite* set of mutually exclusive microstates that is sufficient to model our system. This assumption is based on considerations from Quantum Mechanics where a system can be described as a superposition of a countable set of pure states, typically eigenstates of some Hamiltonian [10].[21] This model aligns with Orwen J. E. Maroney's paper from 2009 [20].

In the course of our discussion, we will expose a complication (section 5.6), which is believed to never have been previously stated in an explicit manner, likely because it is conveniently hidden in most considerations based on specific systems. We then present a solution to the complication (section 5.7), which is also briefly mentioned by Maroney [20], but not as a solution to an identified problem.

[21] Two remarks are in order. First, if we are forced to use a continuous phase space for the system, we can view the countability as originating from a coarse-graining procedure. Second, if the state space is infinite, we can assume that there exists some cut-off energy $E_{\max}$, for which the set of states with lower energy is finite, and where the probability of occupation for any state energy higher than $E_{\max}$ is so low that truncating our state space will be an acceptable approximation.

5.2 Stating Landauer's principle

There are several formulations of *Landauer's principle* in the literature. Colloquially, it can be expressed as "erasure of one bit of information requires a net increase in total entropy of at least $k \ln 2$, or in terms of heat, $kT \ln 2$". But as we shall discuss in section 7, there is a danger that such colloquial phrasing seduces us to apply the principle where it does not belong. Therefore, we shall consider the following, more well defined, proposition.

Proposition 5.1 (Landauer's principle). Any logically *irreversible* and *deterministic* physical manipulation of information that is *entropy-restoring* (as defined in section 5.7), and designed to *decrease* the Shannon entropy of the encoded information, must be accompanied by an increase in the *Gibbs entropy* ΔS (section 3.6) in the total closed system.

$$\langle \Delta S \rangle \geq -k \ln 2 \, \Delta H \qquad \text{where} \qquad \Delta H < 0 \tag{85}$$

If the system can be described as having a thermal reservoir of temperature T, the increase of Gibbs entropy can be realized as an increase of heat ΔQ.

$$\langle \Delta Q \rangle \geq -k T \ln 2 \, \Delta H \qquad \text{where} \qquad \Delta H < 0 \tag{86}$$

This increase in entropy makes this kind of information manipulation a *physically irreversible* process.

We have several remarks to consider about this proposition.

Remark I. Some authors define ΔH with a positive sign when the entropy of the information is *decreasing*. But in this thesis, we will be consistently following the convention that increasing quantities have positive signs, and decreasing quantities have negative signs. □

Remark II. Note that ΔH is evaluated as *Shannon entropy*, in units of *bits* (see equation (12), section 3.5). □

Remark III. Landauer's principle, as stated in proposition 5.1, is discussed in terms of classical systems, and therefore only applies to classical physics. □

Remark IV. We have limited proposition 5.1 to the case of *logically deterministic* manipulations of data. This condition is relaxed by Maroney [20], where also random processes are considered. Note that this limitation is quite tolerable, since most computation devices deal with processes that are logically deterministic. □

Remark V. Demanding that the process is *entropy-restoring* is a requirement not found in the literature cited in this thesis. □

Remark VI. We label any process *physically irreversible* if the total entropy of a closed system is increased as a consequence of the process, since according to the second law of Thermodynamics, you cannot reverse such a process. More comprehensive discussions are carried out for instance by Ladyman et al. [14], and Sagawa [16]. □

Strictly speaking, we will not provide a *proof* of Landauer's principle (proposition 5.1), instead we provide a phenomenological argument for it. In section 6, we will be looking at quantum systems for which more formal rigour can be afforded. Also, to keep the argument more minimal, we will limit the discussion to the case $\Delta H = -1$ bit, and discuss extensions later (section 5.11).

5.3 Defining quantities and concepts

5.3.1 Logical states

We begin by establishing how to represent information. Following the original paper by Landauer [1], we restrict ourselves to information in a binary representation, thus any unit of information takes values from an *alphabet* set with two members, $\{\mathtt{0},\mathtt{1}\}$. For a more general treatment with alphabet sets of arbitrary (finite) size, the reader is referred to Owen Maroney's paper from 2009, [20].

Consider a variable $\mathtt{x}$ that takes values from the alphabet set $\{\mathtt{0},\mathtt{1}\}$ with some probability for each. The probabilities for $\mathtt{x}$ to assume either of the values $\{\mathtt{0},\mathtt{1}\}$ are denoted as $P(\mathtt{x}{=}\mathtt{0})$ and $P(\mathtt{x}{=}\mathtt{1})$. Taken together, these form a "logical state". [22]

$$\big(\mathtt{x} \in \{\mathtt{0},\mathtt{1}\}, \{P(\mathtt{x}{=}\mathtt{0}), P(\mathtt{x}{=}\mathtt{1})\}\big) \tag{87}$$

This notation somewhat bulky, so let us be precise about the meaning of some abbreviated notation, which will be frequently occurring.

$$\big(\mathtt{x}, \{P(\mathtt{0}), P(\mathtt{1})\}\big) \quad \Leftrightarrow \quad \big(\mathtt{x} \in \{\mathtt{0},\mathtt{1}\}, \{P(\mathtt{x}{=}\mathtt{0}), P(\mathtt{x}{=}\mathtt{1})\}\big) \tag{88}$$

$$\mathtt{x} = \mathtt{0} \quad \Leftrightarrow \quad \big(\mathtt{x}, \{P(\mathtt{0}) = 1, P(\mathtt{1}) = 0\}\big) \tag{89}$$

$$\mathtt{x} = \mathtt{1} \quad \Leftrightarrow \quad \big(\mathtt{x}, \{P(\mathtt{0}) = 0, P(\mathtt{1}) = 1\}\big) \tag{90}$$

$$\text{[proposition including x]} \ \forall\, \mathtt{x} \in \{\mathtt{0},\mathtt{1}\} \quad \Leftrightarrow \tag{91}$$

The proposition holds when $\mathtt{x}$ is replaced by any member of $\{\mathtt{0},\mathtt{1}\}$.

5.3.2 Logical operations

To a *logical state* $(\mathtt{x}, \{P(\mathtt{0}), P(\mathtt{1})\})$ we can apply "logical operations". Here we will limit ourselves to *deterministic* logical operations, in the sense that any initial distinct logical state, $\mathtt{x} = \mathtt{0}$ or $\mathtt{x} = \mathtt{1}$, uniquely determines a distinct result, $\mathtt{x} = \mathtt{0}$ or $\mathtt{x} = \mathtt{1}$. Under these restrictions, we can describe any logical operation with a truth table, and for an *alphabet* set with two members there will only be four possible logical operations. For discussions also including operations that are non-deterministic, or *random*, see [20].

[22] The name *logical state* is chosen since it is in agreement with the majority of publications. See for instance [14, 20].

Identity operation

Input	Output
0	0
1	1

Not operation

Input	Output
0	1
1	0

Reset-to-zero operation

Input	Output
0	0
1	0

Reset-to-one operation

Input	Output
0	1
1	1

When "erasure of information" is mentioned in the context of Landauer's principle, one is referring to the behaviour of the last two operations; where the logical output state is certain, irrespective of the initial logical state, and thus entropy—or *information*—is decreased when comparing input to output.

Note that we consider *logical operations* which takes only *one* logical state as input, and produces *one* logical state as output. Maroney considers situations not subject to this limitation, [20].

5.3.3 The information-bearing system, $\mathcal{S}$

We then select some appropriate *free parameter* of a physical system to encode the *logical state*. With "parameter" we mean some *measurable* property, typically a single degree of freedom of the system; such as the energy of an atom, the current through a wire, or the position of a particle, but it can also constitute some partition in a *multidimensional* phase space. This parameter is the assumed to be "free" in the sense that is independent of any other degrees of freedom. Thus the free parameter can be treated as a physical system in itself, separate from its environment, and we will hereafter refer to it as $\mathcal{S}$, or the *information-bearing system*.

5.3.4 Physically encoding logical states in $\mathcal{S}$

As mentioned in the introduction (section 5.1), $\mathcal{S}$ should be such that it can be fully described by a finite set of *mutually exclusive microstates*, $\{\mu\}$. We then select some subset of microstates, $\{\mu_0\} \in \{\mu\}$, as encoding for the *logical state* $\mathtt{x} = \mathtt{0}$ and some other subset, $\{\mu_1\} \in \{\mu\}$, to encode for $\mathtt{x} = \mathtt{1}$. These subsets are chosen to be *non-intersecting* and *collectively exhaustive* in the following sense.

$$\{\mu_0\} \cap \{\mu_1\} = \varnothing \quad ; \quad \{\mu_0\} \cup \{\mu_1\} = \{\mu\} \tag{92}$$

Given some method of measuring whether the *information-bearing system* $\mathcal{S}$ has a microstate in $\{\mu_0\}$ or $\{\mu_1\}$, this construct ensures that a distinct logical state, $\mathtt{x} = \mathtt{0}$ or $\mathtt{x} = \mathtt{1}$, is always well defined if we make a measurement on $\mathcal{S}$.

However in the most general situation—where we typically lack full information about the state of the system[23]—we would describe any state of $\mathcal{S}$ as

[23] In section 4 we discuss the implication of lacking information, related to the *principle of maximum entropy inference*.

a probability distribution over the complete set of microstates $\{\mu\}$, and since the subsets encoding for $\mathtt{x} = \mathtt{0}$ and $\mathtt{x} = \mathtt{1}$ are non-intersecting and collectively exhaustive, according to equation (92), we can take sums over the sets $\{\mu_0\}$ and $\{\mu_1\}$ to find the probabilities for each distinct logical state, $P(\mathtt{0})$ and $P(\mathtt{1})$.

$$P(\mathtt{0}) = \sum_{\{\mu_0\}} P\big(\mu_0\big) \qquad ; \qquad P(\mathtt{1}) = \sum_{\{\mu_1\}} P\big(\mu_1\big) \tag{93}$$

Under the conditions of equation (92), and since the probability distribution over $\{\mu\}$ is normalized to unity, the same holds for $P(\mathtt{0})$ and $P(\mathtt{1})$.

$$P(\mathtt{1}) + P(\mathtt{0}) = 1 \tag{94}$$

5.3.5 The reservoir, $\mathcal{R}$

So far we have introduced the *information-bearing system* $\mathcal{S}$, and its finite set of accessible microstates $\{\mu\}$. However, we shall not consider $\mathcal{S}$ in complete isolation. In addition to the *information-bearing system*, there will be a large number of secondary degrees of freedom, termed the *reservoir*, and denoted by $\mathcal{R}$. This reservoir is assumed to be large enough in order to remain at constant temperature even if small quantities of energy or entropy is exchanged with $\mathcal{S}$.

5.3.6 The closed system, $\mathcal{C}$

Taken together, $\mathcal{S}$ and $\mathcal{R}$ form a closed system, denoted by $\mathcal{C}$, and referred to, simply, as the *closed system*. Note that *the second law of thermodynamics* (axiom 3.4) applies to $\mathcal{C}$.

5.3.7 Logical processes, $\mathcal{P}_\mathcal{S}$

The *free* status of the *information-bearing system* $\mathcal{S}$ can then be momentarily suspended, when a physical interaction between $\mathcal{S}$ and $\mathcal{R}$ is introduced. The purpose of such interactions is to execute *logical operations* (see section 5.3.2) on the *logical state* of $\mathcal{S}$. We will call this type of physical interaction a "logical process" and we label it $\mathcal{P}_\mathcal{S}$—where the subscript is a reminder that the purpose of $\mathcal{P}_\mathcal{S}$ is to manipulate the logical state of $\mathcal{S}$.

Note that the term *process* is here reserved for the physical domain, and the term *operation* refers to the logical domain (see section 5.3.2).

5.3.8 Logical reset processes, $\mathcal{P}_\mathcal{S}^\mathtt{0}$

Landauer's principle concerns the consequences of a particular kind of logical process—one which executes a *reset operation*; by setting the logical state to some predetermined *standard state*, regardless of the initial logical state. Here we let the standard state to be $\mathtt{x} = \mathtt{0}$, thus the logical operation we consider is *reset-to-zero* (see tables in section 5.3.2), but the argument is of course invariant under this choice. We will call any such type of physical process a "logical reset process", and we denote it with $\mathcal{P}_\mathcal{S}^\mathtt{0}$.

5.3.9 Terminology of other authors

Several authors, e.g. [14, 15, 20], use the terms "information-bearing degrees of freedom" or IBDF, and "non-information-bearing degrees of freedom" or NIBDF, to refer to $\mathcal{S}$ and $\mathcal{R}$ respectively. Here, partially in order to ease our subsequent transition into fully quantum-mechanical systems, in section 6, we will favour the terminology in terms of $\mathcal{S}$ and $\mathcal{R}$.

5.4 Example of an information-bearing system $\mathcal{S}$

Before entering into the abstract discussion of a general physical system, it may be comforting to have some concrete ideas to relate to the abstract.

Perhaps the most minimal example of an *information-bearing system* $\mathcal{S}$, is a classical particle in a one-dimensional potential well of infinite height. This example is also discussed by Landauer in his original paper [1]. Our intention is to use the position of the particle to encode our logical state $(\mathtt{x}, P(\mathtt{0}), P(\mathtt{1}))$. We label the position of the particle q, and let the infinite well begin and end at $\pm L$.

Then, divide $q \in [-L, L]$ into tiny intervals to create our set of mutually exclusive microstates $\{\mu\}$.[24] For each state in $\{\mu\}$, if the tiny interval is to the left we will put it into $\{\mu_0\}$, and associate it with the logical state $\mathtt{x} = \mathtt{0}$. Vice versa, if the tiny interval is to the right it will be a member of $\{\mu_1\}$ associated with $\mathtt{x} = \mathtt{1}$.

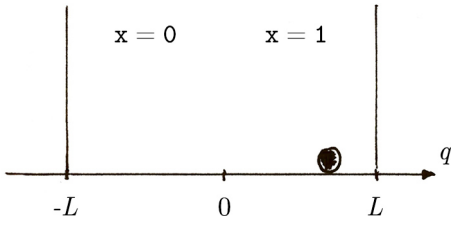


Figure 3: Encoding of a *logical state* in a *information-bearing system* $\mathcal{S}$—a classical particle in a one-dimensional potential well. The interval $q \in [-L, 0]$ is divided into a set of microstates $\{\mu_0\}$, corresponding to the logical state $\mathtt{x} = \mathtt{0}$, and the interval $q \in [0, L]$ is divided into the microstates $\{\mu_1\}$, corresponding to the logical state $\mathtt{x} = \mathtt{1}$. Spontaneous (Brownian) motion is assumed to be negligible.

In this construction, a *logical process* $\mathcal{P}_\mathcal{S}$ can for instance be created by applying forces (which give energy to $\mathcal{S}$) and friction (which dissipate energy into $\mathcal{R}$) to the particle, and thus we can manipulate the logical state of $\mathcal{S}$. Note that spontaneous (Brownian) motion can be assumed to be negligible.

[24] One may argue that space is not quantized in this manner, in which case we can consider this an example using *coarse-grained entropy*. See [5] for further discussions about the coarse-graining of phase spaces.

5.5 Additive entropies

We now intend to construct an argument for *Landauer's principle*. We want to argue for equation (85) in proposition 5.1, i.e. the more general case where the increase in entropy is not *necessarily* in terms of heat. In equation (85) the average change in entropy $\langle \Delta S \rangle$, refers to the entire closed system $\mathcal{C}$, or $\langle \Delta S_{\mathcal{C}} \rangle$. Since entropy is additive for independent states (see section 3.5.4), we can consider the entropies of each subsystem ($\mathcal{S}$ and $\mathcal{R}$) separately, and add them up.

$$S_{\mathcal{C}} = S_{\mathcal{S}} + S_{\mathcal{R}} \qquad \Rightarrow \qquad \Delta S_{\mathcal{C}} = \Delta S_{\mathcal{S}} + \Delta S_{\mathcal{R}} \tag{95}$$

What we are actually interested in, is the *change* in entropy. Therefore we will only define the entropy of each subsystem ($S_{\mathcal{S}}$ and $S_{\mathcal{R}}$) up to an additive constant.

5.6 Complication from undefined entropies

We begin looking for an expression of $S_{\mathcal{S}}$—since we have some idea what happens to the state of this system when the *logical reset process* $\mathcal{P}^{\mathtt{0}}_{\mathcal{S}}$ (as defined in section 5.3.8) is applied. The entropy in $\mathcal{S}$ is simply the entropy over the microstates $\{\mu\}$.

$$S_{\mathcal{S}} = S\big(\{P(\mu)\}\big) \tag{96}$$

Perhaps we can argue that we should be given some initial probability distribution over all microstates $\{P(\mu)\}$, in which case we can calculate the initial values of $S_{\mathcal{S}}$, before $\mathcal{P}^{\mathtt{0}}_{\mathcal{S}}$ has been applied.

We can then apply *the composition law*—axiom 3.3 in section 3.5.3—to separate the entropy *of* the logical state, from the entropy *within* each of the two logical states, as seen in the following relation reproduced by Maroney [20].[25]

$$S_{\mathcal{S}} = S\big(P(\mathtt{0}), P(\mathtt{1})\big) + P(\mathtt{0})\, S\big(\{P(\mu_{\mathtt{0}})\}\big|\mathtt{x}{=}\mathtt{0}\big) + P(\mathtt{1})\, S\big(\{P(\mu_{\mathtt{1}})\}\big|\mathtt{x}{=}\mathtt{1}\big) \tag{97}$$

Here we have employed conditional entropies (see section A.9) to calculate the remaining entropy when a logical state is determined. These entropies will appear frequently, thus it is appropriate to define a more compact notation.

Definition 5.1 (Remaining entropy). The *remaining entropy* when some particular logical state has been determined, $\mathtt{x} = \mathtt{0}$ or $\mathtt{x} = \mathtt{1}$, is calculated from *conditional entropy* (section A.9).

$$S\big(\{P(\mu_{\mathtt{x}})\}\big|\mathtt{x}\big) \tag{98}$$

For any logical state with $P(\mathtt{x}) > 0$, this corresponds to normalizing the probability distribution over $\{\mu_{\mathtt{x}}\}$ such that $\{P(\mu_{\mathtt{x}})\}$ sums to 1, and then calculating the entropy. We define a shorthand notation for this.

$$S(\mu_{\mathtt{0}}) := S\big(\{P(\mu_{\mathtt{0}})\}\big|\mathtt{x}{=}\mathtt{0}\big) \quad ; \quad S(\mu_{\mathtt{1}}) := S\big(\{P(\mu_{\mathtt{1}})\}\big|\mathtt{x}{=}\mathtt{1}\big) \tag{99}$$

$$S(\mu_{\mathtt{x}}) := S\big(\{P(\mu_{\mathtt{x}})\}\big|\mathtt{x}\big) \tag{100}$$

[25] See equation (107) on page 14, where the corresponding formula is expressed as an entropy *difference*, and on a form to account for a larger *alphabet* set.

With this notation we can rewrite equation (97).

$$S_{\mathcal{S}} = S\big(P(\mathtt{0}), P(\mathtt{1})\big) + P(\mathtt{0})\, S(\mu_{\mathtt{0}}) + P(\mathtt{1})\, S(\mu_{\mathtt{1}}) \tag{101}$$

After $\mathcal{P}_{\mathcal{S}}^{\mathtt{0}}$ is applied, the probabilities for the logical state $P(\mathtt{0})$ and $P(\mathtt{1})$ are given, making the first term in equation (101) well defined. Since the *logical operation* is *reset-to-zero* we know the probabilities $P(\mathtt{0}) = 1$ and $P(\mathtt{1}) = 0$, which also sets the last term to zero. However, the entropy over $\{\mu_{\mathtt{0}}\}$ in the second term of equation (101) is not well defined. This is because there are many conceivable *logical processes* that recreate the *large-scale* behaviour of *reset-to-zero*, but are not physically equivalent because they create different probability distributions over the *small-scale* microstates $\{\mu_{\mathtt{x}}\}$ *within* a logical state. Put differently, the *remaining entropy* (from definition 5.1) is not determined.

We can therefore conclude that our assumptions so far are not sufficiently restrictive to relate the entropy of the *physical state*, to the entropy of the information in our *logical state*.

5.7 Logical processes $\mathcal{P}_{\mathcal{S}}$ must be entropy-restoring

There are of course different attitudes one can take to the *complication of undefined entropies*, described in the previous section 5.6.

Approach I. We could demand that we use an *information-bearing system* which only has two mutually exclusive microstates $|\{\mu\}| = 2$, one for each logical state. Then there would be no further *small-scale* structure to account for, i.e the only entropy that exists is in the *logical states*. Clearly, this has the considerable disadvantage that Landauer's principle would not apply to most physical systems currently used for information processing. □

Approach II. We could demand that, even though we allow numerous microstates, any *logical process* $\mathcal{P}_{\mathcal{S}}$ can only map onto a *single* microstate in each of the sets for the logical states, $\{\mu_{\mathtt{0}}\}$ and $\{\mu_{\mathtt{1}}\}$. Then the entropy from $\{P(\mu_{\mathtt{0}})\}$ and $\{P(\mu_{\mathtt{1}})\}$ would be zero after $\mathcal{P}_{\mathcal{S}}$. However, this is still a rather limiting condition which excludes a great number of computation devices. □

We can take the idea from approach II, but relax it further. In fact, it is not necessary to demand *remaining entropy* $S(\mu_{\mathtt{x}})$ to be zero, as long as the entropy assumes some *constant value* on *average*, we will be fine. Thus we define an "entropy-restoring" logical process, loosely stated, as having some fixed *spread* (on average) for the probability distribution over each of the sets of microstates, $\{\mu_{\mathtt{0}}\}$ and $\{\mu_{\mathtt{1}}\}$. The following definition 5.2 expresses this idea formally.

