5-2016

Memory as Bayesian inference: On the connection between memory and the second law of thermodynamics

Wade Daniel Hodson
College of William and Mary

Memory as Bayesian inference

On the connection between memory and the second law of thermodynamics

A thesis submitted in partial fulfillment of the requirement
for the degree of Bachelor of Science in Physics from
The College of William and Mary

by

Wade Daniel Hodson

Accepted for Honors
(Honors or no-Honors)

Eugene Tracy, Advisor

Henry Krakauer, Physics

Peter Vishton, Psychology

Marc Sher, Physics

Williamsburg, VA
May 3, 2016

Memory as Bayesian inference: On the connection between memory and the second law of thermodynamics

Wade D. Hodson

Advisor: Dr. Eugene R. Tracy

May 11, 2016

Abstract

A recent theoretical paper by Leonard Mlodinow and Todd Brun suggests that the functioning of physical records or "memories" is never accompanied by a decrease in entropy, meaning that all memories "align" with the thermodynamic arrow of time. In this thesis, we characterize a class of physical systems as memories in terms of inferences that can be made about the state of the world, given certain information about these systems. Tools from Bayesian probability theory are used to quantify the informativeness and reliability associated with such inferences. Based on consideration of two model systems, one classical and one quantum, we argue in favor of Mlodinow and Brun's claim that the functioning of memory systems is conditioned by thermodynamic constraints. For the classical model, we show that a memory which operates against the thermodynamic arrow, and thus "remembers" a relatively high-entropy state, is much less informative than a similar memory which aligns with the thermodynamic arrow. Our analysis of the quantum model, expressed in the density matrix formalism of quantum mechanics, allows us to consider the inferences that can be made when a quantum system is coupled to a simple type of quantum memory system. We ultimately show that these inferences can be expressed in terms of a probabilistic matrix completion problem.

Contents

1 Questions and Motivation **3**

1.1 Introduction 3

1.2 Memory as inference 4

1.3 Thermodynamics and memory 8

1.4 Prior work of Mlodinow and Brun 10

2 A Classical Memory Model **17**

2.1 Characterization of the model 17

2.2 Memory inferences in the urn model 19

3 A Quantum Memory Model **30**

3.1 Characterization of the model 30

3.2 Probabilities in the quantum model 36

4 Conclusions **41**

5 Appendix A: Bayesian Probability Theory **43**

6 Appendix B: The Density Matrix Formalism **48**

1 Questions and Motivation

1.1 Introduction

There is an extensive scientific canon which posits a link between the second law of thermodynamics, which describes the increase in thermodynamic entropy over time, and physical memories, or systems which record information. Researchers have made the case for such a relationship from a range of perspectives. Some, like psychologist Ryan Smith, approach memory from a neurological standpoint: Smith has noted that the micro-scale processes which are the building blocks of human cognition rely on a predictable increase in entropy.[10] Others have considered the brain and other memory systems as computing systems, and have investigated how the physical instantiation of logical operations is constrained by thermodynamic effects. For example, the famous "Landauer's principle" states that every erasure of a bit, or "clearing" of a bit of memory, entails a net increase in entropy of $k_B \ln 2$.[7] Some authors, such as Lorenzo Maccone, have even searched for a connection between memory and entropy in quantum physics. Maccone has argued that the physics of quantum systems makes it so that all processes which leave a "trail of information" cannot be entropy-decreasing processes.[8]

The goal of this thesis is to explore how the reliability of a physical record is conditioned by the existence of a thermodynamic arrow of time. Our analysis was inspired by the 2014 work of Leonard Mlodinow and Todd Brun, who argued that physical systems can only function precisely and reliably as memories of the thermodynamic past.[9] That is, if a memory accessed at a time t_1 reveals information about the state of the world at another time t_2, then the entropy of the memory and its surroundings at t_1 must be higher than the entropy at t_2, in order for the memory to be reliable or "robust." In our Universe, where all isolated subsystems so far observed exhibit a common thermodynamic arrow of time, this result implies that all memory systems share a common "directionality" in time which aligns with our intuitive, human distinction between the past and the future.

Mlodinow and Brun's analysis is motivated by the property of time-reversal invariance, a characteristic possessed by many modern physical theories, including quantum mechanics and special relativity. The mathematical structure of such theories does not pick out a certain direction in time as obviously past or future. Rather, this temporal ordering is imposed by users of these theories, who already possess a

definite notion of past and future. The question of how our world, governed by such time-symmetric laws, could nevertheless appear so asymmetric when we compare one direction in time to the other, is known as Loschmidt's paradox. Mlodinow and Brun's analysis aims to address Loschmidt's paradox in the particular context of memory systems. That is, they propose an explanation for why all systems that function as memories only appear to operate in one direction of time, despite the underlying time-symmetric dynamics of these systems.

Our contribution to Mlodinow and Brun's work has been to develop two simple models, one classical and one quantum, that we believe can help clarify and extend their analysis. In the remainder of Chapter 1, we offer some theoretical background and motivation for our work, and characterize the terms of the research question more precisely. In particular, we introduce our interpretation of memory as a type of probabilistic inference. This reasoning is applied to a classical model system in Chapter 2. We consider a simple stochastic system of marbles moving between two urns, and show that the informativeness of a "memory-type" inference about certain properties of the marble system is contingent upon changes in the entropy of the system over time. In Chapter 3, we explore a model quantum memory system with a similar approach. Using the density matrix formulation of quantum mechanics, we investigate how inferences about the state of a quantum system are conditioned by observations of a memory device coupled to the system. Chapter 4 concludes our investigation, and highlights some possible avenues for future work. Finally, in Appendix A, we present an overview of the theory of Bayesian probabilistic inference, and in Appendix B we give an introduction to the density matrix formalism.

1.2 Memory as inference

To begin our investigation, a concrete definition of a physical memory system is essential. Here, we wish to understand the category of memories as a broad class of "record" systems, so as to include any system which can function as a record or repository of information about the world. Accordingly, this conception of memory systems encompasses systems like the human brain or computer memories, which can be understood as specifically evolved or designed for the function of recording information. But it also extends beyond these straightforward cases. For example, we would also like to identify objects like fossils as a record of this type, because when given an observation of a fossil's physical properties, we can reasonably extract a wealth of information about the distant past. Of course, the possibility of accessing this information also hinges upon a background of prior information and a pre-existing theoretical framework, which includes scientific

knowledge of geological processes, evidence from similar fossils, and an understanding of biological evolution.

So we wish to define a class of "memory systems" in terms of the information about the world which might be *inferred* from observation or analysis of such systems. With this conception of memory in mind, we put forward the following working definition of a memory system. In this work, we classify a physical system M as a memory of a second system S if:

Given some knowledge of the state or properties of M at some time t_1, some inference about the state of S at another time t_2 can be made with high confidence.

Two important aspects of this characterization of a memory system immediately stand out. First, this definition suggests that memory systems come in degrees, with the reliability of the memory scaling with the confidence that can be placed in the relevant inferences being made. Indeed, almost any system might qualify as a potential memory under this criterion, although the inferences made based on most such systems will be so uninformative as to be practically useless. Although our definition does not provide a strict dividing line between memories and non-memories, it does leave open the possibility that similar memory systems might be compared, based on the relative confidence associated with the inferences they allow.

Second, under this definition, the status of a system as a usable memory depends not only on the raw physical constitution of the memory and its environment, but also on the logical procedure of inference. Of course, the types of inferences which can be made about the world based on a memory will be heavily constrained by physical interactions between the memory and its surroundings. However, we will argue that a purely physical account of memory systems, which does not introduce some notion of inference, is insufficient to account for a range of common systems which function as useful memories.

We emphasize here that our conception of inference should not be construed as an exclusively human activity. Rather, we understand inference as an abstract logical procedure, which could potentially be embodied in a range of physical systems. A system which instantiates an inference procedure take inputs which encode relevant information about the world, processes this information with a set of physical operations corresponding to logical rules, and then produces a physical result associated with this processed information. Consequently, any system capable of encoding and manipulating information in this way could perform inferences about memory systems. Of course, one example of such a system is the human brain, but systems like artifical neural networks and some non-human animals are also capable of inference in this sense.

Proceeding with this definition of memory suggests two related objections. First, why should we even introduce a concept like inference into our definition of a memory? Surely, there must be some way to appraise

a system as a memory solely in terms of its physical properties, without interpretation through an inferential procedure. Second, even if such a definition of memory is put forward, how could it possibly be made precise and unambiguous, given that the state of a single physical record may imply different conclusions under different logical frameworks for inference? We have already noted, for example, that the inferences which could be made about a fossil specimen are heavily conditioned by relevant evidence and scientific knowledge. Though a physical system may, in practice, function as a memory for one observer but not another, this seems to be only a contingent fact about when and how the memory is accessed, and not a rigorous criterion by which we might distinguish memories from non-memories.

In addressing the first objection, we argue that a definition of memory *without* some notion of inference would exclude a broad class of systems which, intuitively and practically, we think of as memories or records. This is because many memory systems, particularly those at a macroscopic scale, function *probabilistically* as memories. That is, these systems work well as memories not because they deterministically record the state of the world with perfect fidelity, but because we judge them to be reasonably reliable under certain conditions. The presence of noise and outlying data does not prevent us from categorizing our experimental instruments as memories; these errors simply temper our trust in these devices. In these cases, it may be difficult or practically impossible to determine whether, in any particular instance, a system acts "properly" as a memory or whether it malfunctions. However, this limitation does not rule out the possibility of building probabilistic theoretical models of these memory systems, which can provide an incomplete and yet useful account of how such systems operate and record information.

Allowing for a probabilistic conception of memory systems is particularly important in the present work, given that we aim to describe memories in the presence of the thermodynamic arrow of time. As we will explain in the next section, results from statistical mechanics suggest that the changes in entropy which determine an arrow of time emerge only statistically: It is not that an observable decrease in entropy is physically impossible, but only that it is extremely unlikely. If we wish to properly account for memory systems which exhibit a thermodynamic arrow of time, we must be able to articulate a meaning for memory in this probabilistic context.

Making this notion of probability more precise will help to address the second objection, concerning the dependence of inference on background information and a pre-existing logical framework. In this work, we follow the Bayesian approach to the assignment and interpretation of probabilities. In the Bayesian formulation of probability, a numerical probability quantifies the degree of certainty that can be associated with

the truth or falsity of a given inference. These probabilities are *assigned*, not observed, based on information relevant to the inference and on requirements of logical consistency. In this way, the Bayesian probability calculus functions as a generalization of standard two-valued formal logic, appropriate for scenarios where incomplete information prevents a definite assignment of "true" or "false" to specific claims.

The logical groundwork of Bayesian probability theory is described in more detail in Appendix A. However, for our present purposes, what is most important to note is that although Bayesian probabilities must be assigned with respect to given prior information, these assignments are done in a systematic and logically consistent manner. To begin with, Bayesian probabilities are constrained by the same basic mathematical axioms that all other theories of probability must satisfy. For example, all probabilities must respect countable additivity, which requires that the probability of observing any of a set of mutually exclusive outcomes is simply the sum of the probabilities of the individual outcomes. However, for any given problem to which Bayesian analysis is applied, certain considerations may not only limit the possible probabilities that an observer might assign, but may single out a *unique* probability assignment as appropriate for the situation. These considerations may have to do with relevant prior information, symmetries of the problem, or constraints related to the rationality or self-consistency of the logical procedure.

But this type of formalism is exactly what we need in order to sharpen our notion of inference. For if the information gathered from a memory system allows some inference to be made about the world, then the tools of Bayesian logic make it possible to assign a probability to the truth value of this inference. This probability quantifies the degree of certainty associated with the inference, and therefore quantifies the extent to which the memory system functions as a useful memory. The most functional and reliable memory systems will permit precise inferences, with associated high probabilities.

So although we acknowledge that any memory inference will depend on relevant prior knowledge and on a given logical framework, Bayesian probability theory allows us to treat this information in a consistent and systematic way. In some instances, the relevant probability assignments can be entirely determined by the given information, combined with constraints of logical consistency encoded in the Bayesian formalism. In these cases, the inference procedure prescribed by Bayesian logic is, in a sense, the only reasonable way to perform inference, since any other method may lead to inconsistent results.

1.3 Thermodynamics and memory

Given this notion of memory as a form of inference, we can now begin to sketch an outline of our investigation. In this work, we aim to analyze memory-type inferences in the context of a thermodynamic arrow of time: A monotonic change in entropy over some period of time. In existing memory systems, this arrow invariably points in the same direction: All known memory systems are records of a past which had a lower entropy than the present. This suggests that entropy changes in a memory system or its surroundings may condition the inferences that observation of a memory system permits. We will argue that memory inferences in the traditional sense, which refer to a low-entropy past state based on observation of a high-entropy present state, can be made much more precise and reliable than the reverse.

In setting the framework for this argument, it is important to be clear about the particular notion of entropy which we refer to in speaking of an "arrow of time." Entropy is a rich concept, with distinct yet related meanings in thermodynamics, statistical mechanics, and information theory. Although colloquially understood as a measure of disorder, the precise definitions of entropy in various fields do not always align with this popular conception. In our analysis, we make use of the statistical mechanical Gibbs entropy, defined as follows.

Suppose a classical system can be measured as having certain macroscopic properties, such as a definite temperature or a given energy. Although these measurements tell an observer something about the state of this system, they do not specify its state completely. A full account of the microscopic configuration of a classical system would consist of knowledge of the positions and velocities of every constituent particle of the system; this is clearly a practically unreachable goal. However, certain microscopic configurations that are *in principle* possible might be incompatible with the given macroscopic measurements, and other configurations might be judged as more or less likely in light of these measured quantities. In defining the Gibbs entropy, we consider the set or "ensemble" of all microscopic configurations which might possibly give rise to the observed macroscopic measurements. If we define p_i as the probability that the i^{th} state in this ensemble is the actual microscopic configuration of the system, then the Gibbs entropy S is expressed as

$$S = -k_B \sum_i p_i \ln p_i. \tag{1}$$

This definition assumes that the ensemble contains a countable number of configurations; in the case of an uncountable ensemble, this sum must be replaced with an integral. If we assume further that the

ensemble has some finite number of states Ω, then we can consider the special case where all these states are equiprobable, such that $p_i = 1/\Omega$. This probability assignment then yields an especially simple form for the entropy, $S = k_B \ln \Omega$. In this simplified formula, the Gibbs entropy appears as a measure of the number of possible microscopic configurations that a system could conceivably attain, given certain macroscopic knowledge or constraints.

In the quantum mechanical case, the Gibbs entropy is generalized as the von Neumann entropy, which is defined in terms of the density operator ρ which describes the system of interest. The von Neumann entropy takes the form

$$S = -k_B \mathrm{Tr}\big[\hat{\rho} \ln \hat{\rho}\big], \tag{2}$$

where the symbol Tr signifies the trace operation, and the operator logarithm $\ln \hat{\rho}$ can be computed in terms of a power series. As described in more detail in Appendix B, the density operator $\hat{\rho}$ contains information about the probabilities associated with the possible outcomes of measurements on a quantum system.

A fundamental result of statistical mechanics, demonstrated in a variety of proofs with varying degrees of generality, is that any decrease in the Gibbs or von Neumann entropy of an isolated system with many interacting degrees of freedom has vanishing probability.[2,4,5,11] Loosely speaking, this result emerges because in a system with a large number of microscopic components, there are far more configurations of the system which correspond to high entropy states than to low entropy states. This fact is manifest in our simplified expression for the Gibbs entropy, $S = k_B \ln \Omega$, where the entropy scales logarithmically with the number of possible microscopic configurations Ω . Therefore, if a system starts in some low entropy state, a very large fraction of the states it could potentially evolve into have higher entropy.