Definition 5.2 (Entropy-restoring process). After some *physical process* $\mathcal{P}_{\mathcal{S}}$ is applied to an information-bearing system $\mathcal{S}$. If the *remaining entropy*, $S(\mu_{\mathtt{x}})$ (see definition 5.1), is distributed around some process-specific average value S_{RES}, for each logical state ($\mathtt{x} = \mathtt{0}$ and $\mathtt{x} = \mathtt{1}$), then this process $\mathcal{P}_{\mathcal{S}}$ is defined as *entropy-restoring*.

$$\big\langle S(\mu_{\mathtt{x}}) \big\rangle = S_{\mathit{RES}} \qquad \forall\, \mathtt{x} \in \{\mathtt{0}, \mathtt{1}\} \text{ where } P(\mathtt{x}) > 0 \tag{102}$$

after $\mathcal{P}_{\mathcal{S}}$ is applied

Remark I. In this thesis we have used an alphabet set with two members, $\{\mathtt{0}, \mathtt{1}\}$, but this definition will of course extend to alphabet sets of any (finite) cardinality. □

Remark II. We will leave some ambiguity about what the average is taken over. But, for example, the average can be produced over an ensemble of several systems, or from a time series. But at its most general, the average can be viewed as a gentle reminder that any one particular realization is not bound to satisfy these relations. □

Definition 5.2 is a reasonable condition to put on our logical processes since any *non-entropy-restoring* process, for which the average *remaining entropy* $\langle S(\mu_{\mathtt{x}})\rangle$ can stray to any value, is practically problematic for a number of reasons. For instance, if the entropy is ever increasing, we may not be able to guarantee that some particular microstates, say in $\{\mu_{\mathtt{0}}\}$, which may have some coupling to microstates of $\{\mu_{\mathtt{1}}\}$ can be kept at a low probability of being populated, thus introducing errors. Secondly, since the entropy over any finite set of events is bounded both from above and below, entropies in some ensemble must have a very exotic behaviour if we want it to avoid converging on some average value as the ensemble grows.

However, examining equation (101), we can see that there is a way to further relax the conditions, briefly discussed by Maroney [20] as *uniform computing*.

Definition 5.3 (Uniform computing). After some *physical process* $\mathcal{P}_{\mathcal{S}}$ is applied to an information-bearing system $\mathcal{S}$. If the entropy over the individual logical states, weighted by their probabilities, is a process specific constant, S_{UNI}, then the process is said to conform to *uniform computing*. With an alphabet set of two members, $\{\mathtt{0}, \mathtt{1}\}$, the condition takes the following form.

$$P(\mathtt{0})\, S(\mu_{\mathtt{0}}) + P(\mathtt{1})\, S(\mu_{\mathtt{1}}) = S_{UNI} \tag{103}$$

Here we will stick with our condition of *entropy-restoring processes*. First, it will be more versatile by considering averages. Second, the *reset process* we want to examine (section 5.3.8) always has $P(\mathtt{0}) = 1$ and $P(\mathtt{1}) = 0$, removing the added freedom in definition 5.3.

We note that is not entirely clear what it means from a physical point of view, to design a *logical process* that operates according to *uniform computing*—where the *remaining entropy* (definition 5.1) depends on the probabilities for logical states. This question is also mentioned in the conclusions section when discussing further work (section 8.1.2).

5.8 Calculating a Landauer bound

We can now return to where we got stuck in the calculation (section 5.6), but we now assume that our logical reset process $\mathcal{P}_{\mathcal{S}}^{\mathtt{0}}$ is *entropy-restoring* according to the definition 5.2.

We will further assume that the initial state of $\mathcal{S}$, is prepared by some process under the same condition of being *entropy-restoring*. This is reasonable since similar processes often sequentially act on information-bearing systems in computation devices.

5.8.1 Entropy in $\mathcal{S}$

With begin by taking an average of $S_{\mathcal{S}}$ in equation (101). Note that we consider some specific logical state, i.e. $\langle P(\mathtt{x})\rangle = P(\mathtt{x}) \;\; \forall\, \mathtt{x} \in \{\mathtt{0},\mathtt{1}\}$.

$$\langle S_{\mathcal{S}}\rangle = S\big(P(\mathtt{0}), P(\mathtt{1})\big) + P(\mathtt{0})\,\big\langle S(\mu_{\mathtt{0}})\big\rangle + P(\mathtt{1})\,\big\langle S(\mu_{\mathtt{1}})\big\rangle \tag{104}$$

Whether or not this is the entropy before or after $\mathcal{P}^{\mathtt{0}}_{\mathcal{S}}$, the state is assumed to prepared by an *entropy-restoring* process. Thus we know that $\langle S(\mu_{\mathtt{x}})\rangle = S_{RES} \;\; \forall\, \mathtt{x} \in \{\mathtt{0},\mathtt{1}\}$.

$$\langle S_{\mathcal{S}}\rangle = S\big(P(\mathtt{0}), P(\mathtt{1})\big) + \big(P(\mathtt{0}) + P(\mathtt{1})\big)\, S_{RES} \tag{105}$$

As we argued in section 5.3.4, equation (94), the probabilities over the alphabet set sum to 1.

$$\langle S_{\mathcal{S}}\rangle = S\big(P(\mathtt{0}), P(\mathtt{1})\big) + S_{RES} \tag{106}$$

As argued in section 5.5, when calculating entropy differences, we only need to define our entropies up to an additive constant, and since S_{RES} is a constant associated with any logical process we can redefine $\langle S_{\mathcal{S}}\rangle$ by subtracting S_{RES}.

$$\langle S_{\mathcal{S}}\rangle = S\big(P(\mathtt{0}), P(\mathtt{1})\big) \tag{107}$$

The probabilities $P(\mathtt{0})$ and $P(\mathtt{1})$, are the probabilities of the *logical states* we use to encode information. Thus it makes sense to express this entropy in terms of *bits*, used in *Shannon entropy* (section 3.5), and in the process, we decouple it from Boltzmann's constant k which is included in S.

$$\boxed{\langle S_{\mathcal{S}}\rangle = k\,\ln 2\; H\big(P(\mathtt{0}), P(\mathtt{1})\big)} \tag{108}$$

Thus, by imposing the arguably reasonable constraint that our logical processes are *entropy-restoring* (definition 5.2) we have been able to make the average entropy in $\mathcal{S}$ dependent only on the probabilities for the logical states, $P(\mathtt{x}=\mathtt{0})$ and $P(\mathtt{x}=\mathtt{1})$. This is exactly what we need if we want to express the change of the entropy in the physical system in terms of change of entropy in information.

5.8.2 Change of entropy in $\mathcal{S}$

We now want to design some physical *reset-to-zero* process $\mathcal{P}^{\mathtt{0}}_{\mathcal{S}}$, with the purpose to reset any *arbitrary* logical state to our *standard* logical state $\mathtt{x} = \mathtt{0}$.

But we need to be explicit about what an *arbitrary* logical state is. More precisely, we mean is that the process $\mathcal{P}^{\mathtt{0}}_{\mathcal{S}}$ has no access to information about the logical state it is going to reset. In this situation, *the principle of maximum entropy inference* (as discussed in section 4) asserts that we should model $\mathcal{S}$ with a state which maximizes the entropy of $\mathcal{S}$, under available constraints. Here, our only constraint is that the state is produced by an *entropy-restoring* process, and therefore the expression derived for $\langle S_{\mathcal{S}}\rangle$ in from equation (108) applies here. Let us say that the process $\mathcal{P}^{\mathtt{0}}_{\mathcal{S}}$ starts at some time t, thus initial entropy is evaluated at time t.

$$\big\langle S_{\mathcal{S}}(t)\big\rangle = k\,\ln 2\, H(t) = \max_{P(\mathtt{0}),P(\mathtt{1})} k\,\ln 2\, H\big(P(\mathtt{0}), P(\mathtt{1})\big) \tag{109}$$

The maximum value is assumed for equal probabilities of each logical state $P(\mathtt{0}) = P(\mathtt{1}) = 1/2$. (We calculate H separately to be able to compare the change of *Shannon entropy* in the information, to the change of *Gibbs entropy* in $\mathcal{S}$.)

$$H(t) = -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) = 1\,\text{bit} \qquad \Rightarrow \tag{110}$$

$$\left\langle S_{\mathcal{S}}(t)\right\rangle = k\,\ln 2 \tag{111}$$

Assuming that $\mathcal{P}_{\mathcal{S}}^{\mathtt{0}}$ requires the time $\Delta t > 0$ to execute, we express the final entropy as $S_{\mathcal{S}}(t+\Delta t)$. Since the purpose of $\mathcal{P}_{\mathcal{S}}^{\mathtt{0}}$ is to set the logical state to $\mathtt{x} = \mathtt{0}$, we know that $P(\mathtt{0}) = 1$ and $P(\mathtt{1}) = 0$.

$$H(t+\Delta t) = -(1\log_2 1 + 0\log_2 0) = 0\,\text{bits} \qquad \Rightarrow \tag{112}$$

$$\left\langle S_{\mathcal{S}}(t+\Delta t)\right\rangle = 0 \tag{113}$$

We then calculate ΔH (the change in *Shannon entropy*), along with $\Delta S_{\mathcal{S}}$ (the change of entropy in $\mathcal{S}$), when a reset-to-zero process $\mathcal{P}_{\mathcal{S}}^{\mathtt{0}}$ (with no prior information about the initial logical state) has executed.

$$\boxed{\Delta H = H(t+\Delta t) - H(t) = -1\,\text{bit}} \tag{114}$$

$$\boxed{\langle \Delta S_{\mathcal{S}}\rangle = \left\langle S_{\mathcal{S}}(t+\Delta t)\right\rangle - \left\langle S_{\mathcal{S}}(t)\right\rangle = -k\,\ln 2} \tag{115}$$

5.8.3 The Landauer bound, change of entropy in $\mathcal{R}$

Since entropy is additive (see section 5.5), the total change in entropy when $\mathcal{P}_{\mathcal{S}}^{\mathtt{0}}$ is applied, is just the change of entropies in $\mathcal{S}$ and $\mathcal{R}$.

$$\langle \Delta S_{\mathcal{C}}\rangle = \langle \Delta S_{\mathcal{S}}\rangle + \langle \Delta S_{\mathcal{R}}\rangle \tag{116}$$

From axiom 3.4, the *generalized second law of Thermodynamics*, we know that the average change in entropy cannot be negative for a closed system, i.e. $\langle \Delta S_{\mathcal{C}}\rangle \geq 0$.

$$\langle \Delta S_{\mathcal{S}}\rangle + \langle \Delta S_{\mathcal{R}}\rangle \geq 0 \tag{117}$$

Since $\langle \Delta S_{\mathcal{S}}\rangle = -k\ln 2$, we can conclude that $\langle \Delta S_{\mathcal{R}}\rangle$ is positive with the lower bound $k\ln 2$. Thus $\mathcal{P}_{\mathcal{S}}^{\mathtt{0}}$ necessarily increases the entropy in $\mathcal{R}$.

$$\boxed{\langle \Delta S_{\mathcal{R}}\rangle \geq k\ln 2 \qquad \text{when} \qquad \Delta H = -1\,\text{bit}} \tag{118}$$

We will label this kind of relation a "Landauer bound", since it sets a lower bound on the entropy increase in the *reservoir* $\mathcal{R}$ when a logical process is executing.[26]

[26] In section 6, when we target the fully quantum mechanical system we will solely focus on finding a similar *Landauer bound* to equation (118), but derived with more mathematical rigour, for an arbitrary change in information $\Delta H < 0$, and not contingent on the second law of thermodynamics.

5.8.4 Remarks about the Landauer bound

The result so far cannot be considered very controversial. We have imposed a condition on our physical processes (definition 5.2) which forces the entropy in a state to be determined from the *logical state* of $\mathcal{S}$. We then *decreased* the entropy in the logical state by 1 bit, and thus found a bound on entropy *increase* in the reservoir $\mathcal{R}$, as expressed by equation (118).

At this point, we have not shown that there is any implication of physical *irreversibility*, as Landauer's principle (proposition 5.1) claims. Put differently, we have not shown that the entropy of the *closed system* $\mathcal{C}$ must strictly increase, instead, it seems like we just moved some entropy from $\mathcal{S}$ to $\mathcal{R}$.

A likely cause for authors claiming that Landauer's principle is not valid is that they conduct an argument which takes us no longer than to this point, see for instance [20]. But to argue for the physical irreversibility we need to consider one final point (see section 5.9).

5.9 Inferring physical irreversibility in $\mathcal{C}$

Consider the identical set-up as before but with one significant change. We will use an "enlightened" logical process $\mathcal{P}_{\mathcal{S}}^{0\prime}$ which also resets to zero, but *does* carry information about the logical state of $\mathcal{S}$; i.e. there exist some correlation between the logical state of $\mathcal{S}$ ($\mathtt{x} = \mathtt{0}$ or $\mathtt{x} = \mathtt{1}$), and some secondary physical system that $\mathcal{P}_{\mathcal{S}}^{0\prime}$ can use to determine how to interact with $\mathcal{S}$.[27] Otherwise, everything else is the same as with our previous "ignorant" process $\mathcal{P}_{\mathcal{S}}^{0}$, in particular, we still demand that $\mathcal{P}_{\mathcal{S}}^{0\prime}$ should be *entropy-restoring*. Under these conditions, new physical mechanisms are made available to $\mathcal{P}_{\mathcal{S}}^{0\prime}$, since it can act differently depending on the logical state, and we can treat each logical state separately.

If the logical state is $\mathtt{x} = \mathtt{0}$, the process $\mathcal{P}_{\mathcal{S}}^{0\prime}$ simply does nothing, which clearly creates no change of entropy in either $\mathcal{S}$ or $\mathcal{R}$, making it a physically reversible process (the entropy of $\mathcal{C}$ is unchanged).

$$\text{If} \quad P(\mathtt{0}) = 1 \quad \Rightarrow \quad \langle \Delta S_{\mathcal{S}} \rangle = 0 \quad ; \quad \langle \Delta S_{\mathcal{R}} \rangle = 0 \tag{119}$$

$$\Rightarrow$$

$$\langle \Delta S_{\mathcal{C}} \rangle = 0 \tag{120}$$

If the logical state is $\mathtt{x} = \mathtt{1}$, the process $\mathcal{P}_{\mathcal{S}}^{0\prime}$ can carry out a logical *not operation* (see section 5.3.2), in which the entropy of the logical state—as calculated from equation (108)—is unchanged. Thus the second law of Thermodynamics implies no lower bound on the entropy increase in $\mathcal{R}$, and again the process is physically reversible.

$$\text{If} \quad P(\mathtt{1}) = 1 \quad \Rightarrow \quad \langle \Delta S_{\mathcal{S}} \rangle = 0 \quad \Rightarrow \quad \langle \Delta S_{\mathcal{R}} \rangle \geq 0 \tag{121}$$

$$\Rightarrow$$

$$\langle \Delta S_{\mathcal{C}} \rangle \geq 0 \tag{122}$$

We now take a step back to compare the two situations—the former *ignorant* process $\mathcal{P}_{\mathcal{S}}^{0}$, versus the latter *enlightened* process $\mathcal{P}_{\mathcal{S}}^{0\prime}$. When we consider

[27] Note that the $\mathcal{S}$ constructed such that whether the logical state is $\mathtt{x} = \mathtt{0}$ or $\mathtt{x} = \mathtt{1}$ is a measurable quality, and each logical state is assumed to be stable over time.

the effect on the microstates $\{\mu\}$ of the information-bearing system $\mathcal{S}$, the two situations are *identical*. Before either process is applied, the logical state was *either* $\mathtt{x} = \mathtt{0}$ *or* $\mathtt{x} = \mathtt{1}$, and after the process, the logical state is $\mathtt{x} = \mathtt{0}$. Also, each process is *entropy-restoring*.[28] But still, we have found conflicting changes of entropy in $\mathcal{S}$.

$$\text{For } \mathcal{P}_{\mathcal{S}}^{0} \ : \ \langle \Delta S_{\mathcal{S}} \rangle = -k \ln 2 \tag{123}$$

$$\text{For } \mathcal{P}_{\mathcal{S}}^{0\prime} \ : \ \langle \Delta S_{\mathcal{S}} \rangle = 0 \tag{124}$$

This contradictory situation can be resolved by attributing the former decrease of entropy—not to some decrease of the space of states for $\mathcal{S}$—but as a decrease in the uncertainty from the point of view of the process $\mathcal{P}_{\mathcal{S}}^{0}$.

We then consider the changes of entropy in $\mathcal{R}$.

$$\text{For } \mathcal{P}_{\mathcal{S}}^{0} \ : \ \langle \Delta S_{\mathcal{R}} \rangle \geq k \ln 2 \tag{125}$$

$$\text{For } \mathcal{P}_{\mathcal{S}}^{0\prime} \ : \ \langle \Delta S_{\mathcal{R}} \rangle \geq 0 \tag{126}$$

Since the processes are different from a physical point of view there is no necessary condition that the effects on $\mathcal{R}$ should be the same. Thus, for the process $\mathcal{P}_{\mathcal{S}}^{0}$, the increase of entropy in $\mathcal{R}$ *can*, and *should*, be considered as enlarging the available space of states. Because if it was not, from the point of view of the process $\mathcal{P}_{\mathcal{S}}^{0}$, the second law of thermodynamics would not hold.

We can view this increase in the entropy of $\mathcal{R}$ as originating from the *ignorance* of the process $\mathcal{P}_{\mathcal{S}}^{0}$, which is not able to operate at the efficiency of the *enlightened* process $\mathcal{P}_{\mathcal{S}}^{0\prime}$ as far as entropy production is concerned.

When we then ask if the ignorant logical process $\mathcal{P}_{\mathcal{S}}^{0}$ (for which $\Delta H = -1\,\text{bit}$) is physically reversible from the perspective of the *closed system* $\mathcal{C}$ itself. It would be incorrect to include the reduction of uncertainty of $\mathcal{P}_{\mathcal{S}}^{0}$ when we evaluate some *physical* change in entropy of $\mathcal{C}$ (related to the size of the state space), and we therefore set $\langle \Delta S_{\mathcal{S}} \rangle = 0$.

$$\langle \Delta S_{\mathcal{C}} \rangle = \langle \Delta S_{\mathcal{S}} \rangle + \langle \Delta S_{\mathcal{R}} \rangle \qquad \Rightarrow \tag{127}$$

$$\boxed{\langle \Delta S_{\mathcal{C}} \rangle \geq k \ln 2 \qquad \text{when} \qquad \Delta H = -1\,\text{bit}} \tag{128}$$

Thus we have shown that the change in entropy in the closed system $\mathcal{C}$ is strictly positive, and the process $\mathcal{P}_{\mathcal{S}}^{0}$ must represent a physically irreversible process, as suggested in proposition 5.1.

Another way to think about it is the following. From the perspective of some physical system $\mathcal{O}$, which is correlated with the initial logical state ($\mathtt{x} = \mathtt{0}$ or $\mathtt{x} = \mathtt{1}$), there is no change in entropy of $\mathcal{S}$ (i.e. $\langle \Delta S_{\mathcal{S}} \rangle = 0$), and this is the correct *perspective* to take, since *reversing* a reset-to-zero process should entail a recovery of the initial logical state.

Claude Shannon makes an independent remark in his seminal paper [9] (page 1), which is surprisingly relevant here. "The system [here, the process $\mathcal{P}_{\mathcal{S}}^{0}$] must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design."

[28] The remaining entropy, for a determined logical state, is on average S_{RES}.

5.9.1 Subtleties and controversies

In order to link the uncontroversial *Landauer bound* from section 5.8.3 to the claim of *physical irreversibility* made by Landauer's principle (proposition 5.1), we need some argument connecting the two. But we should note that the interlinking argument made in section 5.9 is quite subtle, and not entirely robust if one assumes some other point of view towards entropy in physical systems.

However, it is not recommended to examine the argument in complete isolation. Instead, this discussion for a general *information-bearing system* should be considered as a backdrop to a number of publications which reproduces the result when analyzing Landauer's principle in terms of constrained or specific physical systems—among others see [1, 14–19].

We can speculate that it is the delicate nature of the argument (for a general information-bearing system) which have kept others from engaging in it, and we note that the author Owen Maroney (whose argument start out very similar to this one) does not include any arguments of this kind [20]. Generally, we cannot claim that the conclusion of physical irreversibility is backed up by consensus in the literature, but the hope is that arguments and discussions in general terms—such as the one in section 5.9—can expose the correct treatment and work in favour of consensus building.

5.10 A remark about noise

Physical systems are extremely difficult to isolate from random noise. Therefore it is a clear idealization to assume that the *parameter* which was used to encode logical states, should be *free* (see section 5.3.3). In reality, we can assume that there exist some effects which will widen locally concentrated probability distributions with time, thus making the entropy in $\mathcal{S}$ increase over time. However, this effect will only increase the magnitude of the *Landauer bound* on $\langle \Delta S_{\mathcal{R}} \rangle$ from equation (118). Thus $\langle \Delta S_{\mathcal{C}} \rangle \geq k \ln 2$ remains a lower bound, and adding noise will not change our result.

5.11 Generalizing to information reset of arbitrary size

We have constructed an argument for *Landauer's principle* in the case of resetting 1 bit of data, i.e. $\Delta H = -1$ (see equation (128)). But we can easily extend our analysis to any *integer* value for ΔH.

Take some number of physical *parameters* $\{\mathcal{S}_i\}$ and consider our them all together as our *information-bearing system*. Since entropy is additive, the *Landauer bound* in equation (118) will be multiplied by the number of subsystems $|\{\mathcal{S}_i\}|$, and this number will carry over to equation (128).

$$\langle \Delta S_{\mathcal{C}} \rangle \geq -k \ln 2 \, \Delta H \qquad \text{for} \qquad -\Delta H \in \mathbb{N}_1 \tag{129}$$

To generalize to real values, see the 2009 paper by Maroney [20] where an argument is introduced using *alphabet* sets of any finite cardinality.

5.12 Attempting to break Landauer's principle

5.12.1 Measuring the logical state

We have assumed that $\mathcal{P}_S^0$ has no prior information about the logical state, but an immediate concern is that we have not prohibited measurements of the logical state. One might think that we can have a *measuring process*, $\mathcal{P}_S^{0\prime\prime}$, the reset procedure by performing a measurement of the logical state, in order to then operate at the efficiency of $\mathcal{P}_S^{0\prime}$ from section 5.9. But in order for $\mathcal{P}_S^{0\prime\prime}$ to make use of the measurement outcome, a "copy" of the logical state must be encoded somewhere else, and made accessible to $\mathcal{P}_S^{0\prime\prime}$. Then, after $\mathcal{S}$ has been reset, we still have to reset the copy, only pushing the problem of resetting a logical state ahead of ourselves.

This points to a fundamental impracticality of any reset process with prior information, such as $\mathcal{P}_S^{0\prime}$ (section 5.9). In any kind of practical device for information manipulation there would be other processes which may affect the logical state of $\mathcal{S}$, thus rendering the information in $\mathcal{P}_S^{0\prime}$ useless, and a logical process such as $\mathcal{P}_S^{0\prime}$ can at most be used once.

5.12.2 Entropy sinks

When reading literature on Landauer's principle, one often comes across the requirement that some system must operate in a cyclical fashion, returning to the initial state periodically. Indeed, that is also a cornerstone of Classical Thermodynamics (section 3.2). The rationale behind such demands is to avoid introducing *entropy sinks* into the argument, and reaching some incorrect conclusion.

For example, we could imagine a naïve attempt to improve our *measuring process* $\mathcal{P}_S^{0\prime\prime}$ (from section 5.12.1) by supplying a large register for the reset process, such that measurement results, for all practical purposes, never have to be reset. In that case, we introduce an *entropy sink* in our argument, where entropy seems to behave differently because we have moved it to some place of the system where it is hard to see, and there is no longer any *resetting of information* going on; we only move it from one subsystem to another.

The way we detect potential *entropy sinks* is to ask ourselves: "What modifications do we have to introduce in order to make the scheme return to its initial state?".

Indeed it is also possible to regard any process that is not *entropy-restoring* (see definition 5.2 in section 5.7) as introducing a finite *entropy sink*, allowing Landauer's principle to be violated as long as there is room to *hide* additional entropy. Requiring that any process has to be *entropy-restoring* makes the process more cyclical, and prohibits using the microstates $\{\mu_0\}$ and $\{\mu_1\}$ as entropy sinks.

5.13 Comparison to Maroney's 2009 paper

Since the initial setup in this section closely follows that of Owen J. E. Maroney in his 2009 paper [20] it is appropriate to compare the final result to his, and we will point out some differences.

As previously mentioned, Maroney's treatment is initially more general. Two alphabets for logical states are introduced, for *input* and *output*, and they both have some arbitrary finite cardinality (in contrast to this treatment where a single alphabet has only two members). Also, logical operations are not limited to only deterministic operations.

Even though the initial setup in this thesis is contained as a special case of Maroney's treatment, the arguments deviate from each other and we reach contrasting conclusions. Here we will identify which points in Maroney's treatment appear inaccurate and incomplete from the point of view of the current treatment.

In order to deal with the *complication from undefined entropies* (see section 5.6), Maroney attributes the *small-scale* entropy of some logical state, i.e. $S(\mu_{\mathtt{x}})$, to entropy in the "non-information-bearing degrees of freedom". In his paper [20], see the equation (111) and the subsequent paragraph. This is clearly a problematic approach since the entropy $S(\mu_{\mathtt{x}})$ resides in our *information-bearing system* $\mathcal{S}$ (as defined in section 5.3.3), and not the reservoir $\mathcal{R}$, a.k.a. the *non-information-bearing degrees of freedom* (see section 5.3.5 and 5.3.9).

We also note that Maroney does not show that the total entropy of the closed system $\mathcal{C}$ increases, since he does not engage in any argument like the one in section 5.9. This means that Maroney only argues for what we here termed a *Landauer bound* (section 5.8), and does not address whether *physical irreversibility* follows (section 5.9). Of course, the absence of an argument is not a counter-argument. Thus the current treatment can be taken as an extension of Maroney's argument, such that the conclusion—when looking at a general physical system—is the same conclusion as originally put forth by Landauer [1].

6 A Landauer bound in Quantum Mechanics

The claim made by Cabello et al. (see problem II in section 1, and section 7) is backed up by the assertion that Landauer's principle is "considered valid in the quantum domain" [2], and a paper by David Reeb and Michael M. Wolf is cited, [21]. Here we take a look at the assertion, and begin to generalize *Landauer's principle* to Quantum Mechanics—without relying on the *second law of Thermodynamics*—in an easy accessible reconstruction of the most relevant parts of two papers by Reeb and Wolf, authored in 2013 [22] and 2014 [21].

Suppose we have some quantum system $\mathcal{S}$, whose entropy we want to lower. We are free to design any interaction Hamiltonian, but we have limited information about the state of the reservoir $\mathcal{R}$, which $\mathcal{S}$ will interact with. In fact, we only know some average energy of $\mathcal{R}$. Then, what are the limitations in such an interaction?

Compared to section 5, we will not introduce the structure of *logical states*, *logical operations* and *logical processes*, instead, we will look at the bare state of the quantum system. Therefore we will not here make any claims about *Landauer's principle*, but instead, we derive a version of a *Landauer bound* (see section 5.8.3).

6.1 Premises

We begin by defining some premises to base the argument upon. The number of premises is increased by one compared to the formulation by Reeb and Wolf [21], but they correspond to an identical set-up.

Premise 6.1 (System and model). A closed system $\mathcal{C}$ is modelled quantum-mechanically, and it is divided into two sub-systems, $\mathcal{S}$ and $\mathcal{R}$. $\mathcal{S}$ is our *information-bearing system*, and $\mathcal{R}$ is a *reservoir*.

Premise 6.2 (Density operators). $\mathcal{S}$ and $\mathcal{R}$ are both described by finite-dimensional density operators $\hat{\rho}_{\mathcal{S}}$ and $\hat{\rho}_{\mathcal{R}}$ associated with Hilbert spaces of dimensions $d_{\mathcal{S}}$ and $d_{\mathcal{R}}$ respectively (see section 2 axiom 2.1).

Premise 6.3 (Initial state of $\mathcal{R}$). The reservoir $\mathcal{R}$ is initially, at time t_0, in the canonical thermal state (see section 4.2). Let $\hat{H}_{\mathcal{R}}$ be the Hamiltonian of the reservoir, and let $\beta \in [-\infty, \infty]$ be its *reciprocal temperature*.

$$\hat{\rho}_{\mathcal{R}} = \frac{\mathrm{e}^{-\beta \hat{H}_{\mathcal{R}}}}{\mathfrak{Tr}\big[\mathrm{e}^{-\beta \hat{H}_{\mathcal{R}}}\big]} \tag{130}$$

Premise 6.4 (Initial state of $\mathcal{C}$). The entire closed system $\mathcal{C}$ is initially, at time t_0, in a state that has no correlations between $\mathcal{S}$ and $\mathcal{R}$.

$$\hat{\rho}_{\mathcal{C}} = \hat{\rho}_{\mathcal{S}} \otimes \hat{\rho}_{\mathcal{R}} \tag{131}$$

Premise 6.5 (Unitary evolution). At time t_0, the two systems couple through an interaction Hamiltonian $\hat{H}_{\mathcal{SR}}$, and evolve unitarily according to

some Hamiltonian, $\hat{H}_{\mathcal{C}}$, on the entire closed system, until the time $t_0 + \Delta t$. We denote final states with primes, such as $\hat{\rho}'_{\mathcal{C}}$.

$$\hat{H}_{\mathcal{C}} = \hat{H}_{\mathcal{S}} \otimes \mathbb{1}_{\mathcal{R}} \; + \; \mathbb{1}_{\mathcal{S}} \otimes \hat{H}_{\mathcal{R}} \; + \; \hat{H}_{\mathcal{SR}} \tag{132}$$

$$\hat{U}(t) := \mathrm{e}^{-\mathrm{i}\hat{H}_{\mathcal{C}} t} \tag{133}$$

$$\hat{\rho}'_{\mathcal{C}} := \hat{U}(\Delta t)\, \hat{\rho}_{\mathcal{C}}\, \hat{U}^{\dagger}(\Delta t) \quad ; \quad \Delta t > 0 \tag{134}$$

From this construction, we should point out a few things.

Remark I. We make no specific assumptions about the initial state of the information-bearing system $\mathcal{S}$, thus $\hat{\rho}_{\mathcal{S}}$ can have any degree of mixedness; carrying any amount of von Neumann entropy. We do not impose any requirements on the systems Hamiltonian $\hat{H}_{\mathcal{S}}$ either. □

Remark II. Just as in section 5, the *information-bearing system* $\mathcal{S}$ is initially assumed to be independent of the *reservoir* $\mathcal{R}$, until some interaction begins. This means that there are no correlations between $\mathcal{S}$ and $\mathcal{R}$, which is equivalent to modelling the combined system $\mathcal{C}$ with a product state (see premise 6.4). □

6.2 Defining quantities

Our goal is now to relate changes in *von Neumann entropy* (section 3.7) of $\mathcal{S}$ to changes in the average entropy or energy in $\mathcal{R}$. We therefore carry on with further definitions of some useful quantities.

As defined, $\hat{\rho}_{\mathcal{S}}$ is the initial state and $\hat{\rho}'_{\mathcal{S}}$ is the final state, of $\mathcal{S}$. Then let $\langle \Delta S_{\mathcal{S}} \rangle$ be the average change in the entropy carried by $\mathcal{S}$.[29]

$$\langle \Delta S_{\mathcal{S}} \rangle := S_N(\hat{\rho}'_{\mathcal{S}}) - S_N(\hat{\rho}_{\mathcal{S}}) \qquad \Rightarrow \tag{135}$$

$$\langle \Delta S_{\mathcal{S}} \rangle = \mathfrak{Tr}\left[\hat{\rho}_{\mathcal{S}} \ln \hat{\rho}_{\mathcal{S}} - \hat{\rho}'_{\mathcal{S}} \ln \hat{\rho}'_{\mathcal{S}}\right] \tag{136}$$

A state of a subsystem (such as $\mathcal{S}$) is found by *tracing* over other subsystems (here $\mathcal{R}$) on the density operator of the entire closed system (here modelled with $\hat{\rho}_{\mathcal{C}}$). Thus we can find the final state of $\mathcal{S}$ by tracing over $\hat{\rho}'_{\mathcal{C}}$.

$$\hat{\rho}'_{\mathcal{S}} = \mathfrak{Tr}_{\mathcal{R}}\left[\hat{\rho}'_{\mathcal{C}}\right] \tag{137}$$

Then, let $\langle \Delta S_{\mathcal{R}} \rangle$ be the average change in entropy of $\mathcal{R}$.[30]

$$\langle \Delta S_{\mathcal{R}} \rangle := S_N(\hat{\rho}'_{\mathcal{R}}) - S_N(\hat{\rho}_{\mathcal{R}}) \qquad \Rightarrow \tag{138}$$

$$\langle \Delta S_{\mathcal{R}} \rangle = \mathfrak{Tr}\left[\hat{\rho}_{\mathcal{R}} \ln \hat{\rho}_{\mathcal{R}} - \hat{\rho}'_{\mathcal{R}} \ln \hat{\rho}'_{\mathcal{R}}\right] \tag{139}$$

Again, the final state of a subsystem, such as $\mathcal{R}$, is found by tracing over the closed system.

$$\hat{\rho}'_{\mathcal{R}} := \mathfrak{Tr}_{\mathcal{S}}\left[\hat{\rho}'_{\mathcal{C}}\right] \tag{140}$$

[29] Thus positive values correspond to an *increase*, and negative values correspond to a *decrease*, in entropy. Note that this definition is chosen with the opposite sign compared to Reeb and Wolf [21].

[30] Just as with $\langle \Delta S_{\mathcal{S}} \rangle$, positive values correspond to an *increase* in entropy.

To quantify the heat, we will assume that all energy in $\mathcal{R}$ will be in the form of heat. Reed and Wolf state that this is justified since the "energy is not 'ordered' since $\mathcal{R}$ is an initially thermal reservoir, which may absorb energy from $\mathcal{S}$ during the process and spread the energy over many states" [21]. And we choose to define the sign of heat $\langle \Delta Q_{\mathcal{R}} \rangle$ such that heat transferred *into* $\mathcal{R}$ corresponds a positive value.

$$\langle \Delta Q_{\mathcal{R}} \rangle := \big\langle H_{\mathcal{R}}(t+\Delta t) \big\rangle - \big\langle H_{\mathcal{R}}(t) \big\rangle \;=\; \mathfrak{Tr}\big[\hat{\rho}'_{\mathcal{R}} H\big] - \mathfrak{Tr}\big[\hat{\rho}_{\mathcal{R}} H\big] \quad \Rightarrow \tag{141}$$

$$\langle \Delta Q_{\mathcal{R}} \rangle = \mathfrak{Tr}\big[(\hat{\rho}'_{\mathcal{R}} - \hat{\rho}_{\mathcal{R}}) H\big] \tag{142}$$

6.2.1 Remark on averages

Whether we view the classical probability distribution that a density operator can carry (see section 3.7) to mean a statistical distribution in an ensemble, or our best possible description of the system under known conditions, it is clear that some specific interaction between $\mathcal{S}$ and $\mathcal{R}$ may not respect calculated changes, such as $\langle \Delta S_{\mathcal{S}} \rangle$ or $\langle \Delta Q_{\mathcal{R}} \rangle$.[31] It is only in some statistical limit that we expect these relations to hold, and we shall emphasize this point with brackets around our quantities.

6.3 A Landauer bound in terms of entropy

We consider the von Neumann entropies in $\mathcal{S}$ and $\mathcal{R}$. With the definitions of $\langle \Delta S_{\mathcal{S}} \rangle$ and $\langle \Delta S_{\mathcal{R}} \rangle$ from section 6.2, we will show that when entropy changes, the net entropy change will non-negative.

Theorem 6.1 (Landauer entropic bound in finite-dimensional state space). When entropy in either sub-system $\mathcal{S}$ or $\mathcal{R}$ changes, the net change is always non-negative.

$$\langle \Delta S_{\mathcal{S}} \rangle + \langle \Delta S_{\mathcal{R}} \rangle \;=\; I(\hat{\rho}'_{\mathcal{S}} : \hat{\rho}'_{\mathcal{R}}) \;\geq\; 0 \tag{143}$$

Equality holds if and only if the final state is a product state $\hat{\rho}'_{\mathcal{C}} = \hat{\rho}'_{\mathcal{S}} \otimes \hat{\rho}'_{\mathcal{R}}$.

Here, $I(\hat{\rho}'_{\mathcal{S}} : \hat{\rho}'_{\mathcal{R}})$ is the quantum-mechanical generalization of *mutual information*, discussed in section A.10. Note that equality, i.e. $I(\hat{\rho}'_{\mathcal{S}} : \hat{\rho}'_{\mathcal{R}}) = 0$, does not necessarily require both $\langle \Delta S_{\mathcal{S}} \rangle$ and $\langle \Delta S_{\mathcal{R}} \rangle$ to be zero, we can in principle still have entropy transactions between $\mathcal{S}$ and $\mathcal{R}$, as long as our final state $\hat{\rho}'_{\mathcal{C}}$ is a product state.

Proof (Theorem 6.1). We will use the additive property of von Neumann entropy for product states, $S_N(\hat{\rho}_1 \otimes \hat{\rho}_2) = S_N(\hat{\rho}_1) + S_N(\hat{\rho}_2)$ (section 3.5.4), the invariance of entropy under unitary transformations, $S_N(\hat{\rho}) = S_N(\hat{U}\hat{\rho}\hat{U}^\dagger)$ (section A.5), and the non negative property of mutual information, $I(\hat{\rho}_1 : \hat{\rho}_2) \geq 0$ (section A.10).

$$\langle \Delta S_{\mathcal{S}} \rangle + \langle \Delta S_{\mathcal{R}} \rangle = S_N(\hat{\rho}'_{\mathcal{S}}) - S_N(\hat{\rho}_{\mathcal{S}}) + S_N(\hat{\rho}'_{\mathcal{R}}) - S_N(\hat{\rho}_{\mathcal{R}}) = \tag{144}$$

[31] See section 4.1 for an argument on how *our best possible description* can be connected with *statistical outcomes*.

$$= S_N(\hat{\rho}'_{\mathcal{S}}) + S_N(\hat{\rho}'_{\mathcal{R}}) - S_N(\hat{\rho}_{\mathcal{C}}) = \tag{145}$$

$$= S_N(\hat{\rho}'_{\mathcal{S}}) + S_N(\hat{\rho}'_{\mathcal{R}}) - S_N(\hat{U}\hat{\rho}_{\mathcal{C}}\hat{U}^\dagger) = \tag{146}$$

$$= S_N(\hat{\rho}'_{\mathcal{S}}) + S_N(\hat{\rho}'_{\mathcal{R}}) - S_N(\hat{\rho}'_{\mathcal{C}}) \equiv I(\hat{\rho}'_{\mathcal{S}} : \hat{\rho}'_{\mathcal{R}}) \geq 0 \tag{147}$$

■

6.4 A Landauer bound in terms of heat

Here we repeat a similar analysis as in section 6.3, but we compare changes in entropy of $\mathcal{S}$ with changes in *heat* of $\mathcal{R}$. As discussed in section 6.2 we take *heat* to mean the total change in energy of the state.

Theorem 6.2 (Landauer heat bound in finite-dimensional state space). When entropy in system $\mathcal{S}$ changes, the change in heat in $\mathcal{R}$ will respond such that their sum is always non-negative.

$$\langle \Delta S_{\mathcal{S}} \rangle + \beta \langle \Delta Q_{\mathcal{R}} \rangle \;=\; I(\hat{\rho}'_{\mathcal{S}} : \hat{\rho}'_{\mathcal{R}}) + S(\hat{\rho}'_{\mathcal{R}} \| \hat{\rho}_{\mathcal{R}}) \;\geq\; 0 \tag{148}$$

Equality holds if and only if $\langle \Delta S_{\mathcal{S}} \rangle = 0$ and $\langle \Delta Q_{\mathcal{R}} \rangle = 0$.

Here, $I(\hat{\rho}'_{\mathcal{S}} : \hat{\rho}'_{\mathcal{R}})$ is *mutual information* (section A.10), and $S(\hat{\rho}'_{\mathcal{R}} \| \hat{\rho}_{\mathcal{R}})$ is the so called *relative entropy*, (section A.6).