The "thermodynamic arrow of time" points in the direction of this increase in entropy. Although the proof of an arrow of time in statistical mechanics is a probabilistic one, the probabilities involved are so close to unity for macroscopic systems that the increase of entropy can be treated as a practical certainty. Technically, we have only sketched an explanation for the increase of the *statistical mechanical* entropy, not the *thermodynamic* entropy, which can be computed from measurements of temperature and heat flow. However, it is a standard conclusion in statistical mechanics that these two entropies coincide for simple thermodynamic systems, so for our analysis working with the statistical mechanical entropy is sufficient.[5]

With this concept of entropy clearly defined, we can now outline the goals of our analysis in the following two chapters. In this work, we study two model memory systems, one with classical dynamics and the

other with quantum dynamics. In Chapter 2, we introduce our classical model, a probabilistic system which is a modification of the "urn problem" from introductory probability theory. This urn model lends itself to a certain class of memory inferences, in which observation of some property of the system allows an inference to be made about another property of the system at a different time. Using the framework of Bayesian probability theory, we can compute the probabilities associated with this class of inferences. These probabilities act as a metric for the reliability of inferences: A useful, informative memory system will allow for a sharp probability distribution to be assigned over the possible values of a recorded observable. We pay special attention to cases where the entropy of the urn system changes over time. Analysis of these cases allows us to describe how the urn model memory is conditioned by a thermodynamic arrow of time, as identified by a sustained increase in entropy.

For the quantum model, we consider a pair of coupled quantum systems, an abstract memory system and an observed system which is coupled to the memory system. In Chapter 3, we characterize and analyze this model in terms of the density matrix formalism of quantum mechanics. We consider a simple form of coupling between the two systems which permits straightforward memory inferences, and then investigate how measurement of the state of the memory system informs an observer about the possible states of the recorded system.

1.4 Prior work of Mlodinow and Brun

The original impetus for this project was a 2014 paper, "Relation between the psychological and thermodynamic arrows of time," written by physicists Leonard Mlodinow and Todd Brun.[9] In their work, Mlodinow and Brun attempt to explain the direction of the "psychological arrow of time," that is, the temporal direction in which memory devices function, as a consequence of the direction of the thermodynamic arrow. They argue that any useful memory system exhibits a property they term "generality," and then go on to show that any memory system with this property must operate in the same direction as the thermodynamic arrow. In Mlodinow and Brun's analysis, a memory that satisfies generality is one which does not have to be "fine-tuned" to a particular state in order to record each possible outcome. In the remainder of the first chapter, we present a summary of Mlodinow and Brun's approach, and argue that our inferential account of memory systems can supplement their work.

To motivate their work, Mlodinow and Brun begin with a brief note on time reversal invariance. Except for during certain uncommon processes, the microscopic dynamics of both classical and quantum systems

are adequately described by mathematical formalisms which are invariant under time reversal. That is, the equations which govern the time evolution of these systems do not change form when the time variable t is replaced by $-t$. This mathematical symmetry has a physical consequence: Given any possible sequence of states that a system could move through during its evolution, there exists another possible evolution through that same set of states, but in the opposite temporal order. For certain simple systems, this claim seems plausible. A ball thrown in an arc through the air could certainly, in principle, be thrown again so that it moves through the same arc but in the opposite direction. But for most events, this consequence of time reversal invariance flies in the face of everyday experience. It seems ridiculous to suggest that the existence of an egg shattering implies that an egg could spontaneously reassemble itself, or that a drop of dye diffused in water might be suddenly concentrated back into a dense point. These processes appear to us as clearly "one-way" or irreversible, whether they occur naturally or with the help of human intervention.

This apparent paradox can be resolved by recourse to the second law of thermodynamics. From the perspective of statistical mechanics, the second law tells us that processes during which the entropy of a closed system decreases, although not disallowed by the laws of physics, are nevertheless extremely improbable. This means that although we may observe a range of entropy-increasing processes, the time-reversed versions of these processes are effectively impossible due to the entropy deficit that they would generate. Mlodinow and Brun's suggestion is that this asymmetry in time can be used to explain another asymmetry, namely the fact that memory systems only operate as memories in one direction of time. In their view, it is not just a contingent fact that all memory systems function in the same direction of time, but rather a physical consequence of the universal existence of the thermodynamic arrow of time.

As in this thesis, Mlodinow and Brun's notion of a memory is very general, intended to include any system which can function as a record of the state of another system. However, to develop their argument, they first focus on a simple model system that captures the basic content of their result. Mlodinow and Brun's model system consists of two canisters of a classical ideal gas, and a microscopic rotor or "turnstile" placed inside a nozzle which connects the two canisters. This rotor has M possible positions, labeled by the integers 0 to $M-1$, and Mlodinow and Brun define a variable $r(t)$ which corresponds to the rotor position at any particular time t. The nature of the rotor is such that whenever a gas particle moves from left to right, the position of the rotor increments by one, while a flow of particles in the opposite direction causes the rotor to decrement. A diagram of this setup is given in Figure 1.

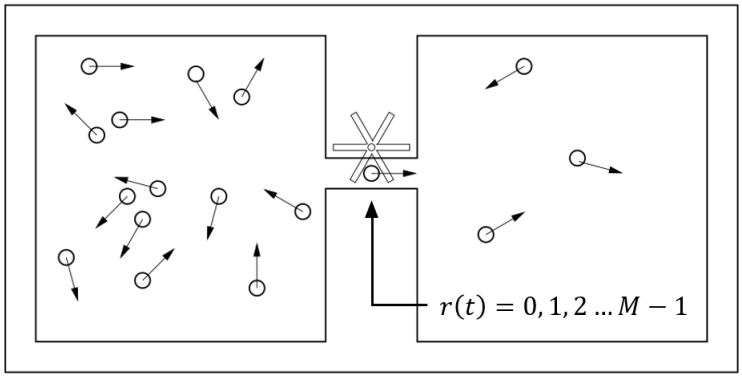

Figure 1: Diagram of Mlodinow and Brun's gas-rotor model memory system. The microscopic configuration of the gas is determined by specifying the position and momentum of each gas molecule. The state of the rotor is specified by the variable $r(t)$, which increments or decrements by 1 whenever a gas molecule passes from one canister to the other. This image is a modified version of a figure that appears in Mlodinow and Brun's 2014 paper.[9]

Given these dynamics, Mlodinow and Brun then introduce a memory "readout" function $f_{read}(r(t))$, which tracks the net number of particles that have passed from left to right between a time t_i and another time t. This function is defined as

$$f_{read}(r(t)) = r(t) - r_{ref}, \tag{3}$$

where $r_{ref} \equiv r(t_i)$ is the position of the rotor at the reference time t_i, and $t_i < t$. As expressed in this function, Mlodinow and Brun's model system has been constructed so that the rotor position $r(t)$ encodes information about the evolution of the gas. In this way, the rotor functions as a memory of changes in the distribution of gas particles between the two canisters. However, Mlodinow and Brun argue that there is no reason that the dynamics of the system cannot be interpreted in a different way, as encoding a record of how many particles *will* pass from left to right between two times t and t_f, where $t_f > t$. Such a "memory" of the future could be expressed in terms of a second readout function,

$$f'_{read}(r(t)) = r'_{ref} - r(t). \tag{4}$$

In this case, $r'_{ref} \equiv r(t_f)$ is the position of the rotor at the *final* reference time t_f. At this stage, there is no compelling reason to deem (3) to be a valid memory readout, and at the same time deny the validity of (4). No characteristic of the system singles out a particular direction in time as "past-like" in any way; the

classical nature of the system insures that the dynamics completely respect time-reversal symmetry. Indeed, our choice of the particular time variable t to track the evolution of the system is somewhat arbitrary, in the sense that the temporal direction in which it increases has no special significance. If we were to re-describe the system in terms of the alternate time variable $-t$, (3) would then appear as a memory of the future. The point here is that no absolute notion of past or future is encoded in the dynamics of the gas-rotor system itself, although observers may impose a convenient time variable in order to track the system's evolution.

However, a definite time asymmetry, and therefore a notion of past and future, does emerge for certain evolutions of Mlodinow and Brun's model. Consider a scenario in which at some reference time, the distribution of particles between canisters is highly asymmetric. For example, there might be far more particles of gas in the left canister than in the right. In such cases, experience tells us that gas will rush into the right canister from the left, until a uniform gas density has been achieved across the two canisters. Specifically, this result is a consequence of the second law of thermodynamics, since the asymmetric distribution at the reference time has a lower entropy than an even distribution. Although this low-entropy condition at the reference time does not *guarantee* an evolution to uniform gas density, the difference in entropy between the two states does make this evolution extremely likely. This preference for trajectories that increase entropy defines the gas-rotor system's thermodynamic arrow of time. In the context of such a tendency, referring to a state of the system as in the "past" of another state simply means that the first state has a lower entropy than the other.

It is under these conditions, when the model system exhibits a thermodynamic arrow of time, that a salient asymmetry between the two memory readout functions (3) and (4) becomes apparent. To understand this distinction, Mlodinow and Brun introduce the concept of "generality." If a memory system satisfies generality, then it must be capable of recording more than one possible state of the world, without having to be set specifically to record each possible state. For Mlodinow and Brun, a memory which does not meet this criterion is hardly a memory at all: For example, they note that if a digital camera required a new chip, specific to each scene, in order to take a photograph, the camera would be of little use as a memory in the first place. Requiring memories to satisfy generality insures that memory systems are not correlated with other systems due to some special "fine-tuning" of the memory state, but due to a definite interaction with the world that reliably effects the memory system.

Armed with the generality criterion, the essential difference between the memory readouts in (3) and (4) can now be explained. Assume that the system exhibits a standard thermodynamic arrow of time, so for

two times $t_1 < t_2$, the entropy of the system at t_1 is lower than at t_2. In the case of the standard "past memory" as defined by the readout in (3), this means that the readout at any time $t > t_i$ will be associated with an increase in entropy between t_i and t. To impose the generality condition for this memory, we need to understand how the memory would record different possible states. Consider the exact microscopic state of the gas at t_i, which can be specified in principle by listing the positions and velocities of all its constituent particles. If the memory satisfies generality, then the readout function $f_{read}(r(t))$ given in (3) ought to still function as a record if this microscopic state differed slightly from its true value. Indeed, this is the case. If we imagine perturbing the system at t_i in this way, as long as we perform this perturbation so that $r_{ref} = r(t_i)$ remains unchanged, $f_{read}(r(t))$ would still offer an accurate account of the flow of gas particles between t_i and t. The exact evolution of the rotor may be slightly different, but this is exactly what we want, since we wish to show that the memory can record multiple possible outcomes. In addition, we see that for such perturbations, the arrow of time we have postulated is generally preserved. This is because although we have altered the microscopic state of the gas at t_i, this microscopic state still corresponds to a low-entropy state of the gas. In Figure 2, several possible evolutions of $f_{read}(r(t))$ are displayed for a low-entropy initial condition, all of which exhibit a definite arrow of time.

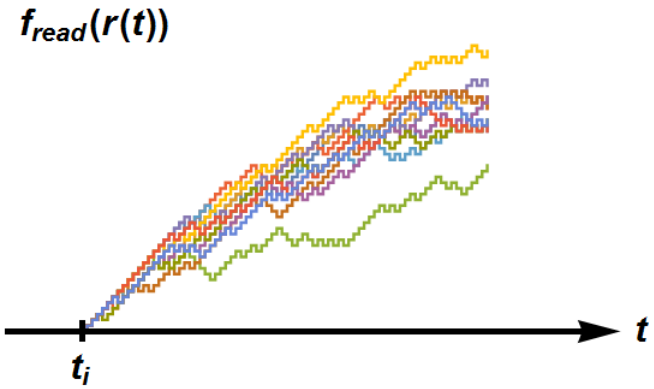

Figure 2: Schematic plots of several possible evolutions of the memory readout function $f_{read}(r(t))$, given a low-entropy initial condition at $t = t_i$. Each of these trajectories corresponds to a different microscopic perturbation of the gas, such that $r_{ref} = r(t_i)$ is left unchanged. All of them exhibit a well-defined arrow of time in the direction of increasing t.

For the "future memory" expressed in the readout function $f'_{read}(r(t))$ in (4), the situation is quite different. Here, if we want to consider the possible states that this memory could record, we can imagine

a similar perturbation of the microscopic state of the gas, this time at t_f. Again, as long as we do not change the value of $r'_{ref} = r(t_f)$ in this perturbation, the record encoded in $f'_{read}(r(t))$ will still remain valid, although the particular value that is read out at t may change. However, we find that upon making such perturbations, the thermodynamic arrow of time we originally proposed will generally be destroyed. This is because, by assumption, the entropy of the gas at t_f is larger than at any time $t < t_f$. But the second law of thermodynamics tells us that given this entropy at the boundary time t_f, the entropy at any other time t is almost certainly higher, which contradicts our assumption. So in attempting to define a memory in terms of the function $f'_{read}(r(t))$, we are forced to conclude that if a thermodynamic arrow of time does exist for the system, it will generally lead us to identify the time t_f with the thermodynamic past. This effect is demonstrated in Figure 3, where we plot a number of possible evolutions of $f'_{read}(r(t))$ from a high-entropy final state, most of which do not pick out an arrow of time in either direction.

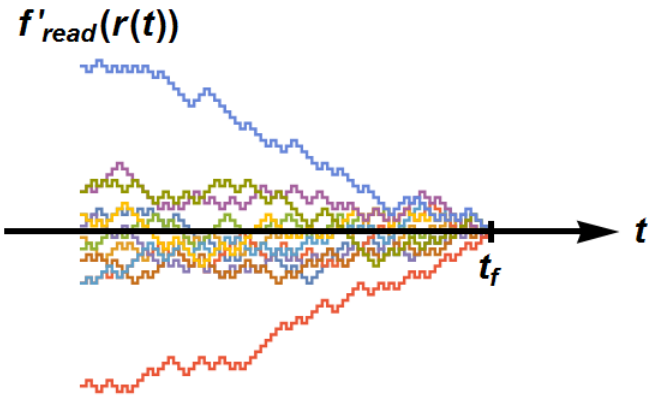

Figure 3: Schematic plots of several possible evolutions of the memory readout function $f'_{read}(r(t))$, given a high-entropy initial condition at $t = t_f$. Each of these trajectories corresponds to a different microscopic perturbation of the gas, such that $r'_{ref} = r(t_f)$ is left unchanged. Most of these evolutions do not exhibit a significant arrow of time.

In our view, this argument proposed by Mlodinow and Brun is essentially sound. However, we claim that without some understanding of memory as a type of inference, some aspects of their proposal remain unclear. The difficulty appears when we try to clarify what is meant by a "generic" evolution of the gas-rotor system. In their work, Mlodinow and Brun claim that a generic evolution of the model system exhibits an arrow of time if, for the vast majority of trajectories possible given some constraints, there is a similar change in entropy over the course of the evolution. First of all, in classical physics, the number of possible

trajectories that satisfy certain conditions will in general be uncountably infinite, since classical systems can exhibit a continuous range of position and momentum values. But more importantly, this metric based on some "number of trajectories" cannot account for probabilistic constraints, which might suggest that certain evolutions are much more likely than others. Indeed, this is precisely the type of constraint that a thermodynamic arrow of time imposes: To say that a system has a thermodynamic arrow of time does not necessarily disallow decreases in entropy, but means that such evolutions of the system are extremely unlikely. In order to precisely account for the probabilistic nature of memory systems in a thermodynamic context, our contribution in this thesis is to bring the tools of Bayesian inference to bear upon the problem.

2 A Classical Memory Model

In this chapter, we introduce a modified version of the urn problem from standard probability theory, as a prototypical example of a classical memory system. The dynamics of this model are simple enough to permit a full analytical treatment of the relevant probabilities, and they allow for a straightforward demonstration of how a thermodynamic arrow of time can emerge purely from statistical considerations. We demonstrate that in the urn model, memory inferences are most informative when made about the thermodynamic past, such that they refer to states of relatively low entropy.

2.1 Characterization of the model

In the urn model, we consider a pair of two urns, labeled arbitrarily as "right" and "left," and a set of N marbles distributed between them. Defining the variables $N_R(t)$ and $N_L(t)$ respectively as the number of marbles in the right and left urns at some time t, we have the constraint $N_R(t) + N_L(t) = N$. The dynamics of the system are as follows: At each of a set of discrete times $t = ... - 2, -1, 0, 1, 2...$, a single marble is removed from one of the urns, and placed into the other. The selection of each marble is "random" in the sense that an observer of the system has no knowledge of the mechanism by which individual marbles are selected. To display the evolution of the system over time, we introduce a sequence of quantities $\{\sigma_i | i \in Z\}$, where each value σ_i corresponds to the i^{th} time-step. The elements of this sequence are defined so that $\sigma_i = +1$ if a marble is moved from right to left at time $t = i$, and $\sigma_i = -1$ if a marble is moved from left to right. So, for each time-step from $t = i$ to $t = i + 1$, the distribution of marbles $(N_R(t), N_L(t))$ undergoes an evolution of the form

$$(N_R(t), N_L(t)) \rightarrow (N_R(t+1), N_L(t+1)) = (N_R(t) - \sigma_i, N_L(t) + \sigma_i). \tag{5}$$

Now, suppose that this system is coupled to a device, the function of which is to record and display the *net* number of marbles that have moved from right to left, between some reference time $t = 0$ and another time $t = n$. Given the definition of the quantities σ_i, the output $f(n)$ of this "memory register" will be a sum of all values of σ_i between $t = 0$ and $t = n$,

$$f(n) = \sum_{m=0}^{n-1} \sigma_m. \tag{6}$$

The evolution rules expressed in (5) and (6) describe all of the observer's knowledge of the dynamics of the system. It is important to note that at this stage in the characterization of our urn model, it is not physically meaningful to define either direction in time as "past" or "future." Nothing in the general dynamics of the urn system itself singles out a direction of time as obviously past or obviously future. Out of a desire to provide a definite temporal label to each marble swap, we are compelled to define *some* time variable t which increases in a particular direction. But the direction affiliated with this choice of labeling is no more meaningful than associating "up" on the Earth's surface with positive values of a height coordinate. In fact, if we specify a suitable summation convention for how to evaluate the sum in (6) for $n \leq 0$, then we can track the evolution of the memory register for $t = -1, -2, -3...$ as well.

However, introducing a concept of entropy for the urn system will allow us to define an arrow of time for particular evolutions of the marble distribution. The statistical entropy associated with this system, for any distribution of marbles $(N_R(t), N_L(t))$, is computed as $S = k_B \log \Omega$, where Ω the number of possible configurations of marbles which would produce that given distribution. For example, in the extreme case where $N_R(t) = 0$, there is only a single way to achieve this distribution, simply by placing every marble in the left urn, and so $S = k_B \log 1 = 0$. The entropy $S = k_B \log \Omega$ is simply a special case of the Gibbs entropy in (1), where we have assumed that all configurations are equiprobable. In the general case, there will be multiple ways to obtain the desired distribution. Suppose we wish to arrange the marbles in the distribution $(N_R(t), N_L(t))$. Given N marbles, this means we must choose $N_R(t)$ marbles out of the total N to place in the right urn. But there are $\Omega = N!/(N_R(t)!(N - N_R(t))!) = N!/(N_R(t)!N_L(t)!)$ ways to make this choice, each corresponding to a different configuration of marbles which has the same overall distribution $(N_R(t), N_L(t))$. The entropy S of the system for this distribution is then just

$$S = \log\left(\frac{N!}{N_R(t)!N_L(t)!}\right) = \log N! - \log N_R(t)! - \log N_L(t)!, \tag{7}$$

where we have set $k_B = 1$. If $N_R(t)$ and $N_L(t)$ are very large, we can apply Stirling's approximation to each term in (7). We find that the entropy per marble S/N is effectively a function of the relative fractions of marbles, $N_R(t)/N$ and $N_L(t)/N$, in the right and left urns, and is given by

$$S/N \approx -\frac{N_R(t)}{N}\log\left(\frac{N_R(t)}{N}\right) - \frac{N_L(t)}{N}\log\left(\frac{N_L(t)}{N}\right). \tag{8}$$

In Figure 4, this expression is plotted as a function of $N_R(t)/N$ alone, which we can compute since $N_R(t)/N + N_L(t)/N = 1$. Inspecting this plot reveals that the entropy drops to zero when either the right urn or the left urn is completely empty, at $N_R(t)/N = 0$ and $N_R(t)/N = 1$, respectively, and that it reaches a maximum when the marbles are evenly distributed at $N_R(t)/N = N_L(t)/N = \frac{1}{2}$. So, for a given time evolution of the marble distribution, we would associate a thermodynamic arrow of time with this evolution if there is a strong tendency for the system to progress from an unevenly distributed state to a balanced state. We will see such a tendency for evolutions for which the urn model begins in a low entropy state.

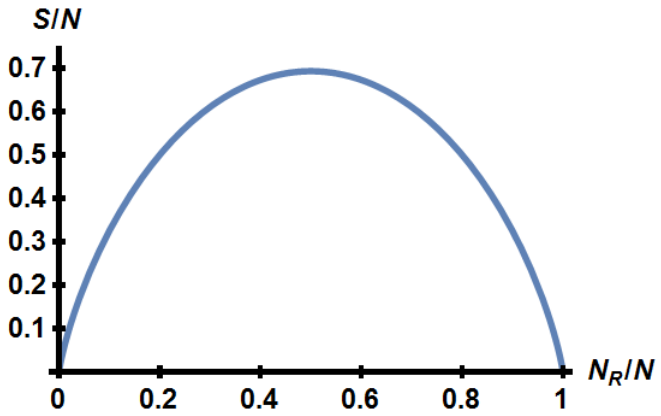

Figure 4: Plot of the approximate entropy per marble S/N, given in (8), as a function of $N_R(t)/N$. S/N reaches a maximum at $N_R/N = \frac{1}{2}$, which corresponds to an even distribution of marbles between the two urns.

2.2 Memory inferences in the urn model

Now that the physical properties and dynamics associated with the urn model have been defined, we can ask questions about the various "memory-type" inferences which can be made about the urn model. Specifically, we would like to know: Given knowledge of $f(n)$, the output of the memory register, at some particular time $t = n$, what can we infer about the distribution of marbles (N_R, N_L) at $t = 0$?

We can formulate this question in terms of Bayesian inference as follows. Suppose that we define $\bar{f} = f(n)$ as the observed output of the register at $t = n$, and we let $N_R = N_R(0)$ and $N_L = N_L(0)$ refer to the number of marbles in the right and left urns, respectively, at $t = 0$. N_R and N_L therefore do not change in time.

If we are making an inference about the distribution of marbles at $t = 0$, this means that we would like to assign probabilities to the various possible distributions (N_R, N_L), conditioned on the value of $\bar{f}$. We denote these probabilities as

$$P(N_R, N_L|\bar{f}, n, N), \tag{9}$$

which translates verbally to "the probability that the marbles have the distribution (N_R, N_L) at $t = 0$, conditioned on the knowledge of $\bar{f}$, n, and N." This is a probability distribution over the possible marble distributions (N_R, N_L), which expresses our confidence in an inference about the marble distribution. $P(N_R, N_L|\bar{f}, n, N)$ reaches a maximum around the values of (N_R, N_L) which are most likely, given knowledge of $\bar{f}$. Moreover, the informativeness of our inference scales with the narrowness of the probability distribution around this peak. A tightly peaked distribution suggests that (N_R, N_L) is almost certainly given by the peak value, while a broad and flat distribution does not strongly single out any particular values of (N_R, N_L).

In order to understand how knowledge of $\bar{f}$ informs our inference about the value of N_R, and thus the marble distribution at $t = 0$, we can compute the probabilities in (9) in terms of a Bayesian updating procedure. As described in detail in Appendix A, this procedure derives from the rule for assigning probabilities to a joint proposition,

$$P(A, B|I) = P(B|A, I)P(A|I) = P(A|B, I)P(B|I). \tag{10}$$

Here, the probability $P(A, B|I)$ that the statements A and B are *both* true is computed in terms of two factors: The probability $P(B|A, I)$ that B is true, given that A is true, and the probability $P(A|I)$ that A is true. All of these probabilities are conditioned on the knowledge of some prior information I.

For our inference, we can identify the proposition A with a claim about the distribution (N_R, N_L) at the time $t = 0$, and associate B with the observed value $\bar{f}$ of the memory register. We take the prior information I as encoding the values of n and N. By making these substitutions into (11) and then solving for $P(A|B, I) = P(N_R, N_L|\bar{f}, n, N)$, we find the relation

$$P(N_R, N_L|\bar{f}, n, N) = \frac{P(\bar{f}|N_R, N_L, n, N)}{P(\bar{f}|n, N)} P(N_R, N_L|n, N). \tag{11}$$

Thus, we can calculate our desired probability $P(N_R, N_L|\bar{f}, n, N)$, which we now refer to as the *posterior*

distribution, in terms of three other probability distributions. These are:

- The *likelihood* function, $P(\bar{f}|N_R, N_L, n, N)$. The likelihood is the probability that a given value $\bar{f}$ is observed as the output of the memory register at $t = n$, given that a particular distribution of marbles (N_R, N_L) is known at $t = 0$.

- The *prior* distribution, $P(N_R, N_L|n, N)$. This gives the probability that the marbles are in the distribution (N_R, N_L) at $t = 0$, *independent* of any knowledge of the value of $\bar{f}$. The prior probability distribution quantifies our initial state of knowledge concerning the marble distribution, before any information about the output of the memory register is obtained.

- The *normalization*, $P(\bar{f}|n, N)$. This gives the probability that a given value $\bar{f}$ is observed as the output of the memory register at $t = n$, independent of any information about the marble distribution (N_R, N_L) at $t = 0$. The normalization is fixed by requiring that the posterior is a normalized probability distribution over the possible values of (N_R, N_L).

We now consider each of these three distributions in turn, in order to understand the behavior of the posterior distribution. In the calculations that follow, we will use abbreviated notation that suppresses any dependence on n and N_L. For example, we will denote the posterior distribution as $P(N_R|\bar{f}, N) \equiv P(N_R, N_L|\bar{f}, n, N)$.

We begin with the likelihood function. To compute $P(\bar{f}|N_R, N)$, we first note that any particular value of $\bar{f}$ can be the outcome of a number of different sequences of swaps $\{\sigma_i\}$. Specifically, whenever a sequence of swaps occurs such that $\bar{f} = f(n) = \sum_{m=0}^{n-1} \sigma_m$, the output on the memory register will be $\bar{f}$. Each of these possible sequences is a mutually exclusive outcome. Therefore the probability $P(\bar{f}|N_R, N)$ is simply the sum of the probabilities of obtaining each individual sequence,

$$P(\bar{f}|N_R, N) = \sum_{}^{'} P(\{\sigma_i\}|N_R, N). \tag{12}$$

Here, the prime over the sum specifies that the sum only runs over sequences $\{\sigma_i\}$ for which $\bar{f} = \sum_{m=0}^{n-1} \sigma_m$. Now, the problem of calculating $P(\bar{f}|N_R, N)$ reduces to computing the probability of the system evolving through various sequences $\{\sigma_i\}$, which we have written as $P(\{\sigma_i\}|N_R, N)$.

To proceed with this calculation, it is easiest to start with a short sequence, and then generalize the procedure for longer evolutions. Consider the situation when $n = 2$. In this case, the evolution of the system

is given by a sequence $\{\sigma_0, \sigma_1\}$, where σ_0 specifies the direction of the marble swap between $t = 0$ and $t = 1$, and σ_1 does the same for the swap between $t = 1$ and $t = 2$. The probability $P(\{\sigma_0\, \sigma_1\}|N_R, N)$ is then the joint probability that the first swap corresponds to the value of σ_0, *and* that the second swap corresponds to the value of σ_1. Expanding this as a conditional probability, we find the relation

$$P(\{\sigma_0, \sigma_1\}|N_R, N) = P(\sigma_0|N_R, N)P(\sigma_1|\sigma_0, N_R, N). \tag{13}$$

So, the probability of the sequence $\{\sigma_0, \sigma_1\}$ occurring is the product of two factors: The probability $P(\sigma_0|N_R, N)$ of attaining the first value σ_0 of the sequence, and the probability $P(\sigma_1|\sigma_0, N_R, N)$ of attaining the second value σ_1 of the sequence, *conditioned* on the fact that σ_0 is given.

We can assign a value to the first factor, $P(\sigma_0|N_R, N)$, by the following logic. Between $t = 0$ and $t = 1$, there are N possible ways in which the system might evolve, since any of the N marbles might be moved into the opposite urn. To an observer, nothing known about the dynamics we have specified gives any reason to single out any particular marble; there is a symmetry in the information that the observer has about the marbles. This suggests that the observer ought to assign an equal probability to each marble being swapped, because otherwise one or more marbles would be arbitrarily distinguished from the others. Therefore, the probability of any given marble being swapped must be $1/N$, so that the total probability that *any* of the marbles is swapped is unity. As described in Appendix A, this justification for a uniform assignment of probabilities is commonly known as the "Principle of Indifference."

However, not all of these possible evolutions will produce a given value of σ_0. For example, the value $\sigma_0 = +1$ could be realized in N_R possible ways, since there are N_R marbles which could be moved from the right urn. Since we have assigned equal probability to each of these mutually exclusive possibilities, the probability that *some* marble will be moved from right to left is N_R/N. But this is just the probability that σ_0 attains the value $+1$, and so we have $P(\sigma_0 = +1|N_R, N) = N_R/N$. By an analogous argument, a probability of $P(\sigma_0 = -1|N_R, N) = N_L/N$ ought to be assigned to the claim that $\sigma_0 = -1$.