Proof (Theorem 6.2). We intend to rewrite $\langle \Delta S_{\mathcal{R}} \rangle$, as defined in equation (139), in terms of change in energy. Our intention is to expose each step, and allow the reader to follow with relative comfort.

$$\langle \Delta S_{\mathcal{R}} \rangle = \mathfrak{Tr}\big[\hat{\rho}_{\mathcal{R}} \ln \hat{\rho}_{\mathcal{R}} - \hat{\rho}'_{\mathcal{R}} \ln \hat{\rho}'_{\mathcal{R}}\big] = \tag{149}$$

$$= \mathfrak{Tr}\left[\hat{\rho}_{\mathcal{R}} \ln \frac{\mathrm{e}^{-\beta \hat{H}}}{\mathfrak{Tr}\big[\mathrm{e}^{-\beta \hat{H}}\big]}\right] - \mathfrak{Tr}\big[\hat{\rho}'_{\mathcal{R}} \ln \hat{\rho}'_{\mathcal{R}}\big] = \tag{150}$$

$$= \mathfrak{Tr}\left[\hat{\rho}_{\mathcal{R}} \Big(-\beta \hat{H} - \ln\Big(\mathfrak{Tr}\big[\mathrm{e}^{-\beta \hat{H}}\big]\Big)\Big)\right] - \mathfrak{Tr}\big[\hat{\rho}'_{\mathcal{R}} \ln \hat{\rho}'_{\mathcal{R}}\big] = \tag{151}$$

$$= -\beta\, \mathfrak{Tr}\Big[\hat{\rho}_{\mathcal{R}} \hat{H}\Big] - \mathfrak{Tr}\left[\hat{\rho}_{\mathcal{R}} \ln\Big(\mathfrak{Tr}\big[\mathrm{e}^{-\beta \hat{H}}\big]\Big)\right] - \mathfrak{Tr}\big[\hat{\rho}'_{\mathcal{R}} \ln \hat{\rho}'_{\mathcal{R}}\big] = \tag{152}$$

$$= -\beta\, \mathfrak{Tr}\Big[\hat{\rho}_{\mathcal{R}} \hat{H}\Big] - \ln\Big(\mathfrak{Tr}\big[\mathrm{e}^{-\beta \hat{H}}\big]\Big)\, \mathfrak{Tr}\big[\hat{\rho}_{\mathcal{R}}\big] - \mathfrak{Tr}\big[\hat{\rho}'_{\mathcal{R}} \ln \hat{\rho}'_{\mathcal{R}}\big] = \tag{153}$$

$$= -\beta\, \mathfrak{Tr}\Big[\hat{\rho}_{\mathcal{R}} \hat{H}\Big] - \ln\Big(\mathfrak{Tr}\big[\mathrm{e}^{-\beta \hat{H}}\big]\Big) - \mathfrak{Tr}\big[\hat{\rho}'_{\mathcal{R}} \ln \hat{\rho}'_{\mathcal{R}}\big] \tag{154}$$

We then add and subtract a term, $\beta\, \mathfrak{Tr}[\hat{\rho}'_{\mathcal{R}} \hat{H}]$.

$$\begin{aligned} &\langle \Delta S_{\mathcal{R}} \rangle = \\ &= \beta\, \mathfrak{Tr}\Big[\hat{\rho}'_{\mathcal{R}} \hat{H}\Big] - \beta\, \mathfrak{Tr}\Big[\hat{\rho}_{\mathcal{R}} \hat{H}\Big] - \ln\Big(\mathfrak{Tr}\big[\mathrm{e}^{-\beta \hat{H}}\big]\Big) - \beta\, \mathfrak{Tr}\Big[\hat{\rho}'_{\mathcal{R}} \hat{H}\Big] - \mathfrak{Tr}\big[\hat{\rho}'_{\mathcal{R}} \ln \hat{\rho}'_{\mathcal{R}}\big] = \end{aligned} \tag{155}$$

$$= \beta\, \mathfrak{Tr}\Big[(\hat{\rho}'_{\mathcal{R}} - \hat{\rho}_{\mathcal{R}}) \hat{H}\Big] - \mathfrak{Tr}\big[\hat{\rho}'_{\mathcal{R}}\big] \ln\Big(\mathfrak{Tr}\big[\mathrm{e}^{-\beta \hat{H}}\big]\Big) - \beta\, \mathfrak{Tr}\Big[\hat{\rho}'_{\mathcal{R}} \hat{H}\Big] - \mathfrak{Tr}\big[\hat{\rho}'_{\mathcal{R}} \ln \hat{\rho}'_{\mathcal{R}}\big] = \tag{156}$$

$$= \beta\,\mathfrak{Tr}\Big[(\hat{\rho}'_{\mathcal{R}}\text{-}\hat{\rho}_{\mathcal{R}})\hat{H}\Big] - \mathfrak{Tr}\Big[\hat{\rho}'_{\mathcal{R}}\ln\Big(\mathfrak{Tr}\big[\mathrm{e}^{-\beta\hat{H}}\big]\Big)\Big] - \beta\,\mathfrak{Tr}\Big[\hat{\rho}'_{\mathcal{R}}\ln\mathrm{e}^{\hat{H}}\Big] - \mathfrak{Tr}\big[\hat{\rho}'_{\mathcal{R}}\ln\hat{\rho}'_{\mathcal{R}}\big] = \quad (157)$$

$$= \beta\,\mathfrak{Tr}\Big[(\hat{\rho}'_{\mathcal{R}}\text{-}\hat{\rho}_{\mathcal{R}})\hat{H}\Big] + \mathfrak{Tr}\left[\hat{\rho}'_{\mathcal{R}}\ln\frac{1}{\mathfrak{Tr}\big[\mathrm{e}^{-\beta\hat{H}}\big]}\right] + \mathfrak{Tr}\Big[\hat{\rho}'_{\mathcal{R}}\ln\mathrm{e}^{-\beta\hat{H}}\Big] - \mathfrak{Tr}\big[\hat{\rho}'_{\mathcal{R}}\ln\hat{\rho}'_{\mathcal{R}}\big] = \quad (158)$$

$$= \beta\,\mathfrak{Tr}\Big[(\hat{\rho}'_{\mathcal{R}}\text{-}\hat{\rho}_{\mathcal{R}})\hat{H}\Big] + \mathfrak{Tr}\left[\hat{\rho}'_{\mathcal{R}}\ln\frac{\mathrm{e}^{-\beta\hat{H}}}{\mathfrak{Tr}\big[\,\mathrm{e}^{-\beta\hat{H}}\big]}\right] - \mathfrak{Tr}\big[\hat{\rho}'_{\mathcal{R}}\ln\hat{\rho}'_{\mathcal{R}}\big] = \quad (159)$$

$$= \beta\,\mathfrak{Tr}\Big[(\hat{\rho}'_{\mathcal{R}} - \hat{\rho}_{\mathcal{R}})\hat{H}\Big] + \mathfrak{Tr}\big[\hat{\rho}'_{\mathcal{R}}\ln\hat{\rho}_{\mathcal{R}}\big] - \mathfrak{Tr}\big[\hat{\rho}'_{\mathcal{R}}\ln\hat{\rho}'_{\mathcal{R}}\big] = \quad (160)$$

$$= \beta\,\mathfrak{Tr}\Big[(\hat{\rho}'_{\mathcal{R}} - \hat{\rho}_{\mathcal{R}})\hat{H}\Big] - \mathfrak{Tr}\Big[\hat{\rho}'_{\mathcal{R}}(\ln\hat{\rho}'_{\mathcal{R}} - \ln\hat{\rho}_{\mathcal{R}})\Big] \quad (161)$$

Here, we can identify the first term as the average change in heat $\langle \Delta Q_{\mathcal{R}}\rangle$ (multiplied by reciprocal temperature β) according to equation (142). The second term is, by definition, the so-called *relative entropy* $S(\hat{\rho}'_{\mathcal{R}}\|\hat{\rho}_{\mathcal{R}})$, discussed in section A.6. We can then rewrite $\langle \Delta S_{\mathcal{S}}\rangle$ in terms of heat, and insert this result into equation (143), from theorem 6.1.

$$\langle \Delta S_{\mathcal{R}}\rangle = \beta\langle \Delta Q_{\mathcal{R}}\rangle - S(\hat{\rho}'_{\mathcal{R}}\|\hat{\rho}_{\mathcal{R}}) \qquad \Rightarrow \quad (162)$$

$$\langle \Delta S_{\mathcal{S}}\rangle + \beta\langle \Delta Q_{\mathcal{R}}\rangle \;=\; I(\hat{\rho}'_{\mathcal{S}}\colon\hat{\rho}'_{\mathcal{R}}) + S(\hat{\rho}'_{\mathcal{R}}\|\hat{\rho}_{\mathcal{R}}) \geq 0 \quad (163)$$

In the last step we have used that both the mutual information $I(\hat{\rho}'_{\mathcal{S}}\colon\hat{\rho}'_{\mathcal{R}})$, and the relative entropy $S(\hat{\rho}'_{\mathcal{R}}\|\hat{\rho}_{\mathcal{R}})$, are non-negative (sections A.10 and A.6). Thus we have proven equation (148) in theorem 6.2, and it remains to prove the conditions for equality, as stated in the same theorem.

For equation (163) to be an equality we need both of the non-negative quantities $I(\hat{\rho}'_{\mathcal{S}}\colon\hat{\rho}'_{\mathcal{R}})$ and $S(\hat{\rho}'_{\mathcal{R}}\|\hat{\rho}_{\mathcal{R}})$ to be zero, and they both impose their individual restrictions. The *mutual information* is zero if and only if the final state is a product state $\hat{\rho}'_{\mathcal{C}} = \hat{\rho}'_{\mathcal{S}}\otimes\hat{\rho}'_{\mathcal{R}}$, meaning there are no correlations between the subsystems. And for the *relative entropy* to be zero, we must require $\hat{\rho}'_{\mathcal{R}} = \hat{\rho}_{\mathcal{R}}$. In conclusion, we have the following two conditions.

$$\hat{\rho}_{\mathcal{C}} = \hat{\rho}_{\mathcal{S}}\otimes\hat{\rho}_{\mathcal{R}} \qquad \text{and} \qquad \hat{\rho}'_{\mathcal{C}} = \hat{\rho}'_{\mathcal{S}}\otimes\hat{\rho}_{\mathcal{R}} \quad (164)$$

Lemma A.4 (see section A.7 in the appendix) then lets us know that the eigenvalues of a product state can be written as a simple multiplication of the eigenvalues from the sub-systems. Let the eigenvalues of $\hat{\rho}_{\mathcal{S}}$ be $\{s_i\}$, the eigenvalues of $\hat{\rho}'_{\mathcal{S}}$ be $\{s'_i\}$, and the eigenvalues of $\hat{\rho}_{\mathcal{R}}$ be $\{r_i\}$.

$$\mathfrak{Eig}[\hat{\rho}_{\mathcal{C}}] = \{s_i r_j\} \qquad ; \qquad \mathfrak{Eig}[\hat{\rho}'_{\mathcal{C}}] = \{s'_i r_j\} \quad (165)$$

Since $\hat{\rho}_{\mathcal{C}}$ and $\hat{\rho}'_{\mathcal{C}}$ are Hermitian matrices related by a unitary transform, according to lemma A.5 (section A.8 of the appendix) we know that they must both have the same eigenvalues.

$$s'_i r_j = s_i r_j \quad (166)$$

Since $\hat{\rho}_{\mathcal{R}}$ has trace 1 (axiom 2.1), we know that there must exist at least one eigenvalue that is larger than zero. This means that for some value(s) of j we can divide equation (166) by r_j.

$$s'_i = s_i \quad (167)$$

From lemma A.5 again, we can conclude that since $\hat{\rho}_{\mathcal{S}}$ and $\hat{\rho}'_{\mathcal{S}}$ are Hermitian, and that they have the same eigenvalues, they must be related by a unitary transform V.

$$\hat{\rho}'_{\mathcal{S}} = \hat{V}\hat{\rho}_{\mathcal{S}}\hat{V}^{\dagger} \tag{168}$$

This implies that the entropy of $\mathcal{S}$ does not change, $\langle \Delta S_{\mathcal{S}} \rangle = 0$, and our initial conditions from equation (164) require $\hat{\rho}_{\mathcal{R}}$ to be unchanged, thus its entropy or heat cannot change either, $\langle \Delta Q_{\mathcal{R}} \rangle = 0$.

The converse—showing that for $\langle \Delta S_{\mathcal{S}} \rangle = 0$ and $\langle \Delta Q_{\mathcal{R}} \rangle = 0$, will imply that $I(\hat{\rho}'_{\mathcal{S}} : \hat{\rho}'_{\mathcal{R}}) = 0$ and $S(\hat{\rho}'_{\mathcal{R}} \| \hat{\rho}_{\mathcal{R}}) = 0$—is a trivial consequence derived from equation (148) since both *mutual information* and *relative entropy* are non-negative quantities. ■

We further point out that for most non-trivial processes there is some energy or entropy exchange (i.e. $\Delta S > 0$ or $\Delta Q > 0$), and then the Landauer heat bound (theorem 6.2) is a strict inequality. We could then ask if there exists some non-zero and increasing function, $f(\Delta S)$, such that we can rewrite equation (148) from $\langle \Delta S_{\mathcal{S}} \rangle + \beta \langle \Delta Q_{\mathcal{R}} \rangle > 0$ to something like the following relation.

$$\langle \Delta S_{\mathcal{S}} \rangle + \beta \langle \Delta Q_{\mathcal{R}} \rangle \;\geq\; f(\Delta S) \tag{169}$$

In section 6.6 we shall return to this question.

6.5 Purifying $\mathcal{S}$ in finite-dimensional state space

Here we construct a lemma that will be useful for a deeper understanding of the purification process. The general idea is that there is a class of mixed states where all the eigenvalues in the density operator are larger than zero (e.g. a maximally mixed state). We can then gain some insight into the purification process by looking at the behaviour of the smallest eigenvalue (corresponding to the least likely pure state). We will show that there exist a bound on how small we can make this eigenvalue when interacting with a thermal reservoir with a finite state space.

Lemma 6.1 (Bound for lowering the smallest pure state probability). Let $p_{\min}(\hat{\rho})$ denote the smallest eigenvalue of any density operator $\hat{\rho}$, and let $E^{\downarrow}_{\mathcal{R}}$ and $E^{\uparrow}_{\mathcal{R}}$ be the lowest and highest energy eigenvalues of the reservoir Hamiltonian $\hat{H}_{\mathcal{R}}$. Then, for the system $\mathcal{S}$, this bounded Hamiltonian $\hat{H}_{\mathcal{R}}$ makes it impossible to transform $p_{\min}(\hat{\rho}_{\mathcal{S}}) > 0$ to zero, according to the following bound.

$$p_{\min}(\hat{\rho}'_{\mathcal{S}}) \;\geq\; \mathrm{e}^{-|\beta|(E^{\uparrow}_{\mathcal{R}} - E^{\downarrow}_{\mathcal{R}})}\, p_{\min}(\hat{\rho}_{\mathcal{S}}) \;\geq\; \mathrm{e}^{-2|\beta|\,\|\hat{H}_{\mathcal{R}}\|}\, p_{\min}(\hat{\rho}_{\mathcal{S}}) \tag{170}$$

Remark I. From this, we can conclude that a reservoir with a finite state space permits "purification" only if the Hamiltonian $\hat{H}_{\mathcal{R}}$ permits energies that are much larger than the typical energy kT.

Proof (Lemma 6.1). Using the spectral theorem (section A.1) we can express the density operator $\hat{\rho}'_{\mathcal{S}}$ in its diagonal form (see section 2). Then, let $p_{\min}(\hat{\rho}'_{\mathcal{S}})$

be its smallest eigenvalue, and let $|s_0\rangle$ be the corresponding normalized eigenstate. Also, let $\{|r_i\rangle\}$ be an orthonormal basis for the reservoir $\mathcal{R}$, and we define the abbreviated notation $|s_0\rangle \otimes |r_i\rangle =: |s_0, r_i\rangle$.

$$p_{\min}(\hat{\rho}'_{\mathcal{S}}) = \langle s_0|\hat{\rho}'_{\mathcal{S}}|s_0\rangle = \langle s_0|\mathfrak{Tr}_{\mathcal{R}}\left[\hat{\rho}'_{\mathcal{C}}\right]|s_0\rangle = \sum_{i=1}^{d_{\mathcal{R}}} \langle s_0, r_i|\hat{\rho}'_{\mathcal{C}}|s_0, r_i\rangle \tag{171}$$

Every matrix element $\langle s_0, r_i|\hat{\rho}'_{\mathcal{C}}|s_0, r_i\rangle \ \forall\, i$ has to be bigger or equal to the smallest eigenvalue of that operator $\hat{\rho}'_{\mathcal{C}}$, thus we rewrite our previous result as an inequality.

$$p_{\min}(\hat{\rho}'_{\mathcal{S}}) = \sum_{i=1}^{d_{\mathcal{R}}} \langle s_0, r_i|\hat{\rho}'_{\mathcal{C}}|s_0, r_i\rangle \geq \sum_{i=1}^{d_{\mathcal{R}}} p_{\min}(\hat{\rho}'_{\mathcal{C}}) = p_{\min}(\hat{U}\hat{\rho}_{\mathcal{C}}\hat{U}^\dagger)\, d_{\mathcal{R}} \tag{172}$$

We then use lemma A.5 (eigenvalues are unaffected by unitary transforms) on this expression, and in the last step below we use lemma A.4 (eigenvalues of a product operator are the products of the eigenvalues).

$$p_{\min}(\hat{\rho}'_{\mathcal{S}}) \geq p_{\min}(\hat{\rho}_{\mathcal{C}})\, d_{\mathcal{R}} = p_{\min}(\hat{\rho}_{\mathcal{S}} \otimes \hat{\rho}_{\mathcal{R}})\, d_{\mathcal{R}} = p_{\min}(\hat{\rho}_{\mathcal{S}})\, p_{\min}(\hat{\rho}_{\mathcal{R}})\, d_{\mathcal{R}} \tag{173}$$

Since we have an explicit expression for $\hat{\rho}_{\mathcal{R}}$—defined as the canonical thermal state in equation (130)—we will be able to find an explicit expression for $p_{\min}(\hat{\rho}_{\mathcal{R}})$ in terms of the *lowest* and *highest* energy eigenvalues of the Hamiltonian $\hat{H}_{\mathcal{R}}$, $E^{\downarrow}_{\mathcal{R}}$ and $E^{\uparrow}_{\mathcal{R}}$ respectively. Let us first assume that the temperature is positive, $\beta > 0$.

$$\begin{aligned} p_{\min}(\hat{\rho}_{\mathcal{R}})\Big|_{\beta>0} &= \frac{\mathrm{e}^{-\beta E^{\uparrow}_{\mathcal{R}}}}{\mathfrak{Tr}\left[\mathrm{e}^{-\beta \hat{H}_{\mathcal{R}}}\right]} \geq \frac{\mathrm{e}^{-\beta E^{\uparrow}_{\mathcal{R}}}}{\mathrm{e}^{-\beta E^{\downarrow}_{\mathcal{R}}}\, d_{\mathcal{R}}} = \\ &= \frac{\mathrm{e}^{-\beta(E^{\uparrow}_{\mathcal{R}} - E^{\downarrow}_{\mathcal{R}})}}{d_{\mathcal{R}}} \geq \frac{\mathrm{e}^{-2\beta\|\hat{H}_{\mathcal{R}}\|}}{d_{\mathcal{R}}} \end{aligned} \tag{174}$$

In the last step, we have used that the operator norm of the Hamiltonian $\|\hat{H}_{\mathcal{R}}\|$ is either $|E^{\uparrow}_{\mathcal{R}}|$ or $|E^{\downarrow}_{\mathcal{R}}|$, depending on which is larger, i.e. $\|\hat{H}_{\mathcal{R}}\| = \sup\{|E^{\downarrow}_{\mathcal{R}}|, |E^{\uparrow}_{\mathcal{R}}|\}$. Then, for negative temperatures, $\beta < 0$ we will just get a minus-sign in the exponent.

$$\begin{aligned} p_{\min}(\hat{\rho}_{\mathcal{R}})\Big|_{\beta<0} &= \frac{\mathrm{e}^{-\beta E^{\downarrow}_{\mathcal{R}}}}{\mathfrak{Tr}\left[\mathrm{e}^{-\beta \hat{H}_{\mathcal{R}}}\right]} \geq \frac{\mathrm{e}^{-\beta E^{\downarrow}_{\mathcal{R}}}}{\mathrm{e}^{-\beta E^{\uparrow}_{\mathcal{R}}}\, d_{\mathcal{R}}} = \\ &= \frac{\mathrm{e}^{\beta(E^{\uparrow}_{\mathcal{R}} - E^{\downarrow}_{\mathcal{R}})}}{d_{\mathcal{R}}} \geq \frac{\mathrm{e}^{2\beta\|\hat{H}_{\mathcal{R}}\|}}{d_{\mathcal{R}}} \end{aligned} \tag{175}$$

Inserting this back into equation (173).

$$p_{\min}(\hat{\rho}'_{\mathcal{S}}) \ \geq \ \mathrm{e}^{-|\beta|(E^{\uparrow}_{\mathcal{R}} - E^{\downarrow}_{\mathcal{R}})}\, p_{\min}(\hat{\rho}_{\mathcal{S}}) \ \geq \ \mathrm{e}^{-2|\beta|\,\|\hat{H}_{\mathcal{R}}\|}\, p_{\min}(\hat{\rho}_{\mathcal{S}}) \tag{176}$$

■

6.6 Finite-size corrections to the Landauer heat bound

From theorem 6.2 we have learned that the total increase the entropy and heat vanish if and only if both $\langle \Delta S_{\mathcal{S}} \rangle$ and $\langle \Delta Q_{\mathcal{R}} \rangle$ are zero—which is a trivial and passive case. For the more interesting case—when entropy and heat are exchanged between $\mathcal{S}$ and $\mathcal{R}$—Reeb and Wolf present a lower bound on the total change that takes the change in entropy $\langle \Delta S_{\mathcal{S}} \rangle$ as its argument [21]. However, only for

decreasing entropies, $\langle \Delta S_{\mathcal{S}} \rangle < 0$, is the bound *tight*[32]. Since in this thesis we are primarily interested in that case (decreasing the *uncertainty* in our state), we will here restrict the following analysis to the case $\langle \Delta S_{\mathcal{S}} \rangle < 0$.