Similar reasoning also motivates the probability assignment for the second factor, $P(\sigma_1|\sigma_0, N_R, N)$. Here, it is given that σ_0 has some known value. This means that at $t = 1$, the distribution of marbles has evolved to $(N_R - \sigma_0, N_L + \sigma_0)$, and we aim to assign probabilities to the occurrence of $\sigma_1 = +1$ and $\sigma_1 = -1$. But this situation is precisely analogous to our calculation of $P(\sigma_0|N_R, N)$: The distribution of marbles at a given time-step is known, and we must assign probabilities to infer the subsequent value in the sequence $\{\sigma_i\}$. Symbolically, we have the identity

$$P(\sigma_1|\sigma_0, N_R, N) = P(\sigma_1|N_R - \sigma_0, N). \tag{14}$$

That is, the probability of a left (or right) swap during the second time-step, given that the direction of the first swap is known, is equal to the probability of a left (or right) swap during the first time-step, but with the initial distribution $(N_R - \sigma_0, N_L + \sigma_0)$ instead of (N_R, N_L). Thus, for $\sigma_1 = +1$, we have $P(\sigma_1 = +1|\sigma_0, N_R, N) = (N_R - \sigma_0)/N$, and for $\sigma_1 = -1$, we have $P(\sigma_1 = -1|\sigma_0, N_R, n) = (N_L + \sigma_0)/N$. After multiplying these results by the probabilities we obtained for $P(\sigma_0|N_R, N)$, we arrive at the probabilities associated with each possible sequence $\{\sigma_0, \sigma_1\}$:

$$P(\{\sigma_0 = +1, \sigma_1 = +1\}|N_R, N) = \frac{N_R}{N}\frac{N_R - 1}{N}, \tag{15}$$

$$P(\{\sigma_0 = +1, \sigma_1 = -1\}|N_R, N) = \frac{N_R}{N}\frac{N_L + 1}{N}, \tag{16}$$

$$P(\{\sigma_0 = -1, \sigma_1 = +1\}|N_R, N) = \frac{N_L}{N}\frac{N_R + 1}{N}, \tag{17}$$

$$P(\{\sigma_0 = -1, \sigma_1 = -1\}|N_R, N) = \frac{N_L}{N}\frac{N_L - 1}{N} \tag{18}$$

For $n = 2$, the probability of obtaining any value of $\bar{f}$ can now be computed by summing over the appropriate probabilities given in (15)-(18), as noted earlier in (12). For example, the value $\bar{f} = 0$ will be observed if either the sequence $\{\sigma_0 = +1, \sigma_1 = -1\}$ or $\{\sigma_0 = -1, \sigma_1 = +1\}$ is realized, so we have $P(\bar{f} = 0|N_R, N) = P(\{\sigma_0 = +1, \sigma_1 = -1\}|N_R, N) + P(\{\sigma_0 = -1, \sigma_1 = +1\}|N_R, N)$.

In the $n = 3$ case, for which we wish to assign a probability to a sequence $\{\sigma_0, \sigma_1, \sigma_2\}$, the generalization is straightforward. Again, we expand the probability of the whole sequence as a product of conditional probabilities, as in (13). Only this time, the result consists of three factors, one for each term in the sequence:

$$P(\{\sigma_0, \sigma_1, \sigma_2\}|N_R, N) = P(\sigma_0|N_R, N)P(\sigma_1|\sigma_0, N_R, N)P(\sigma_2|\{\sigma_0, \sigma_1\}, N_R, N). \tag{19}$$

The first two factors are the same as what we computed for the $n = 2$ case. The third factor can be assigned a probability by an argument similar to that used to assign the value of the second factor. Specifically, the probability of a left (or right) swap during the third time-step, given that the directions of the first two swaps are known, is equal to the probability of a left (or right) swap during the first time-step, but with the initial distribution $(N_R - \sigma_0 - \sigma_1, N_L + \sigma_0 + \sigma_1)$ instead of (N_R, N_L). So in terms of the output of the memory register $f(n)$, we have the expression

$$P(\sigma_2|\{\sigma_0,\sigma_1\},N_R,N) = P(\sigma_2|N_R-\sigma_0-\sigma_1,N) = P(\sigma_2|N_R-f(n=2),N). \tag{20}$$

At this point, a definite pattern has begun to emerge. In the general case, for arbitrary n, we expand the probability $P(\{\sigma_i\}|N_R,N)$ of a sequence $\{\sigma_i\}$ as a product of conditional probabilities. Recall that the probabilities $P(\{\sigma_i\}|N_R,N)$ are what we require to compute the likelihood function, as given in the sum (12). Proceeding with this expansion yields a product of n factors, which takes the form

$$P(\{\sigma_i\}|N_R,N) = \prod_{k=0}^{n-1} P(\sigma_k|\{\sigma_i, 0\leq i<k\},N_R,N). \tag{21}$$

In this expression, each factor $P(\sigma_k|\{\sigma_i, 0\leq i<k\},N_R,N)$ is the probability that the value σ_k will be realized on the $(k+1)^{th}$ swap, given that the truncated sequence of values $\{\sigma_i, 0\leq i<k\}$ is known to have been attained on the previous k swaps. But just as in the $n=2$ and $n=3$ cases, we know that this probability is equal to the probability of σ_k occurring on the *first* swap, given the initial distribution $(N_R-f(k),N_L+f(k))$. Generalizing (14) and (20), we then have the result

$$P(\sigma_k|\{\sigma_i, 0\leq i<k\},N_R,N) = P(\sigma_k|N_R-f(k),N) = \begin{cases} (N_R-f(k))/N & \text{if } \sigma_k=+1, \\ (N_L+f(k))/N & \text{if } \sigma_k=-1. \end{cases} \tag{22}$$

So, the probability $P(\{\sigma_i\}|N_R,N)$ given in (21) is simply a product of factors of the form $(N_R-f(k))/N$ or $(N_L+f(k))/N$, depending on the sign of σ_k.

In order to simplify our expression for $P(\{\sigma_i\}|N_R,N)$ given in (21), we will now assume that the magnitude of the memory register is always much smaller than N, so that $|f(k)| << N$ throughout the evolution of the register. This simply means that we take the change in the number of marbles between urns to be much less than the total number of marbles. Such an approximation implies that the $f(k)$ terms in (22) can be neglected, yielding

$$P(\sigma_k|\{\sigma_i, 0\leq i<k\},N_R,N) \approx \begin{cases} N_R/N & \text{if } \sigma_k=+1, \\ N_L/N & \text{if } \sigma_k=-1. \end{cases} \tag{23}$$

That is, in the limit where $|f(k)| << N$, the probabilities associated with each swap are independent of the results of the previous swaps. This result is intuitively reasonable, given that for very large N, swapping a few marbles from urn to urn will leave the relative fraction of marbles apportioned to each urn practically unchanged.

Now, we can see that under this approximation, each factor in (21) can only have one of two possible values. For the factors for which $\sigma_k = +1$, we have $P(\sigma_k|\{\sigma_i, 0 \leq i < k\}, N_R, N) = N_R/N$, and for the factors for which $\sigma_k = -1$, we have $P(\sigma_k|\{\sigma_i, 0 \leq i < k\}, N_R, N) = N_L/N$. So, if we define n_R as the total number of swaps taken from the right urn between $t = 0$ and $t = n$, and n_L is the total number of swaps taken from the left, then (21) simplifies to

$$P(\{\sigma_i\}|N_R, N) = \left(\frac{N_R}{N}\right)^{n_R} \left(\frac{N_L}{N}\right)^{n_L}. \tag{24}$$

Noting that $n = n_R + n_L$ and $\bar{f} = n_R - n_L$, we can then express this probability in terms of $\bar{f}$ and n as

$$P(\{\sigma_i\}|N_R, N) = \left(\frac{N_R}{N}\right)^{\frac{n+\bar{f}}{2}} \left(\frac{N_L}{N}\right)^{\frac{n-\bar{f}}{2}}. \tag{25}$$

Therefore, under the assumption that $|f(k)| << N$, the probabilities $P(\{\sigma_i\}|N_R, N)$ associated which each possible sequence $\{\sigma_i\}$ are in fact *independent* of the particular order of the terms σ_i, and only sensitive to the net passage of marbles $\bar{f}$ between urns. This means that all the terms in (12), which sum to give us the likelihood function, are in fact identical and equal to (25). So to complete our calculation of the likelihood function, it suffices to determine the number of terms in (12).

Since each term in (12) corresponds to a sequence $\{\sigma_i\}$ for which $\bar{f} = \sum_{m=0}^{n-1} \sigma_m$, the total number of terms is simply the number of possible sequences of length n which satisfy this constraint. We can deduce this number by noting that if a sequence of length n produces an output of $\bar{f}$ on the memory register, then there were $n_R = (n + \bar{f})/2$ swaps from right to left during the sequence. These swaps could have occurred at any time between $t = 0$ and $t = n$, and are marked by the values of the sequence for which $\sigma_k = +1$. Each possible "placement" of right-to-left swaps within the n time-steps corresponds to a distinct sequence for which $\bar{f} = \sum_{m=0}^{n-1} \sigma_m$. Therefore, counting the number of such sequences is equivalent to counting the number of ways that we can "choose" n_R of the n time-steps to have $\sigma_k = +1$. But this is just the binomial coefficient $\binom{n}{n_R} = n!/(n_R!(n - n_R)!) = n!/(n_R!n_L!)$. So, summing the $n!/(n_R!n_L!)$ terms in (12), all equal to (25), yields our final expression for the likelihood function,

$$P(\bar{f}|N_R, N) = \frac{n!}{n_R!n_L!} \left(\frac{N_R}{N}\right)^{n_R} \left(\frac{N_L}{N}\right)^{n_L} = \frac{n!}{\left(\frac{n+\bar{f}}{2}\right)! \left(\frac{n-\bar{f}}{2}\right)!} (x)^{\frac{n+\bar{f}}{2}} (1-x)^{\frac{n-\bar{f}}{2}}, \tag{26}$$

where we have introduced the variable $x \equiv N_R/N$. Note that the likelihood has now taken the form of a binomial distribution in n_R, associated with obtaining n_R successes in n trials.

The first ratio in (26) is an even function of $\bar{f}$, corresponding to a symmetric probability distribution over $\bar{f}$ which is peaked around $\bar{f} = 0$. However, this symmetry is broken by the second pair of factors containing x, which pushes the maximum of the distribution towards positive or negative $\bar{f}$, depending on the initial distribution of marbles (N_R, N_L). In this weight provided by these two factors, we can see evidence of an emergent arrow of time: For asymmetric initial distributions of marbles, we are most likely to see a value of $\bar{f}$ which corresponds to a relaxation back towards an even distribution. Furthermore, the longer that we wait, corresponding to larger values of n, the larger the most likely value of $\bar{f}$. This effect is demonstrated in Figure 5, which shows plots of $P(\bar{f}|N_R, N)$ at $x = \frac{3}{4}$ for various values of n.

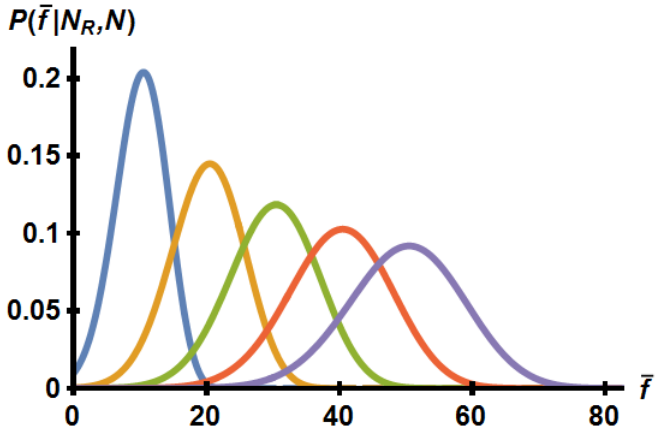

Figure 5: Plots of the likelihood function $P(\bar{f}|N_R, N)$ as a probability distribution over $\bar{f}$, at $x = \frac{3}{4}$, for five values of n. From left to right, the curves plotted are for $n = 20$ (blue), $n = 40$ (orange), $n = 60$ (green), $n = 80$ (red), and $n = 100$ (purple). For the purposes of easy visualization, $P(\bar{f}|N_R, N)$ has been plotted here as a continuous function, although $\bar{f}$ can actually only take on integer values.

However, for the purposes of our memory inference, we are most interested in the behavior of the likelihood as a function of x. This dependence on x determines how the posterior distribution $P(N_R|\bar{f}, N)$ is weighted relative to the prior $P(N_R|N)$, due to the observer's knowledge of $\bar{f}$. Specifically, we would like know the conditions under which the likelihood is tightly peaked around a particular value of N_R. If the likelihood, and therefore the posterior distribution, exhibits a narrow peak for some value of x, this suggests a high confidence in the claim that the initial marble distribution corresponded to this value. In Figure 6, we plot the likelihood as a function of x for various values of $\bar{f}$ at $n = 50$, to demonstrate how the location and sharpness of the peak in the likelihood can vary with changing $\bar{f}$.

A contour plot of the likelihood, as a function of x and $\bar{f}$, is given in Figure 7. Here, we can see that

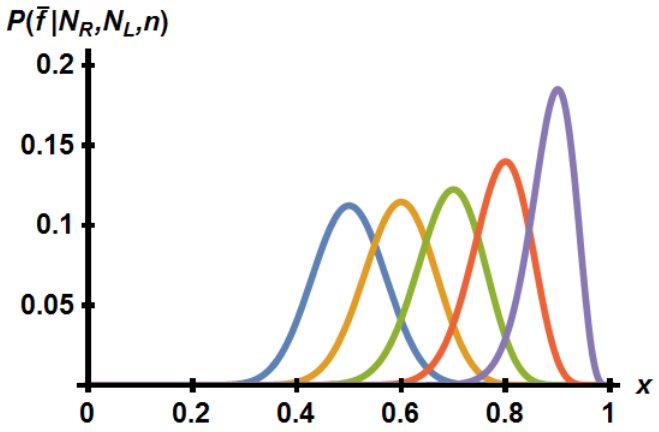

Figure 6: Plots of the likelihood function $P(\bar{f}|N_R, N)$ as a function of x, at $n = 50$, for five values of $\bar{f}$. From left to right, the curves plotted are for $\bar{f} = 0$ (blue), $\bar{f} = 10$ (orange), $\bar{f} = 20$ (green), $\bar{f} = 30$ (red), and $\bar{f} = 40$ (purple). As in Figure 5, for the purposes of easy visualization, $P(\bar{f}|N_R, N)$ has been plotted here as a continuous function, although x can actually only take on values that are multiples of $1/N$.

the likelihood function exhibits a diagonal ridge of maximum values. Each point on this ridge corresponds to the peak value of the likelihood as a a function of x, for a given value of $\bar{f}$. By taking the derivative with respect to x of our expression for the likelihood in (26) and setting it equal to zero, we find that these maximal values lie on the line given by

$$\bar{f}_* = 2n\left(x - \frac{1}{2}\right). \tag{27}$$

To understand how the width of the likelihood as a function of x varies along this ridge, we now consider the limit of large n. Since (26) takes the form of a binomial distribution in n_R, for large n we can approximate the likelihood as a normal distribution in n_R, with mean nx and variance $nx(1-x)$. Proceeding with this approximation, and then making the substitutions $n_R = (n+\bar{f})/2$ and $\bar{f}_* = 2n\left(x - \frac{1}{2}\right)$, yields the expression

$$P(\bar{f}|N_R, N) \approx \frac{1}{\sqrt{2\pi nx(1-x)}} \exp\left[-\frac{(\bar{f} - \bar{f}_*)^2}{8nx(1-x)}\right]. \tag{28}$$

So we see that the likelihood is a normal distribution in $\bar{f}$, with a mean of $\bar{f}_* = 2n\left(x - \frac{1}{2}\right)$ and variance of $4nx(1-x)$. Expressed in terms of x instead of $\bar{f}_*$, this becomes

$$P(\bar{f}|N_R, N) \approx \frac{1}{\sqrt{2\pi nx(1-x)}} \exp\left[-\frac{n\left(x - \frac{n+\bar{f}}{2n}\right)^2}{2x(1-x)}\right]. \tag{29}$$

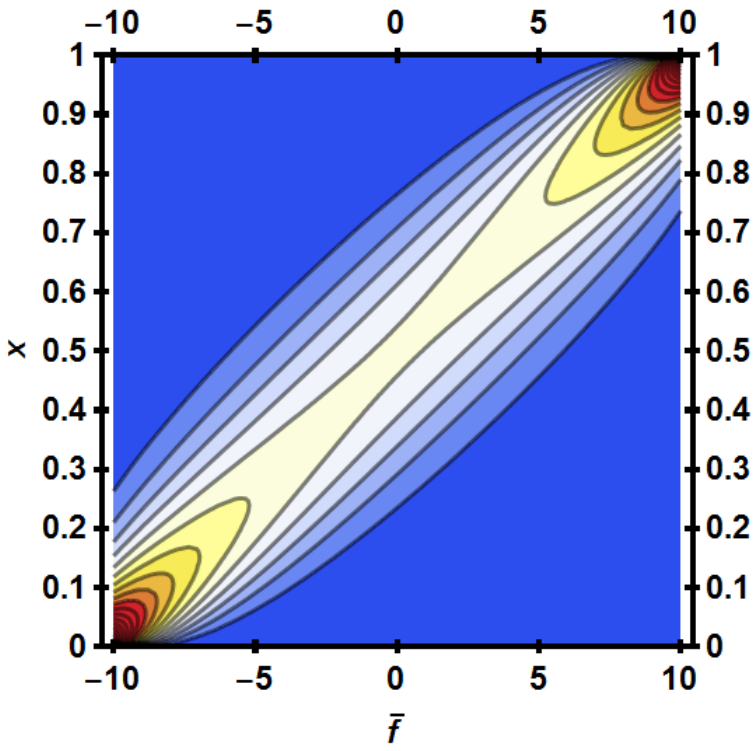

Figure 7: Contour plot of the likelihood function $P(\bar{f}|N_R, N)$, as a function of x and $\bar{f}$ for $n = 10$. The blue shading in the upper left and lower right corners corresponds to small values, and the yellow and red shading along the diagonal ridge corresponds to large values. As in Figures 5 and 6, $P(\bar{f}|N_R, N)$ has been plotted here as a continuous function in $\bar{f}$ and x.

This expression for the likelihood is approximately normally distributed in x, albeit with an effective variance $\sigma_n^2(x) = x(1-x)/n$ that depends on the value of x near the peak of the likelihood. We are now in a position to quantify the informativeness of the inference about x, and therefore the initial marble distribution (N_R, N_L). Substituting our expression for the likelihood in (29) into our equation for the posterior distribution given in (11), we obtain

$$P(N_R|\bar{f}, N) = \frac{1}{\sqrt{2\pi n x(1-x)}} \frac{P(N_R|N)}{P(\bar{f}|N)} \exp\left[-\frac{n\left(x - \frac{n+\bar{f}}{2n}\right)^2}{2x(1-x)}\right]. \tag{30}$$

The informativeness of this posterior function depends on the width of the distribution over the possible values of (N_R, N_L). This width scales with the variance of the posterior, which we denote as σ_{post}^2. To get a sense of how the width of the posterior distribution may vary, suppose for concreteness that the prior is approximately normally distributed, with mean x_0 and variance σ_{prior}^2, so that we have

$$P(N_R|N) \propto \exp\left[-\frac{(x-x_0)^2}{2\sigma^2_{prior}}\right]. \tag{31}$$

Note that now that the prior is given, we have completely specified the dependence of the posterior on x, and consequently its dependence on (N_R, N_L) as well. Computing the normalization $P(\bar{f}|N)$ is not important for our purposes, since it does not vary with x, and therefore does not effect the shape of the posterior as a probability distribution over the possible marble distributions (N_R, N_L).

For the case of the Gaussian prior, the posterior becomes a product of normal distributions, with variance given by

$$\frac{1}{\sigma^2_{post}} \approx \frac{1}{\sigma^2_{prior}} + \frac{1}{\sigma^2_n}(x_0) = \frac{1}{\sigma^2_{prior}} + \frac{n}{x_0(1-x_0)}. \tag{32}$$

Here, we have substituted the value $x = x_0$ into the variance $\sigma^2_n(x_0)$ of the likelihood, since the prior distribution peaks at x_0. Supposing that the variance of the prior is fixed, we can consider this expression for σ^2_{post} in two different limits. First, for a given value of x_0, if we take n to be very large, to compensate we must have a small variance σ^2_{post} for the posterior. This result tells us that as the number of time-steps n increases, the inference about the initial marble distribution becomes more informative. Second, we can look at the behavior of σ_{post} for fixed n, as we allow x_0 to vary between 0 and 1. This is the more interesting limit, as it demonstrates how the informativeness of the inference changes with our knowledge of an arrow of time.

As x_0 approaches either 0 or 1, the likelihood term $\sigma^2_n(x_0)$ in (32) grows without bound. For a value of x_0 near these extremes, the prior identifies the state of the system at $t = 0$ as a low entropy state. In this case, the magnitude of the likelihood term must be compensated by a small value of σ^2_{post}, indicative of a highly informative inference. This corresponds to a "memory of the past" in the thermodynamic sense, where we identify the past with a state of relatively low entropy. In contrast, values of x_0 near $\frac{1}{2}$ yield inferences which are far less informative for a given value of σ^2_{prior} and n, since the denominator of the likelihood term will be much larger than in the $x_0 \to 0, 1$ limit. In these cases, the prior singles out a high-entropy state as most likely at $t = 0$. As a result, the relevant inference is not nearly as sharp; we identify this as a "memory of the future."

3 A Quantum Memory Model

Having completed our analysis of the classical urn model, we now move to develop a quantum mechanical model which will serve as a prototypical quantum memory system. We define this quantum memory to record the net change in some coarse-grained property of a system which it is coupled to, in analogy with Mlodinow and Brun's rotor memory. In order to understand the informativeness of such a memory system, we consider how a measurement of the memory system at one time conditions inferences about the state of the coupled system at a different time. This analysis is carried out in the density matrix formulation of quantum mechanics. Although our work with the quantum memory system is not yet complete, we are able to show that memory inferences in the quantum model can be expressed in terms of a probabilistic matrix completion problem.

3.1 Characterization of the model

We consider two quantum systems: A system S, and a record system M which functions as a memory of some aspect of S. For concreteness, we assume that the Hilbert space of possible states accessible to S is spanned by some orthonormal basis $\{|\mathbf{j}\rangle_S\} = \{|\mathbf{j}_1\rangle_S, |\mathbf{j}_2\rangle_S, |\mathbf{j}_3\rangle_S ...\}$, where the index $\mathbf{j}$ takes on values within some indexing set $\mathbf{J}$. We write $\mathbf{j}$ as a boldface index to emphasize that the indexing set $\mathbf{J}$ could be multidimensional. For example, each value of $\mathbf{j}$ might correspond to an ordered pair $\mathbf{j} = (j_1, j_2)$, where each element of the ordered pair is associated with an independent degree of freedom of S. Suppose that for each possible basis state of S, we can associate a definite integer value of some physical quantity f to this state, so that the value of f corresponding to the state $|\mathbf{j}\rangle_S$ is given by some function $f(\mathbf{j})$.

We define the interaction between S and M so that changes in the value of $f(\mathbf{j})$ can be inferred from knowledge of the state of M. The function $f(\mathbf{j})$ in this quantum memory model will play a role analogous to that of Mlodinow and Brun's readout function $f_{read}(r(t))$, and that of the memory function $f(n)$ in the urn model in Chapter 2. That is, just as the functions $f_{read}(r(t))$ and $f(n)$ do not encode every detail about each particular state of the gas in Mlodinow and Brun's canisters, or the location of every marble in the urns, the value $f(\mathbf{j})$ does not necessarily capture all the properties of a given state $|\mathbf{j}\rangle_S$ of S. Rather, these functions refer to *coarse-grained* properties of their respective systems, which are insensitive to all the precise details

of any given state. Often, such properties correspond to average or aggregate quantities taken over a large number of a system's degrees of freedom. For example, the total spin angular momentum of a many-body quantum system is such a coarse-grained quantity, since many different internal configurations of such a system can correspond to the same net value of the spin.

We can get a clearer sense of the properties of this function $f(\mathbf{j})$ by means of a simple visualization. Consider Figure 8, which represents a subset of the set of basis states $\{|\mathbf{j}\rangle_S\}$ of S in terms of a grid. Each square on the grid corresponds to a basis state with a different value of $\mathbf{j}$. In this example, we have assumed that $\mathbf{j} = (j_1, j_2)$ is a two-component index, so that the degrees of freedom corresponding to j_1 and j_2 can be varied independently by moving parallel to one axis or the other. What we see from this representation is that the function $f(\mathbf{j})$ induces a partition on this set of basis states, with each set in the partition marked by a different color. States with the same color have the same value of the function $f(\mathbf{j})$, and therefore have the same value of whatever property $f(\mathbf{j})$ corresponds to. We would like to construct our memory system M so that it only registers a change when the system S evolves from one coloring in this partition to another, that is, from one value of $f(\mathbf{j})$ to the next. This is analogous to the memory function $f(n)$ in the urn model, which only varies when the coarse-grained variable N_R changes.

To this end, we consider a memory system M with a set of N_M orthonormal basis states $\{|i\rangle_M\} = \{|0\rangle_M, |1\rangle_M ... |N_M - 1\rangle_M\}$, labeled by a quantum number i between 0 and $N_M - 1$. The Hilbert space of the composite system $M \times S$ will then have a basis consisting of the tensor products of basis states of M and S. That is, one useful basis of the Hilbert space of $M \times S$ is defined by

$$\{\ |i\rangle_M \otimes |\mathbf{j}\rangle_S\ \} \equiv \{\ |i\rangle_M \otimes |\mathbf{j}\rangle_S \mid 0 \leq i < N_M,\ \mathbf{j} \in \mathbf{J}\ \}. \tag{33}$$

The coupling between S and M is defined so that when the state of S changes, the change in the value of the memory quantum number i is equal to the change in the value of the quantity f. To see how this constrains the time evolution of the composite system $M \times S$, we consider the evolution of a generic basis state $|i\rangle_M \otimes |\mathbf{j}\rangle_S$ in some definite time interval, between some time $t = t_1$ and a second time $t = t_2$. Assuming that this evolution is defined by the action of some unitary operator $\hat{U}$, it will generically take the form

$$|i\rangle|\mathbf{j}\rangle \rightarrow \hat{U}|i\rangle|\mathbf{j}\rangle = \sum_{i',\mathbf{j}'} U_{i,\mathbf{j},i',\mathbf{j}'}|i'\rangle|\mathbf{j}'\rangle. \tag{34}$$

In this summation and in all subsequent equations, sums over a italicized index can be assumed to range

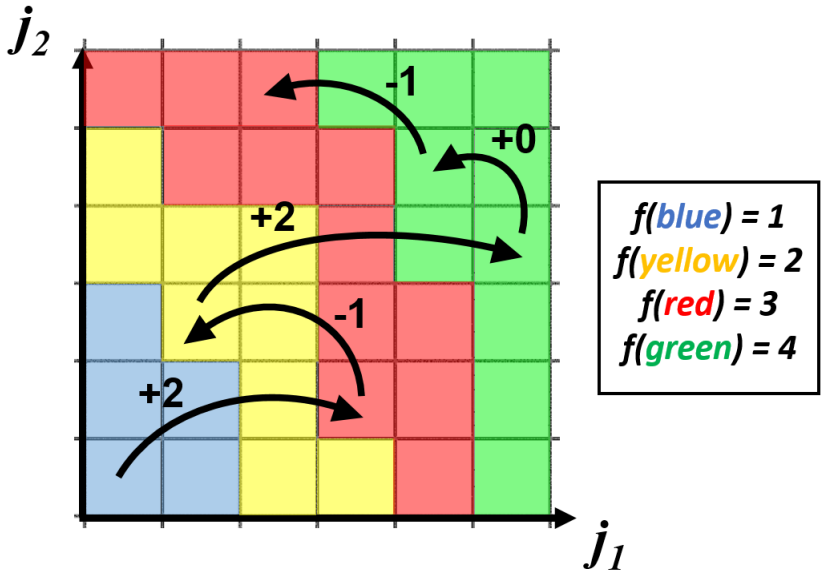

Figure 8: Visualization of the set of basis states of the system S. Each block corresponds to a single basis state, and blocks with the same color have the same value of the quantity $f(\mathbf{j})$, as given in the box on the right. Here, $\mathbf{j} = (j_1, j_2)$ is a two-component index, and each axis corresponds to varying a different degree of freedom of S, associated with either j_1 or j_2. A trajectory through this space is illustrated with arrows; the numbers above the arrows denote how much a memory register coupled to S would increment or decrement during each step.

from 0 to $N_M - 1$ in integer steps, while sums over a boldface index vary over all values in the indexing set $\mathbf{J}$. Here, we have abbreviated the basis states as $|i\rangle|\mathbf{j}\rangle \equiv |i\rangle_M \otimes |\mathbf{j}\rangle_S$, and the coefficients $U_{i,\mathbf{j},i',\mathbf{j}'} \equiv \langle i|\langle \mathbf{j}|\hat{U}|i'\rangle|\mathbf{j}'\rangle$ are the matrix entries of this operator in the basis given in (33). These entries cannot take on arbitrary values, however, since we have specified that any changes in f must be accompanied by equal changes in the memory quantum number. That is, for each term in (34), the change $f(\mathbf{j}') - f(\mathbf{j})$ must be equal to the change $i' - i$, otherwise the term should vanish. This means that for all nonzero terms in (34), we have $i' = i + f(\mathbf{j}') - f(\mathbf{j})$, and consequently the evolution simplifies to the constrained form

$$|i\rangle|\mathbf{j}\rangle \to \hat{U}|i\rangle|\mathbf{j}\rangle = \sum_{\mathbf{j}'} U_{i,\mathbf{j},i+f(\mathbf{j}')-f(\mathbf{j}),\mathbf{j}'}|i + f(\mathbf{j}') - f(\mathbf{j})\rangle|\mathbf{j}'\rangle. \tag{35}$$

Since the matrix elements $U_{i,\mathbf{j},i',\mathbf{j}'}$ vanish when $i' - i \neq f(\mathbf{j}') - f(\mathbf{j})$, we can also write this constraint simply as $U_{i,\mathbf{j},i',\mathbf{j}'} = U_{i,\mathbf{j},i',\mathbf{j}'}\delta_{i'-i,f(\mathbf{j}')-f(\mathbf{j})}$. Beyond the restriction, we allow $\hat{U}$ to be an arbitrary unitary operator. To be precise, we must note that there is one class of evolutions for which (35) is ill-defined. These occur

when $i+f(\mathbf{j}')-f(\mathbf{j})$ either falls below 0 or rises above N_M-1. These are invalid evolutions, since the memory system quantum number can only take on N_M distinct values. To account for this possibility, we require that when the memory "overflows" or "bottoms out" in this manner, the memory quantum number will reset back to 0 or N_M-1. This amounts to replacing the expression $i+f(\mathbf{j}')-f(\mathbf{j})$ in (35) with $(i+f(\mathbf{j}')-f(\mathbf{j}))$ mod N. In the remaining analysis, we will suppress this detail when convenient, since ignoring it will have little effect on our results.