Note that we will now give up the ability to make exact statements about entropy production, which we had in theorem 6.2. Instead, we are looking for a general statement which holds *no matter what* interaction Hamiltonian, and final state $\hat{\rho}'_{\mathcal{C}}$, we have.

To begin with, we define two functions that will be utilized in our theorem.

Definition 6.1 (Binary entropies). Consider two probability distributions, each over some binary choice (i.e. sets of *two* mutually exclusive events), $P := \{p, 1-p \mid 0 \leq p \leq 1\}$ and $Q := \{q, 1-q \mid 0 \leq q \leq 1\}$. We define the *binary entropy* H_2, and *binary relative entropy* D_2 as follows.

$$H_2(p) \quad := S(P) \qquad = p \ln \frac{1}{p} + (1-p) \ln \frac{1}{1-p} \tag{177}$$

$$D_2(p,q) := D_{KL}(P \| Q) = p \ln \frac{p}{q} + (1-p) \ln \frac{1-p}{1-q} \tag{178}$$

See section A.6 for further discussions about *relative entropy*. With this, we are prepared to state our theorem.

Theorem 6.3 (Finite size correction of Landauer heat bound). Given a reservoir, $\mathcal{R}$, described by a finite-dimensional density operator $\hat{\rho}_{\mathcal{R}}$ associated with a Hilbert space of dimension $d_{\mathcal{R}} \geq 2$. If the entropy of the information-bearing system $\mathcal{S}$ is reduced, i.e. $\langle \Delta S_{\mathcal{S}} \rangle < 0$, the Landauer heat bound (theorem 6.2), is tightly bounded from below by a *monotonically increasing* and *convex* function, M.

$$\langle \Delta S_{\mathcal{S}} \rangle + \beta \langle \Delta Q_{\mathcal{R}} \rangle \geq M_{d_{\mathcal{R}}}\big(|\langle \Delta S_{\mathcal{S}} \rangle|\big) \qquad \text{when} \qquad \langle \Delta S_{\mathcal{S}} \rangle < 0 \tag{179}$$

The function M takes $d_{\mathcal{R}}$—the dimensionality for the reservoir's Hilbert space—as a parameter, and it is defined as a minimization of *binary relative entropy*, $D_2(p,q)$ (see definition 6.1), in a bounded region Ω in p-q space.

$$\Omega := 0 \leq p, q \leq \frac{d_{\mathcal{R}} - 1}{d_{\mathcal{R}}} \tag{180}$$

Admissible points in the p-q space are also subject to a second restriction by a transcendental equation, in effect making the minimization problem one-dimensional.

$$\begin{aligned} &M_{d_{\mathcal{R}}}(\Delta S) := \\ &= \min_{\Omega} \Big\{ D_2(p,q) \;\Big|\; H_2(p) - H_2(q) + (p-q)\ln(d_{\mathcal{R}} - 1) = \Delta S \Big\} \end{aligned} \tag{181}$$

[32]The bound is *tight* in the sense that there exist some circumstances under which the inequality becomes an equality.

Note that we must require $d_{\mathcal{R}} \geq 2$ since with $d_{\mathcal{R}} = 1$ we get the trivial case that $\langle \Delta Q_{\mathcal{R}} \rangle = \langle \Delta S_{\mathcal{S}} \rangle = 0$.

The foundation for our proof of theorem 6.3 will be an auxiliary result about a mathematical property of *relative entropy*—from the same authors, Reeb and Wolf—expressed in corollary 6.1 below. [22]

Corollary 6.1. Consider of two density operators $\hat{\rho}$ and $\hat{\sigma}$, both associated with Hilbert spaces $\mathfrak{H}_d$ of finite dimensionality d. Their relative entropy $S(\hat{\sigma} \| \hat{\rho})$ is bounded from below by a *monotonically increasing* and *convex* function M that is defined as the minimization problem of *binary relative entropy*, $D_2(p,q)$, in a bounded region Ω in p-q space. See definition 6.1 for the functions H_2 and D_2.

$$\begin{gathered} S(\hat{\sigma} \| \hat{\rho}) \geq M_d(\Delta S) \\ \text{where} \\ \Delta S_N := S_N(\hat{\sigma}) - S_N(\hat{\rho}) \quad ; \quad d := \dim \hat{\rho} = \dim \hat{\sigma} \\ \Omega := 0 \leq p, q \leq \frac{d_{\mathcal{R}} - 1}{d_{\mathcal{R}}} \\ M_d(\Delta S) := \operatorname*{Min}_{\Omega} \left\{ D_2(p,q) \;\middle|\; H_2(p) - H_2(q) + (p-q)\ln(d-1) = \Delta S \right\} \end{gathered} \tag{182}$$

□

Clearly, corollary 6.1 looks familiar, and a lot of the heavy lifting is indeed accomplished by Reeb and Wolf in [22].

Proof (Theorem 6.3). Starting with equation (148) of the previous theorem 6.2, we have an exact expression of the bound in terms of the final mutual information between the sub-systems $I(\hat{\rho}'_{\mathcal{S}} : \hat{\rho}'_{\mathcal{R}})$ and the relative entropy in $\mathcal{R}$ between the initial and final state $S(\hat{\rho}'_{\mathcal{R}} \| \hat{\rho}_{\mathcal{R}})$.

$$\langle \Delta S_{\mathcal{S}} \rangle + \beta \langle \Delta Q_{\mathcal{R}} \rangle \;=\; I(\hat{\rho}'_{\mathcal{S}} : \hat{\rho}'_{\mathcal{R}}) + S(\hat{\rho}'_{\mathcal{R}} \| \hat{\rho}_{\mathcal{R}}) \tag{183}$$

Then, from equation (143) of theorem 6.1, we find that mutual information can be rewritten in terms of entropy changes, $I(\hat{\rho}'_{\mathcal{S}} : \hat{\rho}'_{\mathcal{R}}) = \langle \Delta S_{\mathcal{S}} \rangle + \langle \Delta S_{\mathcal{R}} \rangle$, and we can thus rewrite equation (183).

$$\beta \langle \Delta Q_{\mathcal{R}} \rangle - \langle \Delta S_{\mathcal{R}} \rangle \;=\; S(\hat{\rho}'_{\mathcal{R}} \| \hat{\rho}_{\mathcal{R}}) \tag{184}$$

From our definitions, see equation (138), we have $\langle \Delta S_{\mathcal{R}} \rangle := S_N(\hat{\rho}'_{\mathcal{R}}) - S_N(\hat{\rho}_{\mathcal{R}})$, and we can then use the result from [22], equation (182).

$$\beta \langle \Delta Q_{\mathcal{R}} \rangle - \langle \Delta S_{\mathcal{R}} \rangle \;\geq\; M_{d_{\mathcal{R}}}(\langle \Delta S_{\mathcal{R}} \rangle) \tag{185}$$

Again we turn to equation (143) of theorem 6.1, but now we make use of the inequality, $\langle \Delta S_{\mathcal{S}} \rangle + \langle \Delta S_{\mathcal{R}} \rangle \geq 0 \Rightarrow \langle \Delta S_{\mathcal{S}} \rangle \geq -\langle \Delta S_{\mathcal{R}} \rangle$, and since M is a monotonically increasing function—as shown in [22]—we can make the right-hand side (potentially) smaller, and the left-hand side (potentially) bigger, using this (non strict) inequality.

$$\beta \langle \Delta Q_{\mathcal{R}} \rangle + \langle \Delta S_{\mathcal{S}} \rangle \;\geq\; M_{d_{\mathcal{R}}}(-\langle \Delta S_{\mathcal{S}} \rangle) \tag{186}$$

Since we assumed $\langle \Delta S_{\mathcal{S}} \rangle < 0$ from the start, we have found the inequality of equation (179), theorem 6.3.

We have now not explicitly shown that M is, in fact, convex, monotonically increasing, and tight as a bound (in the sense that for any $\langle \Delta S_{\mathcal{S}} \rangle < 0$ there exist a $\hat{\rho}'_{\mathcal{C}}$ such that equality is assumed). Some of these discussions get quite involved, and we will not present them as proofs here, instead, the reader is referred to [22]. For convexity, see section 4.2, for monotonicity section 2.1 (remark 5) and for the tightness, section 2.1 (remark 3). ■

6.7 Comparing the Landauer heat bound with finite corrections to the classical counterpart

To gain some intuition about how the function M (theorem 6.3) behaves for different values of $\langle \Delta S_{\mathcal{S}} \rangle < 0$, we can plot the *reservoir's* change in heat $\beta \langle \Delta Q_{\mathcal{R}} \rangle$ against the change in the entropy of our *information bearing system*, $\langle \Delta S_{\mathcal{S}} \rangle < 0$. Note that we have used *von Neumann entropy* in this section, and it does not include Boltzmann's constant k, while the *Gibbs entropy* that we used in section 5 does. This implies that if we express $\langle \Delta S_{\mathcal{S}} \rangle$ in units of *bits*, it will have the same units as ΔH in equation (118) (see section 5.8.3).

Generally, it is safe to assume that the dimensionality of the *reservoir* $\mathcal{R}$ is larger than the *information bearing system* $\mathcal{S}$. We therefore begin by considering a qubit, $d_{\mathcal{S}} = 2$, which we reset by reservoirs of a larger number of dimensions, $d_{\mathcal{R}} = 32$ and $d_{\mathcal{R}} = 1024$.

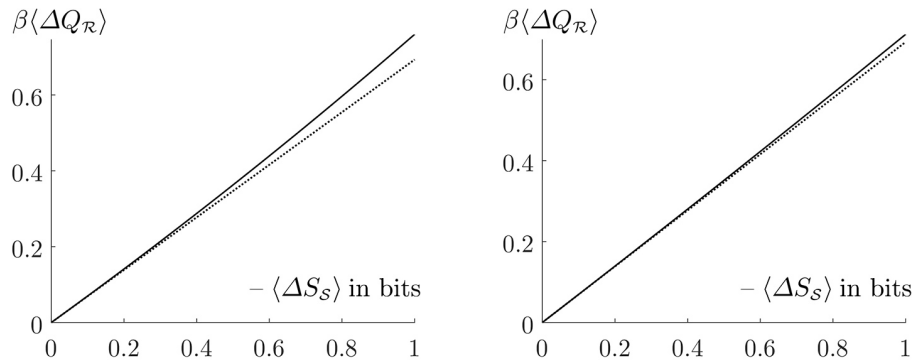


Figure 4: The change in *heat* in the *reservoir*, $\mathcal{R}$, when the entropy of a *qubit* $\mathcal{S}$ ($d_{\mathcal{S}} = 2$) is lowered. The dotted line is the classical linear bound (second law of Thermodynamics), and the solid line include the correction from the function M (theorem 6.3). In the left figure, the reservoir has dimensionality $d_{\mathcal{R}} = 32$, and in the right $d_{\mathcal{R}} = 1024$.

Clearly, as we increase the dimensionality of the reservoir, the behaviour becomes more and more like the linear bound supported by of the *second law of Thermodynamics* (section 3).

For a more complete picture, we compare this to the limiting case when $\mathcal{S}$ and $\mathcal{R}$ have the same number of dimensions. First, we let both the information

bearing system and the reservoir be qubits, $d_{\mathcal{S}} = d_{\mathcal{R}} = 2$. Then, we let both have a larger number of dimensions, $d_{\mathcal{S}} = d_{\mathcal{R}} = 1024$.

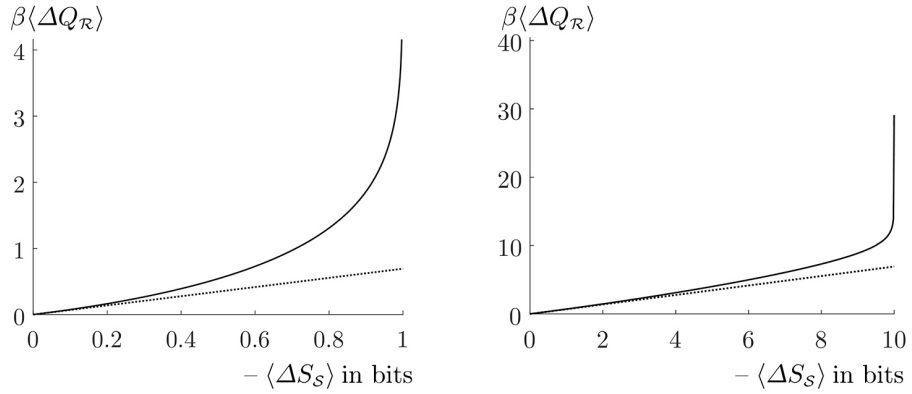


Figure 5: The change in *heat* in the *reservoir*, $\mathcal{R}$, when the entropy of an information bearing system $\mathcal{S}$ is lowered. The dotted line is the classical linear bound (second law of Thermodynamics), and the solid line include the correction from the function M (theorem 6.3). In the left figure both systems ($\mathcal{S}$ and $\mathcal{R}$) are qubits, $d_{\mathcal{S}} = d_{\mathcal{R}} = 2$, and in the right both systems have larger dimensionality, $d_{\mathcal{S}} = d_{\mathcal{R}} = 1024$.

Clearly, if $\mathcal{S}$ and $\mathcal{R}$ have the same dimensionality, the process becomes much less efficient, with the largest deviation from linear behaviour if we want to take $\mathcal{S}$ from a maximally mixed state to a pure state.

7 Objection to "Thermodynamical costs of some interpretations of quantum theory"

The initial motivation to consider *Landauer's principle* was its application in a 2016 paper by Adán Cabello, Mile Gu, Otfried Gühne, Jan-Åke Larsson, and Karoline Wiesner, titled "Thermodynamical costs of some interpretations of quantum theory" [2]. The paper enters into an ambitious enterprise, to attach an experimentally testable prediction to the choice of interpretation for Quantum Mechanics.

In June 2017 (after this work was initiated) a paper by Carina E. A. Prunkl and Christopher G. Timpson was published [23], where problems with the argument by Cabello et al. was pointed out. However, this latter publication was not discovered before its key point was independently reproduced. Even though both objections are in agreement, what is emphasized here diverges somewhat from Prunkl and Timpson.

In this section, we discuss the set-up and argument of Cabello et al. and reproduce their result (section 7.1). Then we point out what the problem of their argument is and offer a replacement argument (section 7.2). Our emphasis will be somewhat different from Prunkl and Timpson, and we deviate from their analysis in one respect (section 7.2.1). Note that the reader familiar with the details of the analysis by Cabello et al. may prefer to skip ahead to section 7.2.

7.1 Reproducing the result from Cabello et al.

7.1.1 Classification of interpretations

The approach by Cabello et al. is to consider a thought experiment, from which seemingly unphysical consequence is derived. They begin by dividing the current set of described interpretations into two classes—type I and II—on the basis of their respective attitude towards the *origin of probabilities* in Quantum Mechanics, as defined below (citing [2] with some clarifying modification of language).

Definition 7.1 (Type I). Type I interpretations consider *quantum probabilities for measurement outcomes* as determined by "intrinsic" (observer-independent) properties of the observed system.

Definition 7.2 (Type II). Type II interpretations consider *quantum probabilities for measurement outcomes* as describing the experiences that the observer has of the observed system, unrelated to "intrinsic" properties of the system.

A number of popular interpretations are then classified in terms of these two definitions. According to Cabello et al., the following interpretations are classified as type I: Einstein's [24], Bohmian mechanics [25,26], many worlds [27, 28], Ballentine's [29], modal interpretations [30,31], Bell's beables [32], collapse theories [33,34], and Spekkens' [35].

In type II interpretations we have: Copenhagen [36, 37], Wheeler's [38], relational [39], Zeilinger's [40], Fuchs and Peres's no interpretation [41], and QBism [42,43].[33]

We note that Prunkl and Timpson argue that classifying interpretations based on the definition of type I and II is not completely unambiguous [23]. We shall however not discuss properties of different interpretations, but instead, focus on the technical aspects of the argument by Cabello et al.

7.1.2 Premises and the considered quantum system

The system we shall consider is comprised of three parts: A single qubit encoded in some physical system with two energy levels, such as a spin-half particle. Second, a measurement device for measuring the spin of the qubit-system along some arbitrary axis. Finally, we have some generator of randomness.

A sequence of measurements is then carried out on the qubit-system, where the measurement direction is determined by the randomness generator, and drawn from a finite set. This set of available measurement directions is chosen such that possible *measurement outcomes* will be isotropically distributed in the x-z plane. Three premises are then assumed to hold.

Premise 7.1. Which measurement is performed on the qubit-system can be chosen randomly and independently of the system. □

Premise 7.2. The qubit-system has limited *memory*. □

Premise 7.3. Landauer's principle holds. □

The eventual intention is to look at the behaviour of this setup when the number of measurement directions grows.

7.1.3 Mathematical structure

The central idea of the paper is that the outcomes from measurements on the qubit-system can be modelled as a *stochastic input-output process*, generated by a so-called *epsilon-transducer*. We will here go through a compact[34] discussion of the necessary concepts and mathematical structure, using *random variables*[35].

[33] Citations for each interpretation are obtained from Cabello et al. [2].

[34] Reader beware. We will not go into enough details for the discussion to be considered self-contained. Some familiarity with *random variables* is assumed. Complementary, but brief, discussions can be found in [2] and [23]. Papers such as [44] discuss some concepts in more detail, but it lacks a prominent pedagogical approach.

[35] A *random variable*, X, is a map between the outcomes of some *process* and an abstract set of labels $\mathcal{X}$, called an *alphabet*. Any particular value which the random variable X can assume is denoted by a lower case letter, $x \in \mathcal{X}$. The *process* underlying X has some random quality (often caused by some physical parameters that are not well understood) that assigns a probability for each outcome in the alphabet, $P(X{=}x)\ \forall\ x \in \mathcal{X}$, often denoted more compact as $P(X)$.

A *stochastic process*, denoted $\overleftrightarrow{Y}$, is a countably infinite set of time-ordered *random variables*.

$$\overleftrightarrow{Y} := \cdots, Y_{-2}, Y_{-1}, Y_0, Y_1, Y_2, \cdots \equiv \{Y_t \,|\, t \in \mathbb{Z}\} \tag{187}$$

Each random variable Y_t can assume values from the countable *alphabet* $\mathcal{Y}$, with a probability distribution $\{P_t(y)\}$ over the members $y \in \mathcal{Y}$, denoted in a more compact notation as $P(Y_t)$. For our purposes, it will be sufficient to consider *finite* alphabets $\mathcal{Y}$, but in principle, they can be countably infinite. We shall use this *stochastic process*, $\overleftrightarrow{Y}$, to represent *outcomes* or *output*, when we take measurements on the qubit-system.

To construct an *input-output process* we pair the *output* $\overleftrightarrow{Y}$ with a similar construct for the *input*, $\overleftrightarrow{X}$, which represents the randomly chosen measurement directions in the x-z plane.

$$\overleftrightarrow{X} := \cdots, X_{-2}, X_{-1}, X_0, X_1, X_2, \cdots \equiv \{X_t \,|\, t \in \mathbb{Z}\} \tag{188}$$

Similarly to the previous case, X_t is a *random variable* that can assume values from the finite alphabet $\mathcal{X}$, with some probability distribution $P(X_t)$.

Taken together $(\overleftrightarrow{X}, \overleftrightarrow{Y})$ represents an *input-output process* where we also impose the condition that the choice of measurement direction, X_t, will affect probabilities of measurement outcomes, thus $P(Y_t)$ becomes $P(Y_t|X_t)$.

Then Cabello et al. argue that it is appropriate to represent the *input-output process* $(\overleftrightarrow{X}, \overleftrightarrow{Y})$ by a minimal representation machine—an *epsilon-transducer* [44]. A central idea is to introduce a (minimal) set $\mathcal{S}$ of so-called *causal states*, with the purpose to find the minimum amount of information needed to be stored in the *epsilon-transducer* in order to accurately mirror the future statistical behaviour of the qubit-system in this *input-output process*. To the causal states $\mathcal{S}$ we also attach a corresponding *random variable* S_t that assumes values from the alphabet $\mathcal{S}$ with some probabilities $P(S_t)$.

Since the causal states are going to distinguish past and future behaviour we introduce a notation to distinguish between the past and future results of the *input-output process*.

$$\overleftarrow{X} := \cdots, X_{-3}, X_{-2}, X_{-1} \quad ; \quad \overrightarrow{X} := X_0, X_1, X_2, \cdots \tag{189}$$

$$\overleftarrow{Y} := \cdots, Y_{-3}, Y_{-2}, Y_{-1} \quad ; \quad \overrightarrow{Y} := Y_0, Y_1, Y_2, \cdots \tag{190}$$

Formally, the *causal states* $\mathcal{S}$ are expressed as a partition of the set of *input-output pasts* $(\overleftarrow{X}, \overleftarrow{Y})$ by an *equivalence class*. In this equivalence class, two members $(\overleftarrow{x}, \overleftarrow{y})$ and $(\overleftarrow{x}', \overleftarrow{y}')$ are equivalent if and only if the probabilities for future outcomes $\overrightarrow{Y}$, as a distribution over different future measurement directions $\overrightarrow{X}$, are equal. Expressed formally as follows.

$$\begin{gathered} P\big(\overrightarrow{Y} \,\big|\, \overrightarrow{X}, \overleftarrow{X} = \overleftarrow{x}, \overleftarrow{Y} = \overleftarrow{y}\big) = P\big(\overrightarrow{Y} \,\big|\, \overrightarrow{X}, \overleftarrow{X} = \overleftarrow{x}', \overleftarrow{Y} = \overleftarrow{y}'\big) \\ \Rightarrow \\ (\overleftarrow{x}, \overleftarrow{y}) \equiv (\overleftarrow{x}', \overleftarrow{y}') \end{gathered} \tag{191}$$

Each member in the equivalence class partition of $(\overleftarrow{X}, \overleftarrow{Y})$ corresponds to a member in the *set of causal states*, $\mathcal{S}$.

Finally, this construction also includes a set of conditional *transition probabilities* $\mathcal{T}_P$ that governs the transitions between the causal states in $\mathcal{S}$. These

transition probabilities depend on the current measurement direction X_t, and of course the current causal state S_t.[36]

$$\mathcal{T}_P := \left\{ P(S_{t+1}{=}s' \,|\, S_t{=}s, X_t{=}x) \right\} \quad \text{where} \quad s, s' \in \mathcal{S} \quad ; \quad x \in \mathcal{X} \tag{192}$$

These transition probabilities $\mathcal{T}_P$ induces a *stationary probability distribution* $\{P(s)\}$ over the causal states $s \in \mathcal{S}$, or in compact notation, $P(S)$. The distribution is called *stationary* since it is taken as a time-independent average over the time parameter t for the probability of each $s \in \mathcal{S}$. The *Shannon entropy* of the *stationary probability distribution*, $H(\{P(s)\}) =: H(\mathcal{S})$, is called the *statistical complexity*, and it represents the minimal amount of information which our *transducer* has to be able to encode in order to produce the desired behaviour. [2]

At this point, it is appropriate to mention that the mathematical machinery presented so far is more powerful than what we actually require for the rest of the argument. Thus, some constructs which can be convoluted when presented in the general case will become clear as we introduce the simplifying circumstances of our particular problem—taking random measurements in the x-z plane on a qubit-system.

7.1.4 Information erasure in the causal states, $\mathcal{S}$

In premise 7.2 we assumed that the qubit-system—as modelled with our *epsilon-transducer*—does not have infinite memory, i.e. it cannot retain all past information about previous *causal states*, so at some point, it has to start *erasing* information. Thus, when the machine is up and running, and has reached its limit in terms of information storage—the average information to be *erased* equals the average amount of information *produced* at each time-step. We can therefore equate these two quantities.

The minimum (average) amount of information produced at each time step is how much entropy remains in S_{t-1}, if we average over all possible measurement directions, measurement outcomes, and causal states; X_t, Y_t, and S_t respectively. This is calculated using *conditional entropy* (see section A.9). Then, let $\langle I_e \rangle$ denote the average amount of information erased at each time step.

$$\langle I_e \rangle = H\big(S_{t-1} \big| X_t, Y_t, S_t\big) \tag{193}$$

In section 7.1.6 we will simplify this and show that for our particular circumstances it is sufficient to only consider a single current causal state, i.e. $\langle I_e \rangle = H(S_{t-1}|S_t = s) \ \forall\, s \in \mathcal{S}(n)$.[37]

7.1.5 Finding the alphabet sets, $\mathcal{X}(n)$, $\mathcal{Y}(n)$, and $\mathcal{S}(n)$

We intend for our *measurement outcomes* of the qubit-system to be isotropically distributed in the x-z plane. Thus we parameterize the x-z plane with an angle θ,

[36] In [2], the transition probabilities $\mathcal{T}_P$ is described as a joint probability distribution over *both* S_t and Y_t, but the simplification made in equation (192) is, in fact, sufficient.

[37] Cabello et al. also includes a condition on the measurement direction X_t, i.e. there $\langle I_e \rangle = H(S_{t-1}|X_t = x, S_t = s)$.

over an open interval that corresponds to a semicircle, where each *measurement outcome* can be either up or down in each *measurement direction*.

$$\theta \in [0, \pi) \tag{194}$$

This interval is then divided into discrete steps of equal size. It turns out that having the number of measurement directions be a power of two, 2^n, will be convenient for later calculations. Thus we define our set of possible measurement directions, a.k.a. the *input alphabet*, $\mathcal{X}(n)$, as follows.

$$\theta = \frac{\pi k}{2^n} \quad ; \quad k \in \{0, 1, \cdots, 2^n - 1\} \quad \Rightarrow \tag{195}$$

$$\mathcal{X}(n) = \left\{ \cos\left(\frac{\pi k}{2^n}\right) \sigma_z + \sin\left(\frac{\pi k}{2^n}\right) \sigma_x \ ; \ k \in \{0, 1, \cdots, 2^n - 1\} \right\} \tag{196}$$

To define an output alphabet $\mathcal{Y}(n)$ we can use the set of all possible quantum states after measurement as labels, and we shall express the states in terms of the eigenstates of the σ_z operator, $|0\rangle$ and $|1\rangle$. Note that for each measurement direction in $\mathcal{X}(n)$ we have two possible outcomes, +1 or –1, thus $|\mathcal{Y}(n)| = 2|\mathcal{X}(n)|$.

$$|0\rangle \quad \Leftrightarrow \quad \text{measuring +1 in the } z\text{-direction} \tag{197}$$

$$|1\rangle \quad \Leftrightarrow \quad \text{measuring –1 in the } z\text{-direction} \tag{198}$$

A state that corresponds to +1 when measured in the x-z plane, at the angle $\chi \in [0, \pi)$ from the z axis, has the following expression.

$$|\psi_\chi\rangle = \cos\left(\frac{\chi}{2}\right)|1\rangle + \sin\left(\frac{\chi}{2}\right)|0\rangle \tag{199}$$

Then, we can expand the angle $\chi \in [0, 2\pi)$ to also include measurements of –1, and parameterize it similar to θ in equation (195).

$$\chi = \frac{\pi j}{2^n} \quad ; \quad j \in \{0, 1, \cdots, 2^{n+1} - 1\} \tag{200}$$

Note that j runs over twice as many values compared to k from equation (195), reflecting the fact that each measurement direction has two outcomes. We then insert this parametrization into equation (199) to find our set of all possible states for the qubit-system.

$$\mathcal{Y}(n) = \left\{ \cos\left(\frac{\pi j}{2^{n+1}}\right)|0\rangle + \sin\left(\frac{\pi j}{2^{n+1}}\right)|1\rangle \ ; \ j \in \{0, \cdots, 2^{n+1} - 1\} \right\} \tag{201}$$

Since the last measurement outcome is sufficient to predict probabilities of future measurements in any direction, we can use the same set as for $\mathcal{Y}(n)$ to label our *causal states*, $\mathcal{S}(n)$.

$$\mathcal{S}(n) = \left\{ \cos\left(\frac{\pi j}{2^{n+1}}\right)|0\rangle + \sin\left(\frac{\pi j}{2^{n+1}}\right)|1\rangle \ ; \ j \in \{0, \cdots, 2^{n+1} - 1\} \right\} \tag{202}$$

Note that *the outputs* Y_t, and *the causal states* S_t, are just the possible physical pure states of the qubit-system. Thus there exist a bijective map (one-to-one) between Y_t, S_t, and the physical state of the system, where knowing one will determine the others.

7.1.6 Calculation

From equation (193) we know that the amount of information that is erased is the entropy in the *causal state* before the measurement S_{t-1}, that we cannot learn from S_t, X_t, and Y_t, after the measurement.

$$\langle I_e \rangle = H\big(S_{t-1}\big|X_t, Y_t, S_t\big) \tag{203}$$

Since there exists a bijective map between Y_t and S_t, no additional information is contained in *the output* Y_t if we have considered *the causal state* S_t. This means that $H(Y_t|S_t) = 0$, and Y_t in equation (203) is redundant.

$$\langle I_e \rangle = H\big(S_{t-1}\big|X_t, S_t\big) \tag{204}$$

Next, there is also a map between S_t and X_t. If we know the causal state S_t, we know the physical state of the system, and thus we know what measurement direction was used, X_t. This means that $H(X_t|S_t) = 0$, and X_t in equation (203) is also redundant.

$$\langle I_e \rangle = H\big(S_{t-1}\big|S_t\big) \tag{205}$$

One further simplification can be made from realizing that the problem is symmetric under discrete rotations in the x-z plane, by any angle from the parameterization of the measurement outcomes, see equation (200). This is because the measurement outcomes are *uniformly* distributed, and the measurement directions are chosen from a *uniform* probability distribution. This rotational symmetry implies that the *conditional entropy* in S_{t-1} will be identical for any particular choice of *current causal state* $S_t = s$, and we do not have to sum over all $s \in \mathcal{S}(n)$.

$$\langle I_e \rangle = H\big(S_{t-1}\big|S_t{=}s\big) \quad \forall\ s \in \mathcal{S}(n) \tag{206}$$

In the appendix, section A.9, we find an explicit expression for *conditional entropy* in terms of *conditional probability*—see equation (274). Since the choice of current causal state, $S_t = s$, does not affect the calculation, we simply select one that is easy to deal with, $S_t = |0\rangle$.

$$\langle I_e \rangle = -\sum_{s_{t-1}\in\mathcal{S}(n)} P\big(S_{t-1}{=}s_{t-1}\big|S_{t-1}{=}|0\rangle\big)\,\log_2\Big(P\big(S_{t-1}{=}s_{t-1}\big|S_{t-1}{=}|0\rangle\big)\Big) \tag{207}$$

The probability $P(S_{t-1}{=}s_{t-1}|S_{t-1}{=}|0\rangle)$ is calculated from multiplying two probabilities. For the causal state s_{t-1} we have the probability that the appropriate measurement direction was chosen, then multiplied by the probability of a transition between the appropriate quantum states, from the Born rule. There are 2^n measurement directions, all with equal probability $1/2^n$. The initial quantum state is assumed to be $|0\rangle$, and the set of all final states is found in equation (202).

$$\begin{aligned} &P(S_{t-1}{=}s_{t-1}|S_t{=}|0\rangle) = \\ &= \frac{1}{2^n}\left|\left(\cos\left(\frac{\pi j}{2^{n+1}}\right)\langle 0| + \sin\left(\frac{\pi j}{2^{n+1}}\right)\langle 1|\right)|0\rangle\right|^2 = \frac{\cos^2\left(\frac{\pi j}{2^{n+1}}\right)}{2^n} \end{aligned} \tag{208}$$

We insert this into equation (207), and from equation (202), we can see that a sum over all $s_{t-1} \in \mathcal{S}(n)$ corresponds to the a sum over $j \in \{0, \cdots, 2^{n+1}-1\}$.

$$\langle I_e \rangle = -\sum_{j=0}^{2^{n+1}-1} \frac{\cos^2\left(\frac{\pi j}{2^{n+1}}\right)}{2^n}\,\log_2\left(\frac{\cos^2\left(\frac{\pi j}{2^{n+1}}\right)}{2^n}\right) \tag{209}$$

To deal with this we will make some approximations. First, we incorporate the minus sign into the logarithm and use that $\cos^2\theta < 1$ for $0 < \theta < \pi$.

$$\langle I_e \rangle = \sum_{j=0}^{2^{n+1}-1} \frac{\cos^2\left(\frac{\pi j}{2^{n+1}}\right)}{2^n} \log_2\left(\frac{2^n}{\cos^2\left(\frac{\pi j}{2^{n+1}}\right)}\right) > \tag{210}$$

$$> \frac{1}{2^n} \sum_{j=0}^{2^{n+1}-1} \cos^2\left(\frac{\pi j}{2^{n+1}}\right) \log_2\left(2^n\right) = \frac{n}{2^n} \sum_{j=0}^{2^{n+1}-1} \cos^2\left(\frac{\pi j}{2^{n+1}}\right) \tag{211}$$

Since we are interested the behaviour of the system as the number of measurement directions 2^n grows, for large n we can convert the sum into an integral by defining $\theta := \pi j/2^{n+1}$, and introduce an interval $\pi/2^{n+1} \to \mathrm{d}\theta$.

$$\langle I_e \rangle > \frac{n}{2^n} \frac{2^{n+1}}{\pi} \sum_{j=0}^{2^{n+1}-1} \cos^2\left(\frac{\pi j}{2^{n+1}}\right) \frac{\pi}{2^{n+1}} \to \tag{212}$$

$$\to \frac{2n}{\pi} \int_0^\pi \cos^2\theta \, \mathrm{d}\theta = \frac{2n}{\pi}\frac{\pi}{2} = n \tag{213}$$

We have thus found an approximate lower bound on the average information erased from the *epsilon-transducer* for each measurement, $\langle I_e \rangle$.

$$\langle I_e \rangle > n \qquad \text{(for large } n\text{)} \tag{214}$$

7.1.7 Unphysical consequence from Landauer's principle

Cabello et al. then applies the colloquial formulation of Landauer's principle (see section 5.2) that "erasure of one bit of information requires a net increase of heat corresponding to $kT \ln 2$".

Under the reasonable assumption that we operate in some temperature $T > 0$, Landauer's principle implies that a finite amount of heat—which is bounded from below—must be dissipated in each measurement, and that the lower bound will tends to infinity linearly with n (where 2^n is the number of measurement directions).

$$\langle \Delta Q \rangle \geq n\, k\, T\, \ln 2 \tag{215}$$

In conclusion, as the number of measurement directions tend to infinity, so does the amount of heat that *necessarily* must be released in each measurement. Such necessary and unbounded heat dissipations appear unphysical, or at least, it is an experimentally testable claim.

7.2 Refuting the result from Cabello et al.

The counterargument to the result of Cabello et al. originates from an understanding of how Landauer's principle is conceived and formalized (discussed in detail through sections 5 and 6).

The classical variety (section 5) introduce a set-up where we *encode* logical states in the physical states of some system, and attempts to draw conclusions about the *physical irreversibility* of certain logical operations. In section 5 we

showed that with this set-up, it is possible to argue for the *physical irreversibility* proposed by Landauer's principle (proposition 5.1), though there are some required assumptions and caveats (see sections 5.7 and 5.9.1).

In this classical setting (again section 5), we partition some set of *mutually exclusive* physical microstates $\{\mu\}$ into logical states, and entropies are calculated from the probability distributions over the microstates $\{\mu\}$. We note that it would not be possible to relax the condition that microstates are mutually exclusive, since *entropy*, in its most abstract mathematical form, is defined from probabilities over a set of *mutually exclusive events* (see section 3.5).

When we later generalize these principles to quantum systems (section 6) we replace our *mutually exclusive* microstates $\{\mu\}$ by a set of basis states $\{|\mu\rangle\}$[38], where the previous condition of mutual exclusivity is transformed into the condition that the basis-states are *orthogonal*. In this framework, *von Neumann entropy* S_N (see section 3.7) can be motivated as the basis for calculating entropies in physical quantum systems, and thus becomes the basis for generalizing the *Landauer bound* and *Landauer's principle* (see sections 5.8 and 5.9) to quantum systems (as we have begun to do in section 6).

$$S_N(\hat{\rho}) := -\mathfrak{Tr}\left[\hat{\rho}\ln\hat{\rho}\right] \tag{216}$$

Thus we should calculate entropy in a quantum system from the density matrix, according to equation (216), and the calculation made by Cabello et al. appear conceptually inaccurate.

When we consider a qubit-system in particular, we only need two orthogonal states to span its Hilbert space. This implies that entropy in this qubit-system is, not only bounded from below by 0, but also bounded from above by 1 bit.[39] Clearly, the maximum *change* in entropy is therefore bounded by 1 bit, implying that the lower bound than Cabello et al. suggested, from equation (215), cannot go to infinity. In fact, any lower bound cannot go above $k\,T\,\ln 2$.

$$\langle \Delta Q\rangle \geq k\,T\,\ln 2 \tag{217}$$

The incorrect conclusion arises for Cabello et al. because the entropy is calculated over the set of *causal states* $\mathcal{S}(n)$ of the *epsilon-transducer*, where the set of causal states *does not* correspond to some set of orthogonal states (as is required when we generalize from *mutually exclusive* states) instead the physical states corresponding to $\mathcal{S}(n)$ are strongly linearly dependent—see equation (202).

In conclusion, it is premise 7.3 (see section 7.1.2) that fails. Not because *Landauer's principle* does not hold, but because it does not *apply* to decreases of entropy associated with *causal states* in this particular set-up. See paragraph (d) in section V. of [2], where Cabello et al. presents a short discussion related to this point.

7.2.1 Substitute argument

Replacing the argument made by Cabello et al. requires much less work than we went through in section 7.1.

[38] Technically, we use the density operator $\hat{\rho}$ that can be considered without choosing a basis, but when the entropy is evaluated we choose some *basis of states* to calculate the trace (further details in sections 3.7 and A.4).

[39] Left to the reader as an exercise.

Because we are making *measurements* on the qubit-system—from the point of view of the measurement device (which is the system interacting with the qubit), the state is always pure after each measurement. Therefore we here argue that the entropy should be calculated from a pure state, becoming zero at each time step, and hence the change in entropy is zero. Then, *Landauer's principle* does no longer demand any minimal heat expenditure, and the difference between the two classes of interpretations of Quantum Mechanics, type I and type II, goes away.

$$\langle \Delta Q \rangle \geq 0 \tag{218}$$

We note that Prunkl and Timpson [23] makes a slightly different argument, claiming that: "It is clear that once the measurement process is up and running, the quantum system will be in a maximally mixed state at each time step, independent of the chosen measurement basis." But here, we argue that a pure state is more appropriate to model the qubit-system after a measurement. But regardless of whether we consider the state to remain pure, or maximally mixed—there is no change of entropy in either approach. We therefore arrive at the same conclusion—that we unfortunately cannot infer any lower bound on heat expenditure from the choice of interpretation of Quantum Mechanics.

8 Conclusions

Since each section in this thesis essentially presents an independent point, there are few all-encompassing conclusions left to discuss. Instead, we shall focus on the question of how some sections (4, 5 and 6) could be further explored in their own right.

We then indulge in some grandiose speculations about the fundamental question that Cabello et al. intended to address (see section 7)—the *measurement problem*—and we conclude this thesis with a proposition for third additional classification—not described by Cabello et al.—for the origin of probabilities in Quantum Mechanics.

8.1 Further work

There are several places in this thesis that present opportunities to conduct further research into various interesting directions. Ultimately, time constraints put boundaries on what lines of thought could be pursued.

8.1.1 Section 4 – The principle of maximum entropy inference

Extension. In section 4.1 we envision a stricter argument to motivate the use of the *principle of maximum entropy inference*. However, this presentation is not complete without some examples of what the *biased assumption* C' will represent in some actual case (see section 4.1). If there is a demand for a stricter motivation of the principle of maximum entropy inference, investigating this line of thought can be an interesting pursuit.

8.1.2 Section 5 – Landauer's principle in Classical Physics

There is still work to be done before consensus about Landauer's principle can be attained. In section 5 we hope the reader can find a contribution to this debate, but it is by no means resolved.

There are a number of different ways to extend the argument. The most obvious is to base the argument on more general assumptions (section 5.3). Clearly, this comes with the risk of obfuscating the core conceptual points, but the whole argument could become applicable in a broader range of situations, and thus more worthwhile to communicate to a larger community.

Extension I. We could distinguish logical states for *input* and *output*, and then allow for *alphabet* sets of any finite size, i.e. $\{\mathtt{0},\mathtt{1}\} \to \{\mathtt{0},\mathtt{1},\cdots,\mathtt{n}\}$ in section 5.3.1. That way the analysis will encompass the logic of any kind of *gate* and not just manipulation of individual bits. □

Extension II. We could have expanded the argument to include *logical operations* that are inherently random to some degree (compare to section 5.3.2) as Maroney does this in his 2009 paper [20]. The consequences of this modification are not trivial to predict, since Maroney's treatment deviates from the structure of our argument (see section 5.13). □

Extension III. It would be desirable to investigate in greater detail the implications of requiring a process to be *entropy restoring*, or conforming to *uniform computing* (see section 5.7). In particular, to compare the constraints with some concrete computing devices, and look for examples of physical processes that would adhere to, or break, the assumptions. A desirable claim would be that *all* computation devices operate under some version of *uniform computing* where Landauer's principle can be derived. □

Extension IV. Finally, the argument in section 5.9 is rather subtle and could be weak to counterarguments—as pointed out in section 5.9.1. It would be beneficial to collect critical peer evaluation to be able to strengthen weak points and address potential concerns. □

8.1.3 Section 6 – A Landauer bound in Quantum Mechanics

Extension I. It is not obvious whether any density operator associated with a finite-dimensional Hilbert space is sufficient for faithfully modelling a thermal reservoir. A required property of any reservoir is to be large enough for its temperature (here reciprocal temperature β) to stay constant under interactions with some system of interest. Clearly, this is not guaranteed to be the case when systems are exchanging entropy and energy, since temperature is defined in terms of these.

$$\beta := \frac{1}{k} \frac{\partial S}{\partial \langle E \rangle} \tag{219}$$

The first step to create a realistic model is to introduce a compact spectrum of numerous energy eigenstates for the Hamiltonian of the thermal reservoir. We may also want to consider a *separable* Hilbert space (see section 2). □

Extension II. The motivation Reeb and Wolf supplies for why energy exchanges between system and reservoir are considered only as *heat* and never as *work* (see section 6.2), may not be entirely conclusive. More careful investigations into how to define *heat* and *work* in Hamiltonian interactions may reveal some caveats. □

Extension III. To begin extending section 6 to the scope of section 5, we would need to define *logical states* in terms of the underlying physical state, $\hat{\rho}_{\mathcal{S}}$ (compare to section 5.3.1). Then we can encode *classical information* in our state for $\mathcal{S}$, and think about what kind of Hamiltonian on the entire closed system $\mathcal{C}$ can compress the state space of $\mathcal{S}$ in order to recreate the *logical reset process* ($\mathcal{P}^0_{\mathcal{S}}$ from section 5.3.8). □

8.2 The measurement problem

Since the initial motivation of this thesis was to investigate the role of probabilities in Quantum Mechanics, we shall briefly discuss some of the author's personal attitudes in relation to the *measurement problem*. Naturally, this section contains some wild speculations, reflecting the attitude of the author at the time of writing. Towards the end, we will argue for a third additional classification for the origin of probabilities, not discussed by Cabello et al. [2].