In order to make the effects of the constraint in (35) more readily apparent, we introduce an ordering of the basis states of the composite system $M \times S$. This ordering will allow us to group the matrix entries of $\hat{U}$ into $(N_M)^2$ different blocks, each of which is a sub-matrix in itself that will take a particularly simple form. Consider the set of basis states $\{|i,\mathbf{j}\rangle\}$ defined as

$$\{|i,\mathbf{j}\rangle\} = \{\ |i,\mathbf{j}\rangle = |(i+f(\mathbf{j})-f(\mathbf{j}_0)) \bmod N_M\rangle|\mathbf{j}\rangle \mid 0 \leq i < N_M,\ \mathbf{j} \in \mathbf{J}\ \}. \tag{36}$$

In this set, $\mathbf{j}_0$ is some arbitrary fixed value in $\mathbf{J}$, which we can choose so that $f(\mathbf{j}_0)=0$ if we wish. This set is in fact the same basis set as defined in (33), only with a new notation to emphasize a particular ordering of the basis vectors. We can think of these basis vectors as arranged as follows. First, we take all the basis vectors $|i,\mathbf{j}\rangle$ with $i=0$, and arrange them in accordance with some arbitrary ordering $\mathbf{j}=\mathbf{j}_1,\mathbf{j}_2,\mathbf{j}_3...$ of the values of $\mathbf{j}$ in the index set $\mathbf{J}$. These are the first vectors in our ordering. We then place all the vectors with $i=1$ after this first subset, organized by the same ordering of $\mathbf{j}$ values as for $i=0$. We continue this until we have reached the subset where $i=N_M-1$, at which point we have arranged all of our basis vectors. The complete ordering will take this form:

$$|i,\mathbf{j}\rangle = |0,\mathbf{j}_1\rangle, |0,\mathbf{j}_2\rangle, |0,\mathbf{j}_3\rangle... \ |1,\mathbf{j}_1\rangle, |1,\mathbf{j}_2\rangle, |1,\mathbf{j}_3\rangle... \ |N_M-1,\mathbf{j}_1\rangle, |N_M-1,\mathbf{j}_2\rangle, |N_M-1,\mathbf{j}_3\rangle... \tag{37}$$

What does this permutation of the basis vectors correspond to physically? For each value of i, this arrangement includes a subset of basis vectors that is that same size as the basis $\{|\mathbf{j}\rangle_S\}$ of S. We will call this the i^{th} block of vectors in the ordered basis. Each basis state $|i,\mathbf{j}\rangle = |(i+f(\mathbf{j})-f(\mathbf{j}_0)) \bmod N_M\rangle|\mathbf{j}\rangle$ in this block corresponds to a term in the superposition obtained when evolving $M \times S$ from the state $|i,\mathbf{j}_0\rangle$.

Ordering the basis elements in this way allows the action of the evolution operator $\hat{U}$ to be expressed in a simple form. The basis states are now separated into N_M different blocks, each containing basis states $|i,\mathbf{j}\rangle$ with a fixed value of i and all possible values of $\mathbf{j}$. As a result, we can now also think of the matrix

entries of U as divided into blocks. In our new basis ordering, the matrix elements of $\hat{U}$ are defined as $U'_{i\mathbf{j},i'\mathbf{j}'} \equiv \langle i,\mathbf{j}|\hat{U}|i',\mathbf{j}'\rangle$, where we have included a prime to distinguish these elements from our original matrix elements $U_{i\mathbf{j},i'\mathbf{j}'}$. Consider the set of all such matrix entries for a given value of i and i'. These entries will form a contiguous square block, since the ordering of our basis insures that basis vectors $|i,\mathbf{j}\rangle$ with the same value of i will be in a single block. This block is itself a matrix, with entries indexed by $\mathbf{j}$ and $\mathbf{j}'$. We refer to this matrix as the block entry of $\hat{U}$ at i and i', and denote it as $\overline{U}_{ii'}$, and we specify its own elements as $(\overline{U}_{ii'})_{\mathbf{j}\mathbf{j}'} \equiv U'_{i\mathbf{j},i'\mathbf{j}'}$. In terms of these block units, the matrix corresponding to the evolution operator $\hat{U}$ in this basis takes the form

$$\hat{U} \sim \begin{bmatrix} \overline{U}_{00} & \overline{U}_{01} & \cdots & \overline{U}_{0,N-1} \\ \overline{U}_{10} & \overline{U}_{11} & \cdots & \overline{U}_{1,N-1} \\ \vdots & \vdots & \ddots & \vdots \\ \overline{U}_{N_M-1,0} & \overline{U}_{N_M-1,1} & \cdots & \overline{U}_{N_M-1,N_M-1} \end{bmatrix} \tag{38}$$

So far, expressing the operator $\hat{U}$ in this form just amounts to a specific ordering and grouping of the matrix elements of $\hat{U}$. However, we can see that this block matrix will simplify considerably by considering the evolution of a basis vector $|i,\mathbf{j}\rangle$ as expanded in the $\{|i,\mathbf{j}\rangle\}$ basis:

$$|i,\mathbf{j}\rangle \to \hat{U}|i,\mathbf{j}\rangle = \sum_{i',\mathbf{j}'} (\overline{U}_{ii'})_{\mathbf{j}\mathbf{j}'}|i',\mathbf{j}'\rangle = \sum_{i',\mathbf{j}'} (\overline{U}_{ii'})_{\mathbf{j}\mathbf{j}'}|i' + f(\mathbf{j}') - f(\mathbf{j}_0)\rangle|\mathbf{j}'\rangle \tag{39}$$

Again, we have suppressed the mod N_M operator in $|i,\mathbf{j}\rangle = |(i+f(\mathbf{j})-f(\mathbf{j}_0)) \bmod N_M\rangle|\mathbf{j}\rangle$ for brevity. We can now ask how the constraint expressed in (35), which specifies how the states of the memory system M co-evolve with the states of S, constrains the terms in this expansion. From (35), we know that a state $|i\rangle|\mathbf{j}\rangle$ evolves into a superposition with terms only of the form $|i + f(\mathbf{j}') - f(\mathbf{j})\rangle|\mathbf{j}'\rangle$, so that the quantum number of the memory system state increases by an amount equal to the change in the value of f. This means that the state $|i,\mathbf{j}\rangle = |i + f(\mathbf{j}) - f(\mathbf{j}_0)\rangle|\mathbf{j}\rangle$ that is evolved in (39) can only evolve into a superposition with terms of the form:

$$|(i + f(\mathbf{j}) - f(\mathbf{j}_0)) + f(\mathbf{j}') - f(\mathbf{j})\rangle|\mathbf{j}'\rangle = |i + f(\mathbf{j}') - f(\mathbf{j}_0)\rangle|\mathbf{j}'\rangle = |i,\mathbf{j}'\rangle \tag{40}$$

So, a basis element $|i,\mathbf{j}\rangle$ under the evolution operator $\hat{U}$ evolves into a superposition of *only* basis vectors with the same value i. That is, the evolution induced by $\hat{U}$ does not send states from one block of our basis

ordering into another. Specifically, this means that all terms in (39) must vanish except for those for which $i' = i$, and so the entries $(\overline{U}_{ii'})_{\mathbf{jj'}}$ are zero unless $i' = i$. This simplifies the representation of $\hat{U}$ in our ordered basis into a block-diagonal matrix, where each off-diagonal block element is just the zero matrix $\overline{0}$ with all vanishing entries. Using the abbreviation $\overline{U}_i \equiv \overline{U}_{ii}$ for the diagonal elements, we then have:

$$\hat{U} \sim \begin{bmatrix} \overline{U}_0 & \overline{0} & \dots & \overline{0} \\ \overline{0} & \overline{U}_1 & \dots & \overline{0} \\ \vdots & \vdots & \ddots & \vdots \\ \overline{0} & \overline{0} & \dots & \overline{U}_{N-1} \end{bmatrix} = diag(\overline{U}_0, \overline{U}_1 ... \overline{U}_{N-1}) \tag{41}$$

Moreover, the assumption that $\hat{U}$ is a unitary operator implies that each of these diagonal block entries $\overline{U}_i$ is also a unitary matrix. We can see this be computing the product $\hat{U}\hat{U}^\dagger$ in our ordered basis, which is equal to the identity operator $\hat{\mathbb{1}}$ since $\hat{U}$ is unitary. Noting that $\hat{U}^\dagger = diag(\overline{U}_0^\dagger, \overline{U}_1^\dagger ... \overline{U}_{N-1}^\dagger)$, we can see that the block entries $(\hat{U}\hat{U}^\dagger)_{kk'}$ of the product $\hat{U}\hat{U}^\dagger$ are given by

$$(\hat{U}\hat{U}^\dagger)_{kk'} = \sum_i (\hat{U})_{ki} (\hat{U}^\dagger)_{ik'}. \tag{42}$$

Here, we have performed the matrix multiplication blockwise. In a blockwise multiplication, the block entries of the product are computed in terms of the block entries of the factors, in the same form as a standard matrix multiplication. We have denoted the block element of $\hat{U}$ with indices k and i as $(\hat{U})_{ki}$, and the block element of $\hat{U}^\dagger$ with indices i and k' as $(\hat{U}^\dagger)_{ik'}$. We will continue to use this notation in the next section: Unless otherwise specified, we surround a matrix with parentheses and two indices in order to signify the block entry of that matrix specified by those indices.

Notation aside, we find that the block entries of $\hat{U}\hat{U}^\dagger$ given in (42) simplify because $\hat{U}$ and $\hat{U}^\dagger$ are both block-diagonal. Specifically, the sum in (42) collapses to a single term, since $(\hat{U})_{ki}$ and $(\hat{U}^\dagger)_{ik'}$ are equal to the zero matrix unless $k = i = k'$. Noting that $(\hat{U})_{kk} = \overline{U}_k$ and $(\hat{U}^\dagger)_{k'k'} = \overline{U}_{k'}^\dagger$, we have the result

$$(\hat{U}\hat{U}^\dagger)_{kk'} = \delta_{kk'} (\hat{U})_{kk} (\hat{U}^\dagger)_{k'k'} = \delta_{kk'} \overline{U}_k \overline{U}_k^\dagger. \tag{43}$$

However, we know that since $\hat{U}\hat{U}^\dagger = \hat{\mathbb{1}}$, the block entries of $\hat{U}\hat{U}^\dagger$ are given by identity matrices $\overline{I}$ along the diagonal, and by $\overline{0}$ otherwise. This means that $(\hat{U}\hat{U}^\dagger)_{kk'} = \delta_{kk'}\overline{I}$, and equating this expression with (43) yields $\overline{U}_k \overline{U}_k^\dagger = \overline{I}$. This is just the condition for the unitary of the matrix $\overline{U}_k$.

To conclude this section, we note that there is no reason why the various block entries of $\hat{U}$ should not be equal to one another. Under the action of such an operator, all blocks of our ordered basis would evolve in the exact same manner. In this case, given an initial state of the system S, the states of S in the superposition produced by the action of $\hat{U}$ will be independent of the initial state of M. So in fact, an operator $\hat{U}$ for which the block entries $\overline{U}_i$ are the same or almost the same could be said to correspond to the most useful sort of memory, since this constraint means that the evolution of S is hardly effected by the coupling to M. However, for our analysis, we will still consider the general case with nonidentical block entries of $\hat{U}$, since an interaction of some system with a memory need not leave that system completely undisturbed.

3.2 Probabilities in the quantum model

In analogy with our analysis of the classical urn model, we now wish to understand how observation of the quantum memory system M at one time allows for inference about the state of the system S at another time. To this end, we describe the systems M and S primarily in terms of density operators, as opposed to the state vectors that we have been working with until now. An overview of the density operator formalism is given in Appendix B. This approach is useful in our present context for two reasons. First, the density operator description easily accommodates a probabilistic analysis of the state of S. A general density operator encodes not only probabilities arising from fundamental quantum indeterminism, but also probabilities which are assigned based on a lack of exact information about a system. Second, it allows us to describe the systems M and S in terms of two separate mathematical objects, namely each system's reduced density operator. This is not possible when working with state vectors alone, since if M and S become entangled, a state vector can only be meaningfully assigned to the composite system $M \times S$.

To begin, suppose that at some reference time t_1, the composite system $M \times S$ is in some separable state, so that the total density operator $\hat{\rho}(t_1) = \hat{\rho} = \hat{\rho}^M \otimes \hat{\rho}^S$ for $M \times S$ is simply the tensor product of the density operators $\hat{\rho}^M$ and $\hat{\rho}^S$, corresponding to the systems M and S respectively. While the state of S is described by some unknown density operator $\hat{\rho}^S$, we assume that M is in a known pure state $\hat{\rho}^M = |m_0\rangle_M \langle m_0|_M$.

Now, under the evolution described in the previous section, the composite system $M \times S$ will evolve into a new state $\hat{\rho}(t_2) = \hat{U}\hat{\rho}\hat{U}^\dagger$ by some second time t_2. Suppose now that a measurement of the quantum number associated with the memory system M is made, so that M collapses back into a new pure state $|m\rangle_M \langle m|_M$. Given the initial pure state $\hat{\rho}^M = |m_0\rangle_M \langle m_0|_M$ of M and the evolution operator $\hat{U}$, we would like to understand how the measurement of M at t_2 informs any inference we can make about the possible

entries of $\hat{\rho}^S$, and therefore the state of S, at t_1.

Our first task is to calculate how the density operator $\hat{\rho}$ will evolve between t_1 and t_2. This evolution will look simplest in the ordered basis given in (36). From the analysis in the previous section, we know that $\hat{U}$ takes a block diagonal form in this basis. We can obtain the entries of $\hat{\rho}$ in this basis as follows. First, we expand our definition $\hat{\rho} = \hat{\rho}^M \otimes \rho^S$ in terms of basis states of M and S, yielding

$$\hat{\rho} = \hat{\rho}^M \otimes \hat{\rho}^S = |m_0\rangle_M\langle m_0|_M \otimes \left(\sum_{\mathbf{j},\mathbf{j}'} \rho^S_{\mathbf{j}\mathbf{j}'} |\mathbf{j}\rangle_S \langle \mathbf{j}'|_S \right) = \sum_{\mathbf{j},\mathbf{j}'} \rho^S_{\mathbf{j}\mathbf{j}'} |m_0\rangle|\mathbf{j}\rangle\langle m_0|\langle \mathbf{j}'|. \tag{44}$$

Here, the coefficients $\rho^S_{\mathbf{j}\mathbf{j}'} \equiv \langle \mathbf{j}|_M \hat{\rho}^S |\mathbf{j}'\rangle_M$ are the entries of $\hat{\rho}^S$ in the $\{|\mathbf{j}\rangle_S\}$ basis. However, we also know that we can expand $\hat{\rho}$ as the sum of the reordered basis states

$$\hat{\rho} = \sum_{i,i',\mathbf{j},\mathbf{j}'} (\overline{\rho}_{ii'})_{\mathbf{j}\mathbf{j}'} |i,\mathbf{j}\rangle\langle i',\mathbf{j}'| = \sum_{i,i',\mathbf{j},\mathbf{j}'} (\overline{\rho}_{ii'})_{\mathbf{j}\mathbf{j}'} |i + f(\mathbf{j}) - f(\mathbf{j}_0))\rangle|\mathbf{j}\rangle\langle i' + f(\mathbf{j}') - f(\mathbf{j}_0)|\langle \mathbf{j}'|. \tag{45}$$

Just as we expressed the entries of $\hat{U}$ in terms of the block matrices $\overline{U}_{ii'}$, we define $(\overline{\rho}_{ii'})_{\mathbf{j}\mathbf{j}'} \equiv \langle i,\mathbf{j}|\hat{\rho}|i',\mathbf{j}'\rangle$ as the entry at row $\mathbf{j}$ and column $\mathbf{j}'$ of the block matrix $\overline{\rho}_{ii'}$. In terms of these block entries, the density operator is represented as:

$$\hat{\rho} \sim \begin{bmatrix} \overline{\rho}_{00} & \overline{\rho}_{01} & \cdots & \overline{\rho}_{0,N-1} \\ \overline{\rho}_{10} & \overline{\rho}_{11} & \cdots & \overline{\rho}_{1,N-1} \\ \vdots & \vdots & \ddots & \vdots \\ \overline{\rho}_{N-1,0} & \overline{\rho}_{N-1,1} & \cdots & \overline{\rho}_{N-1,N-1} \end{bmatrix} \tag{46}$$