In short, the *measurement problem* stems from the conflict between the superposition principle in Quantum Mechanics—allowing systems to be in any

superposition of states—and the limitation of a much smaller state space in Classical Physics, where no superpositions exist. It is clear that the quantum description of the world should be considered more fundamental than classical theories, but still, on macroscopic scales we observe a behaviour that appears restricted to the classical framework. [45]

At the heart of this conflicts sits the "collapse of the wave function", i.e. an event wherein the unitary evolution of quantum systems no longer agrees with experiments. Shoot an electron at a double slit setup, and the Schrödinger equation predicts the evolution perfectly, up until the point when the particle hits the detection screen, and a superposition of positions can no longer be an acceptable *experimental outcome*, instead the electron has to end up at some *particular* position with some probabilities determined from the *Born rule*. We label these unitarity-breaking situation as "measurements", and allow for a different *projective evolution* to take place. Colloquially, and in the language of the *Copenhagen interpretation*, we can say that *the wave function collapses*.

But there is no consensually agreed-upon formal definition of what actually counts as a *measurement*, and there is no agreed-upon *mechanism* for a state to go from the superposition of outcomes, to one definite outcome. The latter problem is termed the *problem of definite outcomes*, it only constitutes half of the *measurement problem* [45], and it will be our focus here. The other half, the *problem of the preferred basis*, will not be discussed.

8.2.1 Interpretation versus explanation

There is a multitude of *interpretations* of Quantum Mechanics, trying to account for the measurement problem. While some interpretations provide accounts on a more philosophical level—supplying some "attitude" towards the measurement problem, without providing any testable predictions. Here we shall argue that we ought to be looking for an "explanation"—meaning some set of ideas that (at least in principle) can produce testable predictions. The *decoherence program* [46], aligns somewhat with such ambitions (expressing a part of the process in terms of *physical interactions*), but here we will argue that the idea fails to provide a complete physical description of a transition to from a decohered state to a definite measurement outcome.

Then, lacking a completely satisfactory physical model we can apply an "interpretational band-aid", and say that the world splits into a myriad of different outcomes (the *many-worlds interpretation*). But such *interpretations* do not supply testable predictions (the worlds are assumed to be fundamentally inaccessible after an *outcome* is registered).

8.2.2 Proposition for a third additional class of interpretations of Quantum Mechanics

If one consider *von Neumann entropy* S_N (section 3.7) as a quantum mechanical counterpart of thermodynamic entropy, to which something like the second law of thermodynamics (axiom 3.4) should apply, and then adopts the framework of decoherence [46], it is possible to argue that quantum states evolve from an initial pure state with low entropy, to a mixed state with some classical probability distribution of different outcomes, having high entropy. However,

actual measurement outcomes are of course some particular outcome from the probability distribution, with low entropy again.

$$\text{Initial pure state} \longleftrightarrow S = 0$$

$$\text{State after decoherrence} \longleftrightarrow S > 0$$

$$\text{Single measurement outcome} \longleftrightarrow S = 0$$

The transition from a pure state to a mixed one is not that problematic—it rhymes with the second law of thermodynamics, which allows for entropy to increase. However, the transition from a decohered mixed state to a pure measurement outcome is more problematic since it should obey a *Landauer bound*. Whether we say that the bound should be upheld due to the *second law of thermodynamics* (axiom 3.4), or we simply consider a derivation as the one carried out in section 6, the entropy of the total system should not decrease. Thus when the entropy of the state decreases, it is not so clear where the complementary increase should take place.

The proposition here is that during decoherence, we do not expect there to be some increase in the number of quantum states that are somehow occupied by the system, but instead it is our *ignorance* about the actual state that increases. Or put differently, we expect the state to *always* be pure, and the increase in entropy indicates that we have lost some information about which pure state is occupied at all times.

This can seem strange since we had full information about the state to begin with, but we can argue that the uncertainty is introduced as soon as the particle starts interacting with the measurement device—a macroscopic object to which we have no chance of knowing its precise state and Hamiltonian. Therefore their joint evolution inevitably contains some uncertainty which spills over to the quantum system as a mixed state.

The *conjecture* is therefore that—given the actual pure state for the entire measurement device, and the time-dependent Hamiltonian for the combined system at all times, we should *in principle* be able to predict what the measurement outcome will be, however no model of an interaction between a quantum system, and a macroscopic object big enough to be a measurement device (completely modelled in Quantum Mechanics), has yet been analyzed in enough detail.

Comparing this idea with the two classes proposed by Cabello et al. (see section 7.1.1) it is clear that we can conjecture an additional third class.

Definition 8.1 (Type III). Type III *explanations* consider *quantum probabilities for measurement outcomes* as originating from *partially unknown* interactions with, typically large, *external* systems in *unknown* states. □

8.3 Acknowledgements

I would like to thank my supervisor Gunnar Björk for his assistance and keen eye for interesting questions. I would also like to thank the assisting supervisors, Jonas Larson, and Supriya Krishnamurthy for their guidance. I am particularly grateful for the generosity with which Supriya offered her time to discuss ideas and provide feedback. Thank you all!

9 References

[1] R. Landauer, "Irreversibility and heat generation in the computing process", IBM Journal of Research and Development, vol. 5, no. 3, pp. 183–191, July 1961. https://ieeexplore.ieee.org/document/5392446/

[2] A. Cabello, M. Gu, O. Gühne, J.-Å. Larsson, and K. Wiesner, "Thermodynamical cost of some interpretations of quantum theory", Phys. Rev. A, vol. 94, p. 052127, Nov 2016. https://link.aps.org/doi/10.1103/PhysRevA.94.052127

[3] C. H. Bennett, "Demons, engines and the second law", Scientific American, vol. 257, no. 5, pp. 108–117, 1987. http://www.jstor.org/stable/24979551

[4] E. T. Jaynes, "The second law as physical fact and as human inference", unpublished manuscript, 1998. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.40.8986

[5] R. Frigg and C. Werndl, "Entropy – a guide for the perplexed", Probabilities in Physics, Oxford University Press, 2011. http://philsci-archive.pitt.edu/8592/

[6] L. Landau and E. Lifshitz, "Statistical physics". Elsevier Science, 2013, vol. 5. https://books.google.se/books?id=VzgJN-XPTRsC

[7] E. T. Jaynes, "Information theory and statistical mechanics", Phys. Rev., vol. 106, pp. 620–630, May 1957. https://link.aps.org/doi/10.1103/PhysRev.106.620

[8] E. T. Jaynes, "Information theory and statistical mechanics II", Phys. Rev., vol. 108, pp. 171–190, Oct 1957. https://link.aps.org/doi/10.1103/PhysRev.108.171

[9] C. E. Shannon, "A mathematical theory of communication", The Bell System Technical Journal, vol. 27, no. 3, pp. 379–423, July 1948. https://ieeexplore.ieee.org/document/6773024/

[10] C. Van den Broeck, "Stochastic thermodynamics: A brief introduction", Proceedings of the International School of Physics "Enrico Fermi", vol. 184, no. Physics of Complex Colloids, p. 155–193, 2013. http://ebooks.iospress.nl/publication/33636

[11] J. von Neumann, "Thermodynamik quantenmechanischer gesamtheiten", Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse, vol. 1927, pp. 273–291, 1927. http://eudml.org/doc/59231

[12] D. Petz, "Entropy, von Neumann and the von Neumann entropy". Dordrecht: Springer Netherlands, 2001, pp. 83–96. https://doi.org/10.1007/978-94-017-2012-0_7

[13] O. J. E. Maroney, "The (absence of a) relationship between thermodynamic and logical reversibility", Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics, vol. 36, no. 2, pp. 355 – 374, 2005. http://www.sciencedirect.com/science/article/pii/S1355219805000031

[14] J. Ladyman, S. Presnell, A. J. Short, and B. Groisman, "The connection between logical and thermodynamic irreversibility", Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics, vol. 38, no. 1, pp. 58 – 79, 2007. http://www.sciencedirect.com/science/article/pii/S1355219806000682

[15] C. 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, vol. 34, no. 3, pp. 501–510, 2003. https://philpapers.org/rec/BENNOL-2

[16] T. Sagawa, "Thermodynamic and logical reversibilities revisited", Journal of Statistical Mechanics: Theory and Experiment, no. 3, p. P03025, 2014. http://stacks.iop.org/1742-5468/2014/i=3/a=P03025

[17] C. H. Bennett, "The thermodynamics of computation—a review", International Journal of Theoretical Physics, vol. 21, no. 12, pp. 905–940, Dec 1982. https://doi.org/10.1007/BF02084158

[18] B. Piechocinska, "Information erasure", Phys. Rev. A, vol. 61, p. 062314, May 2000. https://link.aps.org/doi/10.1103/PhysRevA.61.062314

[19] S. Hilt, S. Shabbir, J. Anders, and E. Lutz, "Landauer's principle in the quantum regime", Phys. Rev. E, vol. 83, p. 030102, Mar 2011. https://link.aps.org/doi/10.1103/PhysRevE.83.030102

[20] O. J. E. Maroney, "Generalizing landauer's principle", Phys. Rev. E, vol. 79, p. 031105, Mar 2009. https://link.aps.org/doi/10.1103/PhysRevE.79.031105

[21] D. Reeb and M. M. Wolf, "An improved landauer principle with finite-size corrections", New Journal of Physics, vol. 16, no. 10, p. 103011, 2014. http://stacks.iop.org/1367-2630/16/i=10/a=103011

[22] D. Reeb and M. M. Wolf, "Tight bound on relative entropy by entropy difference", IEEE Transactions on Information Theory, vol. 61, no. 3, pp. 1458–1473, March 2015. https://ieeexplore.ieee.org/document/7001656/

[23] C. E. A. Prunkl and C. G. Timpson, "On the thermodynamical cost of some interpretations of quantum theory", Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics, 2018. http://www.sciencedirect.com/science/article/pii/S1355219817300928

[24] A. Einstein, "Physics and reality", Journal of the Franklin Institute, vol. 221, no. 3, pp. 349 – 382, 1936. http://www.sciencedirect.com/science/article/pii/S0016003236910475

[25] D. Bohm, "A suggested interpretation of the quantum theory in terms of 'hidden' variables. I", Phys. Rev., vol. 85, pp. 166–179, Jan 1952. https://link.aps.org/doi/10.1103/PhysRev.85.166

[26] S. Goldstein, "Bohmian mechanics", in The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University, 2017. https://plato.stanford.edu/archives/sum2017/entries/qm-bohm/

[27] H. Everett, "'Relative state' formulation of quantum mechanics", Rev. Mod. Phys., vol. 29, pp. 454–462, Jul 1957. https://link.aps.org/doi/10.1103/RevModPhys.29.454

[28] L. Vaidman, "Many-worlds interpretation of quantum mechanics", in The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University, 2016. https://plato.stanford.edu/archives/fall2016/entries/qm-manyworlds/

[29] L. E. Ballentine, "The statistical interpretation of quantum mechanics", Rev. Mod. Phys., vol. 42, pp. 358–381, Oct 1970. https://link.aps.org/doi/10.1103/RevModPhys.42.358

[30] R. G. Colodny and A. Fine, "Paradigms & paradoxes: The philosophical challenge of the quantum domain", R. G. Colodny, Ed. University of Pittsburgh Press, 1972. https://digital.library.pitt.edu/islandora/object/pitt:31735057893376

[31] O. Lombardi and D. Dieks, "Modal interpretations of quantum mechanics", in The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University, 2017. https://plato.stanford.edu/archives/spr2017/entries/qm-modal/

[32] J. S. Bell, "The theory of local beables", Epistemological Lett., vol. 9, 1976. https://cds.cern.ch/record/980036/files/197508125.pdf

[33] G. C. Ghirardi, A. Rimini, and T. Weber, "Unified dynamics for microscopic and macroscopic systems", Phys. Rev. D, vol. 34, pp. 470–491, Jul 1986. https://link.aps.org/doi/10.1103/PhysRevD.34.470

[34] G. Ghirardi, "Collapse theories", in The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University, 2016. https://plato.stanford.edu/archives/spr2016/entries/qm-collapse/

[35] R. W. Spekkens, "Evidence for the epistemic view of quantum states: A toy theory", Phys. Rev. A, vol. 75, p. 032110, Mar 2007. https://link.aps.org/doi/10.1103/PhysRevA.75.032110

[36] N. Bohr, "The philosophical writings of niels bohr", J. Faye and H. J. Folse, Eds. Ox Bow Press, Woodbridge, 1987. https://philpapers.org/rec/BOHTPW

[37] J. Faye, "Copenhagen interpretation of quantum mechanics", in The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University, 2014. https://plato.stanford.edu/archives/fall2014/entries/qm-copenhagen/

[38] J. A. Wheeler, "Quantum theory and measurement", J. A. Wheeler and W. H. Zurek, Eds. Princeton University Press, Princeton, 1983. https://press.princeton.edu/titles/806.html

[39] C. Rovelli, "Relational quantum mechanics", International Journal of Theoretical Physics, vol. 35, no. 8, pp. 1637–1678, Aug 1996. https://doi.org/10.1007/BF02302261

[40] A. Zeilinger, "A foundational principle for quantum mechanics", Foundations of Physics, vol. 29, no. 4, pp. 631–643, Apr 1999. https://doi.org/10.1023/A:1018820410908

[41] C. A. Fuchs, "Quantum theory needs no 'interpretation'", Physics Today, vol. 53, no. 3, p. 70, 2000. https://doi.org/10.1063/1.883004

[42] C. A. Fuchs, "Qbism, the perimeter of quantum bayesianism", 2010. https://arxiv.org/abs/1003.5209

[43] C. A. Fuchs, N. D. Mermin, and R. Schack, "An introduction to qbism with an application to the locality of quantum mechanics", American Journal of Physics, vol. 82, no. 8, pp. 749–754, 2014. https://doi.org/10.1119/1.4874855

[44] N. Barnett and J. P. Crutchfield, "Computational mechanics of input–output processes: Structured transformations and the epsilon-transducer", Journal of Statistical Physics, vol. 161, no. 2, pp. 404–451, Oct 2015. https://doi.org/10.1007/s10955-015-1327-5

[45] M. Schlosshauer, "Decoherence, the measurement problem, and interpretations of quantum mechanics", Rev. Mod. Phys., vol. 76, pp. 1267–1305, Feb 2005. https://link.aps.org/doi/10.1103/RevModPhys.76.1267

[46] W. H. Zurek, "Decoherence and the transition from quantum to classical—revisited", 2003. https://arxiv.org/cits/quant-ph/0306072

[47] N. Higham, "Functions of matrices: Theory and computation". Society for Industrial and Applied Mathematics, 2008. https://epubs.siam.org/doi/book/10.1137/1.9780898717778

[48] A. Wehrl, "General properties of entropy", Rev. Mod. Phys., vol. 50, pp. 221–260, Apr 1978. https://link.aps.org/doi/10.1103/RevModPhys.50.221

[49] S. Stenholm and K.-A. Suominen, "Quantum approach to informatics". John Wiley & Sons, Inc., 2005. https://onlinelibrary.wiley.com/doi/book/10.1002/0471739367

A Appendix

A.1 The spectral theorem for finite matrices

Here we prove a theorem about complex *normal* (square) matrices. A normal matrix A satisfies $AA^\dagger = A^\dagger A$, and therefore, both *Hermitian* and *unitary* matrices are subsets of normal matrices.

Theorem A.1 (Spectral theorem for normal matrices). Let A be an $n \times n$ normal matrix, i.e. $AA^\dagger = A^\dagger A$. Then A can be decomposed in terms of a diagonal matrix Λ, containing the eigenvalues of A on the diagonal, and a unitary matrix U.

$$A = U\Lambda U^\dagger \tag{220}$$

The proof we provide here will assume *Schur decomposition*.

Lemma A.1 (Schur decomposition). Let A be an $n \times n$ complex matrix. Then A can be expressed in terms of an upper triangular matrix T, which has the eigenvalues of A on its diagonal, and an unitary matrix U.

$$A = UTU^\dagger \tag{221}$$

□

Proof (Theorem A.1). We can use Schur decomposition (lemma A.1) to rewrite our normal matrix A in terms of an upper triangular matrix T and an unitary matrix U. Then it remains to show that T is, in fact, diagonal.

$$A = UTU^\dagger \quad \Rightarrow \quad T = U^\dagger AU \quad \Rightarrow \quad T^\dagger = U^\dagger A^\dagger U \tag{222}$$

We first show that the upper triangular matrix T is *normal*.

$$TT^\dagger = U^\dagger AUU^\dagger A^\dagger U = U^\dagger AA^\dagger U = U^\dagger A^\dagger AU = U^\dagger A^\dagger UU^\dagger AU = T^\dagger T \tag{223}$$

Then we consider the element in first row and first column of $T^\dagger T$.

$$(T^\dagger T)_{11} = \sum_j t^*_{j1} t_{j1} \tag{224}$$

Since T is upper triangular we have that $t_{ab} = t^*_{ab} = 0$ if $a > b$. This means that in the sum above, only one terms survives.

$$(T^\dagger T)_{11} = t^*_{11} t_{11} = |t_{11}|^2 \tag{225}$$

However, considering the same position in $TT^\dagger$ gives a different result.

$$(TT^\dagger)_{11} = \sum_j t_{1j} t^*_{1j} = |t_{11}|^2 + |t_{12}|^2 + \cdots \tag{226}$$

Since $TT^\dagger = T^\dagger T$, according to equation (223), all entries in the first *row*, except t_{11}, must vanish; i.e. $t_{1i} = 0 \ \forall \ i \neq 1$. And we already knew that all entries in the first *column* (except t_{11}) are zero.

Then we look at the element $(T^\dagger T)_{22}$.

$$(T^\dagger T)_{22} = t_{22}^* t_{22} = |t_{22}|^2 \tag{227}$$

And also $(TT^\dagger)_{22}$, using the result from the first row, i.e. $t_{1i} = 0 \ \forall \ i \neq 1$, we find the following.

$$(TT^\dagger)_{22} = \sum_j t_{2j} t_{2j}^* = |t_{22}|^2 + |t_{23}|^2 + \cdots \tag{228}$$

Then every entry in the *second row* of T (except t_{22}) must vanish. Repeating this procedure n times, we see that when we are done T will be diagonal. ■

A.2 Obtaining the canonical expression for $\hat{\rho}$

The corresponding finite-dimensional case of *the spectral theorem*, theorem A.1 in section A.1, applies to *density operators* defined according to axiom 2.1. However the spectral theorem is most commonly found in matrix representation, and in Quantum Mechanics there is another prolific representation, as seen below (see equation (5) of section 2).

$$\hat{\rho} = \sum_{i=1}^{n} P_i \, |\phi_i\rangle\langle\phi_i| \tag{229}$$

Here we will demonstrate how to think about the transition between the two.

First, consider some basis of $\mathfrak{H}_n$ such that the operator $\hat{\rho}$ can be expressed as a matrix. According to the spectral theorem (see section A.1), we know that $\hat{\rho}$ can be written as a diagonal matrix Λ with the eigenvalues on the diagonal, sandwiched between a unitary matrix U and its Hermitian conjugate.

$$\hat{\rho} \doteq U \Lambda U^\dagger \tag{230}$$

Note that we will put a dot above equal signs when we make some transition between operator notation (common in Quantum Mechanics), and matrix notation given some basis of $\mathfrak{H}_n$.

From further theorems about unitary matrices, we know that all eigenvalues of U lies on the complex unit circle, and we also know that U has an orthogonal set of eigenvectors $\{|\phi_i\rangle\}$ (which we consider normalized) and they span the whole $\mathfrak{H}_n$. Thus, $\{|\phi_i\rangle\}$ is an orthonormal basis for $\mathfrak{H}_n$. Then, consider some arbitrary vector $|\Psi\rangle$ is expressed in the basis for U.[40]

$$\Psi \doteq \sum_{i=1}^{n} |\phi_i\rangle\langle\phi_i|\Psi\rangle \tag{231}$$

Applying (say) $U^\dagger$ to this vector will be very straight forward; we just multiply each component with its corresponding eigenvalue, $\{\, \mathrm{e}^{-\mathrm{i}\theta_i}\}$.

$$U^\dagger \Psi \doteq \sum_{i=1}^{n} \mathrm{e}^{-\mathrm{i}\theta_i} \, |\phi_i\rangle\langle\phi_i|\Psi\rangle \tag{232}$$

[40] Expressing $|\Psi\rangle$ in terms of a basis corresponds to projections onto that basis using the inner product.