Now, after equating terms in (44) and (45) with the same values of $\mathbf{j}$ and $\mathbf{j}'$, we find the relation

$$\rho^S_{\mathbf{j}\mathbf{j}'} |m_0\rangle\langle m_0| = \sum_{i,i'} (\overline{\rho}_{ii'})_{\mathbf{j}\mathbf{j}'} |i + f(\mathbf{j}) - f(\mathbf{j}_0))\rangle\langle i' + f(\mathbf{j}') - f(\mathbf{j}_0)|. \tag{47}$$

Comparing the terms of this sum with $\rho^S_{\mathbf{j}\mathbf{j}'}|m_0\rangle\langle m_0|$, we can now see that the only nonzero term will be that for which $m_0 = i + f(\mathbf{j}) - f(\mathbf{j}_0) = i' + f(\mathbf{j}') - f(\mathbf{j}_0)$. This means that if these equalities are satisfied, then we simply have $(\overline{\rho}_{ii'})_{\mathbf{j}\mathbf{j}'} = \rho^S_{\mathbf{j}\mathbf{j}'}$. Otherwise, we must have $(\overline{\rho}_{ii'})_{\mathbf{j}\mathbf{j}'} = 0$ if the appropriate terms are to vanish. Symbolically, our result is:

$$(\overline{\rho}_{ii'})_{\mathbf{j}\mathbf{j}'} = \rho^S_{\mathbf{j}\mathbf{j}'} \delta_{m_0, i+f(\mathbf{j})-f(\mathbf{j}_0)} \delta_{m_0, i'+f(\mathbf{j}')-f(\mathbf{j}_0)}. \tag{48}$$

Now that all the entries of $\hat{\rho}$ in our ordered basis have been specified, determining the evolution of $\hat{\rho}$ is just a matter of computing the product $\hat{U}\hat{\rho}\hat{U}^\dagger$. Our goal is to calculate the individual entries of this product, $((\hat{U}\hat{\rho}\hat{U}^\dagger)_{\alpha\alpha'})_{\boldsymbol{\beta}\boldsymbol{\beta}'} \equiv \langle \alpha, \boldsymbol{\beta}|\hat{U}\hat{\rho}\hat{U}^\dagger|\alpha', \boldsymbol{\beta}'\rangle$, where α and α' take on the integer values from 0 to $N_M - 1$, and $\boldsymbol{\beta}$ and $\boldsymbol{\beta}'$ are in the indexing set $\mathbf{J}$. As an intermediate step, we first compute the block entry of $\hat{U}\hat{\rho}\hat{U}^\dagger$ at row α and column α', denoted $(\hat{U}\hat{\rho}\hat{U}^\dagger)_{\alpha\alpha'}$. As noted previously, this computation has the exact same form as a standard matrix multiplication, but with the individual entries given as matrices instead of single numbers. Proceeding this way, we obtain the result

$$(\hat{U}\hat{\rho}\hat{U}^\dagger)_{\alpha\alpha'} = \sum_{k,l} \overline{U}_{\alpha k}\overline{\rho}_{kl}\overline{U}^\dagger_{l\alpha'} = \overline{U}_{\alpha}\overline{\rho}_{\alpha\alpha'}\overline{U}^\dagger_{\alpha'}, \tag{49}$$

where the sum collapses to a single term because of the block diagonal form of $\hat{U}$, which ensures that $\overline{U}_{ii'} = \overline{U}_i$ is the zero matrix unless $i = i'$. Proceeding to the computation of $((\hat{U}\hat{\rho}\hat{U}^\dagger)_{\alpha\alpha'})_{\boldsymbol{\beta}\boldsymbol{\beta}'}$, we find that

$$((\hat{U}\hat{\rho}\hat{U}^\dagger)_{\alpha\alpha'})_{\boldsymbol{\beta}\boldsymbol{\beta}'} = (\overline{U}_{\alpha}\overline{\rho}_{\alpha\alpha'}\overline{U}^\dagger_{\alpha'})_{\boldsymbol{\beta}\boldsymbol{\beta}'} = \sum_{\mathbf{k},\mathbf{l}} (\overline{U}_{\alpha})_{\boldsymbol{\beta}\mathbf{k}}(\overline{\rho}_{\alpha\alpha'})_{\mathbf{k}\mathbf{l}}(\overline{U}^\dagger_{\alpha'})_{\mathbf{l}\boldsymbol{\beta}'}. \tag{50}$$

Noting that $(\overline{U}^\dagger_{\alpha'})_{\mathbf{l}\boldsymbol{\beta}'} = (\overline{U}_{\alpha'})^*_{\boldsymbol{\beta}'\mathbf{l}}$, where $*$ denotes complex conjugation, and substituting our expression for $(\overline{\rho}_{ii'})_{\mathbf{j}\mathbf{j}'}$ given in (48), our final expression for $((\hat{U}\hat{\rho}\hat{U}^\dagger)_{\alpha\alpha'})_{\boldsymbol{\beta}\boldsymbol{\beta}'}$ is

$$((\hat{U}\hat{\rho}\hat{U}^\dagger)_{\alpha\alpha'})_{\boldsymbol{\beta}\boldsymbol{\beta}'} = \sum_{\mathbf{k},\mathbf{l}} (\overline{U}_{\alpha})_{\boldsymbol{\beta}\mathbf{k}}\rho^S_{\mathbf{k}\mathbf{l}}\delta_{m_0,\alpha+f(\mathbf{k})-f(\mathbf{j}_0)}\delta_{m_0,\alpha'+f(\mathbf{l})-f(\mathbf{j}_0)}(\overline{U}_{\alpha'})^*_{\boldsymbol{\beta}'\mathbf{l}}. \tag{51}$$

Given these entries for the evolved density matrix, we can now evaluate the probabilities $P(m|\hat{\rho}^S)$ of obtaining various values m when the measurement of M is made at t_2. As described in Appendix B, these probabilities are given by $P(m|\hat{\rho}^S) = \mathrm{Tr}\left[\hat{P}_m\hat{U}\hat{\rho}\hat{U}^\dagger\right]$, where the operator $\hat{P}_m$ is the projection operator onto the space of states of M with quantum number m, and the symbol Tr denotes the trace operation. In this case, the basis state $|m\rangle_M$ is the unique state described by this quantum number, so $\hat{P}_m$ is simply $|m\rangle_M\langle m|_M$. The trace of $|m\rangle_M\langle m|_M\hat{U}\hat{\rho}\hat{U}^\dagger$, expressed as a sum over the original basis states, is given by

$$P(m|\hat{\rho}^S) = \mathrm{Tr}\left[|m\rangle_M\langle m|_M\hat{U}\hat{\rho}\hat{U}^\dagger\right] = \sum_{i,\mathbf{j}} \langle i|_M\langle \mathbf{j}|_S \left(|m\rangle_M\langle m|_M\hat{U}\hat{\rho}\hat{U}^\dagger\right)|i\rangle_M|\mathbf{j}\rangle_S. \tag{52}$$

Given the orthogonality of the basis states $\{|i\rangle_M\}$, the terms in this sum are zero unless $i = m$, and so we have

$$P(m|\hat{\rho}^S) = \sum_{\mathbf{j}} \langle m|_M \langle \mathbf{j}|_S \hat{U}\hat{\rho}\hat{U}^\dagger |m\rangle_M |\mathbf{j}\rangle_S. \tag{53}$$

Now, the matrix $\hat{U}\hat{\rho}\hat{U}^\dagger$, expressed in terms of its entries $((\hat{U}\hat{\rho}\hat{U}^\dagger)_{\alpha\alpha'})_{\boldsymbol{\beta}\boldsymbol{\beta}'}$, takes the form:

$$\hat{U}\hat{\rho}\hat{U}^\dagger = \sum_{\alpha,\alpha',\boldsymbol{\beta},\boldsymbol{\beta}'} ((\hat{U}\hat{\rho}\hat{U}^\dagger)_{\alpha\alpha'})_{\boldsymbol{\beta}\boldsymbol{\beta}'} |\alpha, \boldsymbol{\beta}\rangle\langle\alpha', \boldsymbol{\beta}'|, \tag{54}$$

$$\hat{U}\hat{\rho}\hat{U}^\dagger = \sum_{\alpha,\alpha',\boldsymbol{\beta},\boldsymbol{\beta}'} ((\hat{U}\hat{\rho}\hat{U}^\dagger)_{\alpha\alpha'})_{\boldsymbol{\beta}\boldsymbol{\beta}'} |\alpha + f(\boldsymbol{\beta}) - f(\mathbf{j}_0)\rangle_M |\boldsymbol{\beta}\rangle_S \langle\alpha' + f(\boldsymbol{\beta}') - f(\mathbf{j}_0)|_M \langle\boldsymbol{\beta}'|_S. \tag{55}$$

Upon substituting this expression for $\hat{U}\hat{\rho}\hat{U}^\dagger$ into the equation for $P(m)$ in (53), we note that the only non-vanishing terms will be those for which $m = \alpha + f(\boldsymbol{\beta}) - f(\mathbf{j}_0) = \alpha' + f(\boldsymbol{\beta}') - f(\mathbf{j}_0)$ and $l = \boldsymbol{\beta} = \boldsymbol{\beta}'$, due to the orthogonality of the basis states associated with both M and S. Thus we are left with

$$P(m|\hat{\rho}^S) = \sum_{\mathbf{j}} ((\hat{U}\hat{\rho}\hat{U}^\dagger)_{m-f(\mathbf{j})+f(\mathbf{j}_0),m-f(\mathbf{j})+f(\mathbf{j}_0)})_{\mathbf{j}\mathbf{j}}. \tag{56}$$

This probability only depends on the diagonal entries of $\hat{U}\hat{\rho}\hat{U}^\dagger$. From (51), we can see that these entries take the values

$$((\hat{U}\hat{\rho}\hat{U}^\dagger)_{\alpha\alpha})_{\boldsymbol{\beta}\boldsymbol{\beta}} = \sum_{\mathbf{k},\mathbf{l}} (\overline{U}_\alpha)_{\boldsymbol{\beta}\mathbf{k}} \rho^S_{\mathbf{k}\mathbf{l}} \delta_{m_0,\alpha+f(\mathbf{k})-f(\mathbf{j}_0)} \delta_{f(\mathbf{k}),f(\mathbf{l})} (\overline{U}_\alpha)^*_{\boldsymbol{\beta}\mathbf{l}}. \tag{57}$$

Here, we have replaced the product $\delta_{m_0,\alpha+f(\mathbf{k})-f(\mathbf{j}_0)}\delta_{m_0,\alpha'+f(\mathbf{l})-f(\mathbf{j}_0)}$ in (51) with $\delta_{m_0,\alpha+f(\mathbf{k})-f(\mathbf{j}_0)}\delta_{f(\mathbf{k}),f(\mathbf{l})}$, since for $\alpha = \alpha'$ these quantities are equal. Finally, substituting (57) into our expression for $P(m|\hat{\rho}^S)$ in (56), with $\alpha = m - f(\mathbf{j}) + f(\mathbf{j}_0)$ and $\boldsymbol{\beta} = \mathbf{j}$, we find the result

$$P(m|\hat{\rho}^S) = \sum_{\mathbf{j},\mathbf{k},\mathbf{l}} (\overline{U}_{m-f(\mathbf{j})+f(\mathbf{j}_0)})_{\mathbf{j}\mathbf{k}} \rho^S_{\mathbf{k}\mathbf{l}} \delta_{m_0,m+f(\mathbf{k})-f(\mathbf{j})} \delta_{f(\mathbf{k}),f(\mathbf{l})} (\overline{U}_{m-f(\mathbf{j})+f(\mathbf{j}_0)})^*_{\mathbf{j}\mathbf{l}}. \tag{58}$$

Now, collecting the terms with common values of $\rho^S_{\mathbf{k}\mathbf{l}}$, we have:

$$P(m|\hat{\rho}^S) = \sum_{\mathbf{k},\mathbf{l}} \left(\sum_{\mathbf{j}} (\overline{U}_{m-f(\mathbf{j})+f(\mathbf{j}_0)})_{\mathbf{j}\mathbf{k}} (\overline{U}_{m-f(\mathbf{j})+f(\mathbf{j}_0)})^*_{\mathbf{j}\mathbf{l}} \delta_{m_0,m+f(\mathbf{k})-f(\mathbf{j})} \delta_{f(\mathbf{k}),f(\mathbf{l})} \right) \rho^S_{\mathbf{k}\mathbf{l}}, \tag{59}$$

$$P(m|\hat{\rho}^S) = \sum_{\mathbf{k},\mathbf{l}} a^{(m)}_{\mathbf{k}\mathbf{l}} \rho^S_{\mathbf{k}\mathbf{l}}. \tag{60}$$

We have defined the coefficient $a^{(m)}_{\mathbf{k}\mathbf{l}}$ as the sum over $\mathbf{j}$ in parentheses in (59). So, we see that the probability $P(m)$ is a linear combination of the entries of $\hat{\rho}^S$, the initial density operator of the system S.

The coefficients $a^{(m)}_{\mathbf{kl}}$ are known to us, since they depend on the entries of $\hat{U}$, the form of $f(\mathbf{j})$, and the value of m_0, all of which we assume are given. Given this expression $P(m|\hat{\rho}^S)$, which encodes our knowledge about the measurement of M at the time t_2, we can now ask how about how this conditions any inference we can make about $\hat{\rho}^S$, which describes the state of S at the time t_1.

Since $\hat{\rho}^S$ is a density matrix, this already places some constraints on the values of its entries. First, $\hat{\rho}^S$ is necessarily a Hermitian operator, and therefore has fewer independent degrees of freedom than an arbitrary complex operator. To see this, let us assume for simplicity that the matrix representation of $\hat{\rho}^S$ is a finite $N \times N$ matrix. In order to be Hermitian, this matrix must have N real diagonal entries, and $N(N-1)/2$ independent complex values in off-diagonal entries. This means that $\hat{\rho}^S$ is described by $N + N(N-1) = N^2$ real values, each corresponding to an independent degree of freedom. This is half the degrees of freedom available to a general complex $N \times N$ matrix. Second, the eigenvalues of $\hat{\rho}^S$ must all be real numbers between zero and one, and they must sum to one. This constraint insures that the eigenvalues of $\hat{\rho}^S$ can be interpreted as probabilities. Without any additional information, these two restrictions alone vastly underdetermine the values of the entries $\rho^S_{\mathbf{kl}}$. The task of assigning these entries in the most appropriate way, given such constraints, constitutes a *matrix completion* problem. For the purposes of our inference, the solution to this problem is also informed by the sum (60), which relates the matrix entries to one another in terms of the known coefficients $a^{(m)}_{\mathbf{kl}}$ and the probability $P(m|\hat{\rho}^S)$.

So far, we have only just begun to consider our quantum memory model from the standpoint of a matrix completion problem. One way to analyze this problem may be through a Bayesian updating procedure. In this case, the aim is to make the best inference about the values of the matrix elements of $\hat{\rho}^S$. We begin with a prior distribution $P(\hat{\rho}^S)$ over the possible values of the entries $\rho^S_{\mathbf{kl}}$, conditioned by our knowledge that $\hat{\rho}^S$ is a density operator, as well as by any relevant prior information about the systems M and S. Then, after a particular value m of the memory quantum number is obtained in a measurement, we use this value to update our prior distribution over the possible matrix entries, to obtain a posterior distribution $P(\hat{\rho}^S|m)$. Here, the role of the likelihood function is played by $P(m|\hat{\rho}^S)$, since it encodes the probability of a measurement outcome conditioned on knowledge of $\hat{\rho}^S$. Expressed in this form, the memory inferences in our quantum model appear quite analogous to the inferences in the classical urn model.