We are now ready to look at the density operator $\hat{\rho}$. Let us denote the eigenvalues in Λ as $\{P_i\}$, and consider the situation where $\hat{\rho}$ is acting on $|\Psi\rangle$.

$$\begin{aligned} \hat{\rho}\,|\Psi\rangle &\doteq U\Lambda U^\dagger \Psi \doteq \\ \doteq U\Lambda \sum_{i=1}^{n} \mathrm{e}^{-\mathrm{i}\theta_i}\,|\phi_i\rangle\langle\phi_i|\Psi\rangle &\doteq U \sum_{i=1}^{n} P_i\,\mathrm{e}^{-\mathrm{i}\theta_i}\,|\phi_i\rangle\langle\phi_i|\Psi\rangle \doteq \\ \doteq \sum_{i,j} \mathrm{e}^{\mathrm{i}\theta_i} P_i\,\mathrm{e}^{-\mathrm{i}\theta_i}\,|\phi_j\rangle\langle\phi_j|\phi_i\rangle\langle\phi_i|\Psi\rangle &= \sum_{i=1}^{n} P_i\,|\phi_i\rangle\langle\phi_i|\Psi\rangle \end{aligned} \tag{233}$$

After the last step we have an expression for $\hat{\rho}\,|\Psi\rangle$, but we are looking for an expression for the operator $\hat{\rho}$ alone. We can simply leave the last slot in the inner product empty.

$$\hat{\rho} = \sum_{i=1}^{n} P_i\,|\phi_i\rangle\langle\phi_i|\cdot\rangle \tag{234}$$

Here, the dot means that we should insert whatever $\hat{\rho}$ is applied on in its place. However, it is more common to leave out the dot and the closing bracket—for a more compact and aesthetic notation.

$$\hat{\rho} = \sum_{i=1}^{n} P_i\,|\phi_i\rangle\langle\phi_i| \tag{235}$$

The notation $|\phi_i\rangle\langle\phi_i|$ is then interpreted as a scalar product of the vector $|\phi_i\rangle$ and the inner product $\langle\phi_i|\cdot\rangle$ taken with whatever $\hat{\rho}$ is applied to.

A.3 The logarithm of an operator

The natural logarithm of an operator $\hat{A}$, associated with a finite Hilbert space $\mathfrak{H}_n$, is defined as the operator $\hat{X}$, whose exponential series expansion produces $\hat{A}$.

$$\begin{gathered} \text{Find } \hat{X} \text{ such that } \mathrm{e}^{\hat{X}} = \hat{A} \\ \text{where} \\ \mathrm{e}^{\hat{X}} := \sum_{n=0}^{\infty} \frac{1}{n!}\,\hat{X}^n \end{gathered} \tag{236}$$

We then make the following definition.

$$\ln \hat{A} := \hat{X} \tag{237}$$

We note that there is generally no guarantee that such $\hat{X}$ exists, and if it does exist there may be many. There are, however, theorems that allows one to determine if none, one, or several solutions exist. Here we will simply state the theorems and provide references for the reader.

Theorem A.2. An operator has $\hat{A}$ has at least one logarithm if and only if it is invertible.

See Theorem 1.27 in the book "Functions of Matrices: Theory and Computation" by Nicholas Higham, [47].

Theorem A.3. An operator has $\hat{B}$ has a unique logarithm if it has no eigenvalues on the negative real axis.

See Theorem 1.31 in the book "Functions of Matrices: Theory and Computation" by Nicholas Higham, [47].

A.4 Invariance of the trace

Lemma A.2. Consider a *normal* operator, i.e. $\hat{A}\hat{A}^\dagger = \hat{A}^\dagger\hat{A}$, associated with the finite Hilbert space $\mathfrak{H}_n$. The trace of $\hat{A}$ is invariant under the choice of basis in which the trace is taken.

Proof (Lemma A.2). By definition, the trace of some normal operator $\hat{A}$, is the sum over an arbitrarily chosen orthonormal and complete basis $\{|\varphi_i\rangle\}$ for $\mathfrak{H}_n$, where we take the inner products with with each basis vector.

$$\mathfrak{Tr}\big[\hat{A}\big] := \sum_{i=1}^{n} \langle\varphi_i|\hat{A}\varphi_i\rangle \tag{238}$$

Since $\hat{A}$ is assumed to be *normal*, we can use the *spectral theorem* (see section A.1 and A.2) to rewrite the operator on its diagonal form, in some diagonalizing basis $\{|\phi_i\rangle\}$.

$$\hat{A} = \sum_{j=1}^{n} \lambda_j \, |\phi_j\rangle\langle\phi_j| \tag{239}$$

Inserting equation (239) into (238), we can express the trace of $\hat{A}$ explicitly.

$$\mathfrak{Tr}\big[\hat{A}\big] = \sum_{i,j} \lambda_j \, \langle\varphi_i|\phi_j\rangle \, \langle\phi_j|\varphi_i\rangle \tag{240}$$

We rearrange the inner products and identify a unit operator (possible since $\{|\varphi_i\rangle\}$ is a complete orthonormal basis).

$$\mathfrak{Tr}\big[\hat{A}\big] = \sum_{j=1}^{n} \lambda_j \sum_{i=1}^{n} \langle\phi_j|\varphi_i\rangle \, \langle\varphi_i|\phi_j\rangle = \sum_{j=1}^{n} \lambda_j \, \langle\phi_j|\phi_j\rangle = \sum_{j=1}^{n} \lambda_j \tag{241}$$

Thus $\mathfrak{Tr}\big[\hat{A}\big]$ is invariant under the choice of $\{|\varphi_i\rangle\}$. ■

A.5 Invariance of von Neumann entropy under unitary transformations

Lemma A.3 (Invariance of von Neumann entropy). The von Neumann entropy $S_N(\hat{\rho}) := \mathfrak{Tr}\big[\hat{\rho}\ln\hat{\rho}\big]$ of a density operator $\hat{\rho}$, as defined in axiom 2.1 (see section 2), is invariant under unitary transformations.

$$S_N(\hat{U}\hat{\rho}\,\hat{U}^\dagger) = S_N(\hat{\rho}) \tag{242}$$

Proof (Lemma A.3). First, it is easy to show[41] that $\mathrm{e}^{\hat{U}\hat{X}\hat{U}^\dagger} = \hat{U}\,\mathrm{e}^{\hat{X}}\hat{U}^\dagger$, using the definition of the exponential (see section A.3). We will then prove that $\ln(\hat{U}\hat{\rho}\,\hat{U}^\dagger) = \hat{U}\ln(\hat{\rho})\,\hat{U}^\dagger$. From section A.3 we know that $\ln(\hat{U}\hat{\rho}\,\hat{U}^\dagger) =: \hat{X}$ for the $\hat{X}$ that solves the following.

$$\mathrm{e}^{\hat{X}} = \hat{U}\hat{\rho}\,\hat{U}^\dagger \quad \Rightarrow \quad \hat{U}^\dagger\,\mathrm{e}^{\hat{X}}\hat{U} = \hat{\rho} \quad \Rightarrow \quad \mathrm{e}^{\hat{U}^\dagger\hat{X}\hat{U}} = \hat{\rho} \tag{243}$$

Then $\hat{U}^\dagger\hat{X}\hat{U}$, is the logarithm of $\hat{\rho}$.

$$\ln(\hat{\rho}) = \hat{U}^\dagger\hat{X}\hat{U} = \hat{U}^\dagger\ln(\hat{U}\hat{\rho}\,\hat{U}^\dagger)\,\hat{U} \quad \Rightarrow \tag{244}$$

$$\hat{U}\ln(\hat{\rho})\hat{U}^\dagger = \ln(\hat{U}\hat{\rho}\,\hat{U}^\dagger) \tag{245}$$

With this, and the fact that the trace is invariant under unitary transformations (equivalent of choosing a different basis, see section A.4), we can complete our proof.

$$S_N(\hat{U}\hat{\rho}\,\hat{U}^\dagger) = \mathfrak{Tr}\Big[\hat{U}\hat{\rho}\,\hat{U}^\dagger\ln\big(\hat{U}\hat{\rho}\,\hat{U}^\dagger\big)\Big] = \mathfrak{Tr}\Big[\hat{U}\hat{\rho}\,\hat{U}^\dagger\hat{U}\,\ln\big(\hat{\rho}\big)\hat{U}^\dagger\Big] = \tag{246}$$

$$= \mathfrak{Tr}\Big[\hat{U}\,\hat{\rho}\,\ln\big(\hat{\rho}\big)\,\hat{U}^\dagger\Big] = \mathfrak{Tr}\big[\hat{\rho}\ln\hat{\rho}\big] = S_N(\hat{\rho}) \tag{247}$$

■

A.6 Relative entropy

Theorem A.4 (Properties of relative entropy). Consider two density operators $\hat{\rho}_1$ and $\hat{\rho}_2$, both associated with the same finite-dimensional Hilbert space $\mathfrak{H}_n$ (see axiom 2.1). Then the relative entropy $S(\hat{\rho}_1\|\hat{\rho}_2)$ as defined below, assumes only non-negative values, and is zero if and only if $\hat{\rho}_1 = \hat{\rho}_2$.

$$S(\hat{\rho}_1\|\hat{\rho}_2) := -\mathfrak{Tr}\big[\hat{\rho}_1\,(\ln\hat{\rho}_2 - \ln\hat{\rho}_1)\big] \tag{248}$$

We will not provide proofs for the properties of relative entropy, but simply mention that the they are supported by the so called *Klein's inequality*. See Alfred Wehrl's paper for further discussions and proofs, [48].

Intuitive understanding of *relative entropy* is most straight forward to build in a classical framework (with classical probability distributions), where *relative*

[41] Left to the reader as an exercise.

entropy is called *Kullback-Leibler divergence*. Consider two probability distributions $P = \{P_i\}$ and $Q = \{Q_i\}$ over the same *finite* set of events. The classical relative entropy, D_{KL}, is very similar as in Quantum Mechanics, equation (248).

$$D_{KL}(P||Q) := -\sum_i P_i\big(\ln Q_i - \ln P_i\big) \tag{249}$$

Most resources discuss the *relative entropy* as a *distance-like* measure, answering the question: "How much information is lost if we approximate a *true* probability distribution $\{P_i\}$ with an approximate distribution $\{Q_i\}$?". Note however that it cannot be an *actual* distance measure, since relative entropy is not symmetric with respect to $\{P_i\}$ and $\{Q_i\}$.

Another quirk of relative entropy that is worth our attention is that it is not bounded, even on a finite set of events. If the probability distribution Q contains some events that has a zero probability, but the corresponding event in P is non zero we get an infinite term.

$$-P_i\big(\ln 0 - \ln P_i\big) = \infty \qquad \text{for} \qquad 0 < P_i \leq 1 \tag{250}$$

A.7 Spectra of product operators

Here we want to determine what we can say about the *spectra* (the set of eigenvalues) of a finite dimensional operator, if it happens to be a *product operator* of two operators with known spectra.

Lemma A.4 (Spectra of product operators). Consider two *normal* operators $\hat{A}$ and $\hat{B}$, associated with the finite Hilbert spaces $\mathfrak{H}_m$ and $\mathfrak{H}_n$ respectively. Let the spectra of $\hat{A}$ be $\{a_i\}$, and the spectra of $\hat{B}$ be $\{b_i\}$. Construct their product operator $\hat{P}$, associated with the Hilbert space $\mathfrak{H}_{m\times n}$.

$$\hat{P} = \hat{A} \otimes \hat{B} \tag{251}$$

The spectra of $\hat{P}$ is $\{a_i b_j\}$.

Proof (Lemma A.4). Since $\hat{A}$ and $\hat{B}$ are *normal*, we can use the spectral theorem (see sections A.1 and A.2) and express them on their diagonal form.

$$\hat{A} = \sum_{i=1}^{m} a_i \, |a_i\rangle\langle a_i| \tag{252}$$

$$\hat{B} = \sum_{j=1}^{n} b_j \, |b_j\rangle\langle b_j| \tag{253}$$

According to the *spectral theorem*, $\{|a_i\rangle\}$ is an orthonormal basis for $\mathfrak{H}_m$, and $|b_j\rangle$ is an orthonormal basis for $\mathfrak{H}_n$. We can then form a basis for $\mathfrak{H}_{m\times n}$ by taking the tensor product of the two bases.

$$\big\{\, |a_i\rangle \otimes |b_j\rangle \,\big\} \quad \text{is an orthonormal basis for} \quad \mathfrak{H}_{m\times n} \tag{254}$$

We apply $\hat{P}$ to the set of basis vectors $\{|a_i\rangle \otimes |b_j\rangle\}$.

$$\hat{P}\big(\,|a_i\rangle \otimes |b_j\rangle\,\big) =$$
$$= \left(\left(\sum_k a_k\,|a_k\rangle\langle a_k|\right) \otimes \left(\sum_\ell b_\ell\,|b_\ell\rangle\langle b_\ell|\right)\right)\big(\,|a_i\rangle \otimes |b_j\rangle\,\big) = \tag{255}$$

$$= a_i b_j\big(\,|a_i\rangle \otimes |b_j\rangle\,\big) \tag{256}$$

We see that $|a_i\rangle \otimes |b_j\rangle$ is an eigenvector of $\hat{P}$, with the eigenvalue $a_i b_j$. ■

A.8 Normal matrices under unitary transformations

Lemma A.5 (Eigenvalues under unitary transformations). Let A be a complex normal matrix, $AA^\dagger = A^\dagger A$. Let B be the result of a invertible transform V, applied to A, such that B is also normal.

$$B := VAV^{-1} \qquad ; \qquad BB^\dagger = B^\dagger B \tag{257}$$

V is unitary, if and only if, the eigenvalues of B are the same as the eigenvalues of A.

Proof (Lemma A.5). First, we prove that if V is unitary, the eigenvalues are the same. Since A is a normal matrix, we know from the *spectral theorem* (section A.1) that it can be written as a decomposition with a diagonal matrix Λ containing the eigenvalues of A, and a unitary matrix U.

$$A = U\Lambda U^\dagger \tag{258}$$

Then B becomes the following.

$$B = VU\Lambda U^\dagger V^{-1} = VU\Lambda U^\dagger V^\dagger = VU\Lambda(VU)^\dagger \tag{259}$$

We define $W := VU$, and rewrite B.

$$B = W\Lambda W^\dagger \tag{260}$$

We then prove that W is unitary, from the knowledge that U and V are unitary.

$$WW^\dagger = VU\,(VU)^\dagger = VUU^\dagger V^\dagger = \mathbb{1} \tag{261}$$

$$W^\dagger W = (VU)^\dagger VU = U^\dagger V^\dagger VU = \mathbb{1} \tag{262}$$

This proves that the eigenvalues of B are the the same as for A; the diagonal elements of Λ.

Next, we prove that if A and B have the same eigenvalues, V is unitary. Since A and B are normal, we can use the *spectral theorem* again (section A.1). Since we assume the eigenvalues of A and B to be equal, we can use the same diagonal matrix, Λ, for both decompositions.

$$A = U\Lambda U^\dagger \qquad ; \qquad B = W\Lambda W^\dagger \tag{263}$$

But according to equation (257), B can also be written in terms of A.

$$W\Lambda W^\dagger = B \overset{(257)}{=} VAV^{-1} = VU\Lambda U^\dagger V^{-1} \tag{264}$$

Comparing the left hand side and the right hand side, we see that $W = VU$ and $W^\dagger = U^\dagger V^{-1}$. From the latter relation, since W is unitary, $U^\dagger V^{-1}$ is unitary.

$$U^\dagger V^{-1}(U^\dagger V^{-1})^\dagger = \mathbb{1} \quad \Rightarrow \tag{265}$$

$$U^\dagger V^{-1}(V^{-1})^\dagger U = \mathbb{1} \quad \Rightarrow \tag{266}$$

$$V^{-1}(V^{-1})^\dagger = \mathbb{1} \quad \Rightarrow \tag{267}$$

$$(V^{-1})^\dagger = V \quad \Rightarrow \tag{268}$$

$$V^{-1} = V^\dagger \tag{269}$$

Thus V is unitary. ■

A.9 Conditional entropy

Note, this section and the next assumes some familiarity with *random variables*.

Say we want to look at the *Shannon entropy* (section 3.5) in a certain random variable X, belonging to the alphabet $\mathcal{X}$, but there exist some correlation with some other random variable Y, of the alphabet $\mathcal{Y}$. So the value of Y does not necessarily determine X, but it can affect the probabilities. We can then talk about the entropy that remains in X, for any specific value $y \in \mathcal{Y}$.

$$H\big(X \,|\, Y{=}y\big) \tag{270}$$

Clearly, to evaluate this using the definition of *Shannon entropy*—see equation (12) in section 3.5—we require some description of the probabilities of $X = x$ for every case of $Y = y$. This probability of x, given y, is called the *conditional probability*, and denoted as $P(x|y)$.

$$H\big(X \,|\, Y{=}y\big) = -\sum_{x\in\mathcal{X}} P(x|y)\, \log_2\big(P(x|y)\big) \tag{271}$$

This can be interpreted as quantifying the entropy that remains in X, for some particular value $Y = y$. Then, from this construct we can create the so called *conditional entropy*, $H(X|Y)$, by simply averaging over every y, weighted with its probability.

$$H\big(X \,\big|\, Y\big) := \sum_{y\in\mathcal{Y}} P(y)\, H\big(X \,\big|\, Y{=}y\big) = \tag{272}$$

$$= -\sum_{y\in\mathcal{Y}} \left(P(y) \sum_{x\in\mathcal{X}} P(x|y)\, \log_2\big(P(x|y)\big) \right) \quad \Rightarrow \tag{273}$$

$$H\big(X \,|\, Y\big) = -\sum_{x,y} P(y)\, P(x|y)\, \log_2\big(P(x|y)\big) \tag{274}$$

We can rewrite this using a so called *probability mass function* $P(x, y)$ (a.k.a. *joint probability distribution*, or *diagram probabilities* in Shannon [9]) that treats both random variables, X and Y, symmetrically.

$$P(x,y) = P(x)\,P(y|x) = P(y)\,P(x|y) \tag{275}$$

$$\Rightarrow$$

$$H\big(X\,|\,Y\big) = -\sum_{x,y} P(x,y)\,\log_2\left(\frac{P(x,y)}{P(y)}\right) \tag{276}$$

Note that probabilities of individual variables, $P(x)$ and $P(y)$, can be obtained from the *probability mass function* $P(x,y)$.

$$P(x) = \sum_{y\in\mathcal{Y}} P(x,y) \qquad ; \qquad P(y) = \sum_{x\in\mathcal{X}} P(x,y) \tag{277}$$

Whether we use equation (274) or (276) is a matter of what probabilities are most conveniently accessible.

A.10 Mutual information

Let the outcome of a random variable X have some correlation with the outcome another random variable, Y. The purpose of the so called *mutual information* is to quantify this correlation. Here we derive and discuss the expression for this *mutual information* denoted by $I(X{:}Y)$.

Starting from quantifying the uncertainty, a.k.a. entropy, of a random variable X, we can ask how much the entropy is reduced, ΔH if we learn that Y has some specific value y.

$$\Delta H\big(Y{=}y\big) = H\big(X\big) - H\big(X\,\big|\,Y{=}y\big) \tag{278}$$

The *mutual information* is simply average over all possible values, $y \in \mathcal{Y}$, weighted by their probabilities.

$$\begin{aligned} I(X{:}Y) &= \sum_{y\in\mathcal{Y}} P(y)\Big(H\big(X\big) - H\big(X\,|\,Y{=}y\big)\Big) = \\ &= H\big(X\big) - \sum_{y\in\mathcal{Y}} P(y)\,H\big(X\,|\,Y{=}y\big) \end{aligned} \tag{279}$$

The last term is the *conditional entropy*, $H(X\,|\,Y)$, as discussed in section A.9.

$$I(X{:}Y) = H\big(X\big) - H\big(X\,|\,Y\big) \tag{280}$$

Using the definition of *Shannon entropy* from section 3.5, equation (12), and equalities from section A.9, equations (276) and (277), we can rewrite this expression in a more useful format, based on the *probability mass function*, $P(x,y)$.

$$\begin{aligned} I(X{:}Y) &= H(X) - H\big(X\,|\,Y\big) = \\ &= -\sum_{x\in\mathcal{X}} P(x)\,\log_2\big(P(x)\big) + \sum_{x,y} P(x,y)\,\log_2\left(\frac{P(x,y)}{P(y)}\right) = \\ &= -\sum_{x,y} P(x,y)\,\log_2\big(P(x)\big) + \sum_{x,y} P(x,y)\,\log_2\left(\frac{P(x,y)}{P(y)}\right) \qquad \Rightarrow \end{aligned} \tag{281}$$

$$I(X{:}Y) = \sum_{x,y} P(x,y)\, \log_2 \left(\frac{P(x,y)}{P(x)\, P(y)} \right) \tag{282}$$

Here, we can clearly see that mutual information is symmetric with respect to the two random variables X and Y.

Further is is also possible to show that it is non-negative, equal to zero if and only if X and Y are independent, and if the mutual information equals the entropy of X, then X is completely determined by Y, see [49].