4 Conclusions

In our analysis, we have attempted to bring out the role of inference in the functioning of physical systems as memories. Any system is capable of interacting with its environment, and when it does so, information about the state of the world is encoded in its internal degrees of freedom. However, this information may be recorded in an unpredictable way, dispersed throughout a set of practically inaccessible degrees of freedom, or quickly scrambled by further environmental influence. Fortunately, in some special cases, the coupling of a system to its surroundings in a regular way allows for predictable correlations between certain properties of the system and certain aspects of the environment. If these correlations can be harnessed to make precise, informative inferences about the state of the world at one time, given observation of some property of the system in question at another time, then we say then the system functions as a memory.

In the case of our classical urn model system, we have carried through an analysis of inferences about the marble distribution between urns, given the output of a coupled memory device at a different time. This has allowed us to understand how such memory inferences depend on the thermodynamic arrow of time associated with the the marble system. Specifically, we find that inferences about the low-entropy past, given information from the relatively high-entropy future, are much more informative in general than inferences in the reverse direction. For our quantum model, although more work still needs to be done, we have developed a framework for understanding how measurements of a quantum memory register allow inferences to be made about a system coupled to the register. We have shown that such inferences can be couched in terms of a matrix completion problem, where the entries of the density matrix of the coupled system must be inferred based on the outcome of a measurement of the memory.

This work builds on that of Mlodinow and Brun in two key ways. First, as we have previously argued, a definition of memory in terms of inference allows us to make sense of Mlodinow and Brun's notion of "generality." For Mlodinow and Brun, a memory system satisfies generality if it still functions as a memory under certain perturbations of the state of the memory's environment; these perturbations may correspond to a range of different possible recorded values. We can then say that such a memory aligns with the thermodynamic arrow of time if "typical" evolutions of the memory under these perturbations all exhibit similar changes in entropy. However, to understand which possible evolutions of the memory and its surroundings

are most likely to occur, and therefore most typical, we have to be able to pose some inference about the how the memory will interact and evolve over time. To specify such inferences in a regular and quantitative manner, we have applied the formalism of Bayesian inference in the context of memory systems.

Second, we have begun to extend this reasoning to quantum systems. In Mlodinow and Brun's paper, the systems which they consider are explicitly classical: The states of both the memory system and its environment are described by points in classical phase space, which precisely specify the positions and velocities of all the constituent parts of these systems. This state of affairs no longer obtains in a quantum memory system. Importantly, the quantum states of a memory system and its environment will generally not be specifiable individually, but only as an entangled quantum state of the composite system composed of the memory and its surroundings. As Mlodinow and Brun point out, their argument assumes that the memory and its environment are separable, in the sense that the state of each system can be specified individually. However, the interactions between two quantum systems which encode information about one system in the degrees of freedom of the other will generally entangle these systems, so that a state vector cannot be assigned to the systems individually. So in the quantum realm, this assumption of separability must be dispensed with. Nevertheless, we ultimately hope to show that this assumption is not critical to the thrust of Mlodinow and Brun's argument.

Given these developments in our work, several possible paths are left open for further study. The most obvious next step is to complete our investigation of the quantum memory system. This would entail an analysis of the matrix completion problem that we described at the end of Chapter 3, in order to determine the most reasonable entries to assign to the density matrix for the system S, given a measurement of the memory system M. Another direction of future research might then involve considering the von Neumann entropy of the quantum memory system and its surroundings. We would like to show, as we have for the urn model, that changes in the entropy of the memory and its surroundings over time condition the possible inferences that can be made about these systems. Finally, it may prove fruitful to generalize the classical results obtained with our urn model. This would involve describing the memory inferences about a classical system more abstractly, without reference to any particular system like the urn model. Our analysis of the urn model was useful because of the simple probabilistic calculations that it permitted, but approaching the problem in more general terms may highlight the essential structure of our arguments.

5 Appendix A: Bayesian Probability Theory

Bayesian probability theory, in the form developed by authors such as E. T. Jaynes, is a mathematical prescription for assigning *degrees of belief* to various inferences, or claims about the world made on the basis of relevant information. These degrees of belief, which quantify the extent to which inferences are supported by the given evidence or information available, are expressed as probabilities between 0 and 1. The standard rules of probability theory allow these degrees of belief to be manipulated and compared. In particular, the Bayesian approach to Bayes' rule of conditional probability is a method by which degrees of belief may be systematically updated, given the existence of new information. This "Bayesian updating" procedure is especially relevant for our analysis of memory systems, since it can be used to understand how inferences about the state of some observed system can change upon access to some "memory" device that is coupled to it. In this appendix, we present an overview of three important concepts in Bayesian theory: The interpretation of probability statements, the Bayesian updating method, and the assignment of "prior" or baseline probabilities. This presentation is guided by the writings of E. T. Jaynes concerning Bayesian probability, particularly his book *Probability Theory: The Logic of Science*.[6]

We begin with a brief note on the meaning of probabilities in Bayesian theory. Bayesian probabilities, as degrees of belief, are not empirical quantities. This interpretation of probabilities may be unfamiliar to some readers, who are accustomed to understanding probability as the frequency with which a given event occurs, relative to some set of other possible outcomes. In such "frequentist" accounts of probability, probabilities are estimated based on the real or hypothetical frequencies of specific events, and then manipulated with the mathematical rules governing probabilities. In contrast, a Bayesian probability is *assigned* to an inference as the result of a logical procedure, which takes as inputs the information relevant to making that inference. In this way, probabilities in the Bayesian formalism are not facts or properties of the world, but rather a metric by which inferences can be quantified and then compared.

However, although the probabilities employed in Bayesian logic are not empirical quantities, they are also more than just subjective judgments or beliefs of some human observer. First of all, Bayesian probabilities cannot be assigned arbitrarily, but must obey both the standard rules of probability and a standard of logical consistency. In some cases, a careful imposition of such constraints can single out a *unique* set of

probabilities as the correct assignment for a given inference. In such situations, the resultant probabilities are completely objective or "user-independent" in the sense that there is only one way to assign them in a logically consistent manner. In fact, Cox's Theorem, first proved in Richard T. Cox's 1961 work *Algebra of Probable Inference*, is discussed extensively in the first two chapters of Jaynes' book. This theorem proves that under very general assumptions, any self-consistent method for assigning numerical values to degrees of belief about analytic statements of the form "A is true," can be mapped to standard probability theory.[3] Here, A can be a proposition like "A = The mass of the electron is X," or "A = G is a human," or any proposition which can be assigned a definite binary truth value.

Second, the notion of inference employed in Bayesian theory does not rely on the existence of some human agent or observer who can formulate inferences. Rather, since Bayesian inference is a clearly defined logical procedure, it might be instantiated in all sorts of physical processes, from the cognitive processes of animals to the bit operations of computers. The dynamics of such processes carry out the inference as a type of computation or sequence of logical operations; no human intervention need be invoked. To emphasize this very point, Jaynes even introduces a hypothetical "inference robot" in the first chapter of his book, which demonstrates how each logical step in Bayesian probability theory can be precisely formulated and then automated.

Given this approach to probabilities, we can now explain how Bayesian probabilities are assigned. The procedure of assigning probabilities is conventionally divided into two steps: The assignment of initial probabilities, called "prior probabilities", and the updating of these probabilities based on new information or evidence. We will first discuss the Bayesian updating procedure, and then conclude with a short explanation of prior probabilities.

To introduce the Bayesian updating procedure, we first must define some basic notation. Consider two claims A and B, which are propositions about the world. Specifically, these claims must be analytic statements, i.e. they obey the Law of the Excluded Middle. They are either true or false. Typically, these will be claims of as-of-yet undetermined truth value, such as statements about the future. Now suppose we would like to assign a probability to the claim B, *conditioned* on the fact that A is given to be true. We denote this probability as

$$P(B|A). \tag{61}$$

All probabilities we will consider will take this form. Specifically, in the Bayesian formalism it is meaning-

less to define probabilities like $P(A)$, that is, the probability that A is true given no background information. Any statement A about the world with any content always presupposes certain conditions, at least enough to make the meaning of the statement intelligible. The notation $P(B|A)$ makes this conditioning explicit.

The one rule of probability we will need to derive the Bayesian updating method is the rule governing joint probabilities, or probabilities associated with more than one proposition being true simultaneously. Suppose we wish to assign a probability to the claim that two statements A and B are *both* true, given a statement C. If we denote this joint claim as (A, B), then this probability is expressed simply as $P(A, B|C)$. The rule for assigning joint probabilities like $P(A, B|C)$ is as follows. First, we compute the probability of one of the claims, for example A, given that C is true. Then, we compute the probability that the second claim B is true, assuming that C *and* the first claim A are true. The joint probability is then given by the product of these two probabilities,

$$P(A, B|C) = P(B|A, C)P(A|C). \tag{62}$$

Note that this probability could equally be expressed as $P(A, B|C) = P(A|B, C)P(B|C)$, since the joint statement (A, B) is symmetrical: (A, B) is the same as (B, A). By equating these two expressions for $P(A, B|C)$ and solving for $P(A|B, C)$, we arrive at Bayes' Theorem, expressed as

$$P(A|B, C) = \frac{P(B|A, C)}{P(B|C)}P(A|C). \tag{63}$$

The method of Bayesian updating can now be expressed simply in these terms. Suppose that, given some background information encoded in a set of statements I, we have formulated some inference H, which we will call the hypothesis. We wish to associate a probability with this inference, given that the information I and some additional claim E are true. We call this statement E the evidence, and interpret it as some new information which has been acquired. Given these definitions and Bayes' Theorem in (63), the probability of interest $P(H|E, I)$ can be expressed as

$$P(H|E, I) = \frac{P(E|H, I)}{P(E|I)}P(H|I). \tag{64}$$

Now, it becomes clear why we have referred to this method as an "updating" procedure. (64) tells us that, given some initial assignment of a probability to our hypothesis H, expressed as $P(H|I)$, we can then reassess this probability upon consideration of the new information provided by the evidence E. This updated

probability is given by $P(H|E,I)$. In Bayesian terminology, the various terms in this updating procedure are typically described as follows:

- $P(H|I)$: The prior probability. This represents the best assessment of the hypothesis, without any of the information expressed in the additional evidence E. Prior probabilities can be assigned in various ways, depending on the specific nature of the prior information I; a few of these methods will be touched on later.

- $P(H|E,I)$: The posterior probability. Obtaining this value is the goal of our updating procedure, as it corresponds to a new assessment of the hypothesis based on the additional evidence E.

- $P(E|H,I)$: The likelihood. This probability must be computed in order to evaluate the posterior, assuming some prior is already assigned. One way to interpret this factor is to consider the hypothesis H as corresponding to some model about the world. $P(E|H,I)$ then expresses the probability that the given evidence E is obtained, given that the model expressed by H is a valid one.

- $P(E|I)$: The normalization. This factor must also be assigned a value in order to compute the posterior. However, the normalization is usually not computed directly, but is instead fixed by the constraint that the probability of H and its logical negation must sum to 1.

Given the posterior probability in terms of these values, the only remaining obstacle to implementing a Bayesian updating procedure is to specify a prior probability. The particular method of assigning this prior will depend on the nature of the hypothesis and prior information at hand, and in general there is not a single universal method for assigning priors. However, with certain types of prior information, invoking certain criteria of logical consistency can greatly constrain the possible priors which may be assigned. A few of these potential constraints are described briefly below.

One criterion for assigning prior probabilities is the Principle of Indifference. This principle states that, given N possible hypotheses $H_i = H_1, H_2 ... H_N$, if the given prior information does not distinguish them in any way except for their labeling, then equal probabilities ought to be assigned to each hypothesis: $P(H_i|I) = 1/N$. This principle can be useful in cases when each possible outcome can be expressed in terms of some base set of possibilities, each of which can be assigned an equal probability. Another method is to invoke symmetry considerations. For example, if a hypothesis H concerns some physical system which has rotational symmetry, then any probability assignment associated with this system should assign identical

values to statements which are indistinguishable up to a rotation of the system. One other approach, pioneered by Jaynes, is the "Maximum Entropy" principle. This criterion identifies the prior distribution with the probability distribution that maximizes the Gibbs entropy defined in (1), with respect to any constraints on the probabilities given in the prior information. With no constraints on the probability distribution, and a finite number of possible outcomes, the Maximum Entropy principle assigns the same probabilities as the Principle of Indifference.

6 Appendix B: The Density Matrix Formalism

In this appendix, we develop the concepts and mathematics behind the density matrix formulation of non-relativistic quantum mechanics. The advantages of the density matrix formalism are twofold. First, it permits a natural treatment of quantum systems in a *statistical mixture* of states, that is, systems which are in an indefinite state not due to intrinsic quantum indeterminism, but because of uncertainty in the system's preparation. Second, the density matrix formalism allows for a useful description of individual subsystems of an entangled quantum system. The framework developed in this appendix is primarily drawn from the textbook *Entangled Systems: New Directions in Quantum Physics*, by Jürgen Audretsch.[1]

The appearance of randomness in the standard account of quantum mechanics is well known. In general, the value of a quantum observable cannot be predicted with certainty prior to measurement: Only probabilities can be assigned to the various possible outcomes. Despite this fundamentally stochastic character of quantum systems, in quantum theory a definite state is nevertheless assigned to an isolated quantum system at all times. This is the state vector $|\psi\rangle$, which encodes the probabilities of obtaining various possible measurement outcomes from the system. In quantum mechanics, knowledge of the state vector amounts to the most complete description possible of a quantum system, although it does not deterministically dictate the results of measurement outcomes.

However, a description of a quantum system may also require probabilities for another reason, independent of the indeterministic foundation of quantum physics. This can occur when the system under consideration has been prepared in one of a set of possible states, but it is not known which one. We refer to this ensemble of possible states as a *statistical mixture*, or simply a mixed state. In such situations, although the system might be described by a single state vector in principle, without further interaction with the system this exact state cannot be specified. However, under these circumstances, it may be useful to assign a probability to each possible state in the mixture, which expresses the likelihood that the system is actually in that particular state. This is where the density matrix formalism becomes useful: It offers a compact description of the possible quantum states in a statistical quantum, along with their corresponding probabilities. As we will demonstrate later, the density matrix description can also be employed in the analysis of entangled quantum systems, for which it is impossible to meaningfully assign a definite state vector to either entangled

subsystem.

To begin building the density matrix picture of quantum mechanics, we first consider the special case in which a quantum system *does* in fact have a definite state vector. In contrast to mixed states, in this case we say that the quantum system is in the pure state $|\psi\rangle$. We will assume that this vector is normalized. We define the density operator associated with such a pure state as the outer product of the state vector $|\psi\rangle$ and its adjoint, $\langle\psi|$,

$$\hat{\rho} \equiv |\psi\rangle\langle\psi|. \tag{65}$$

This operator encodes the same probabilities concerning measurement outcomes as the state vector, but in a way which generalizes to the description of statistical mixtures in a straightforward manner. To see this, suppose there exists some quantum observable associated with the system under consideration, with a corresponding operator $\hat{Q}$. Each possible outcome of a measurement of this observable will be an eigenvalue of $\hat{Q}$. We denote the i^{th} potential measurement outcome as q_i, where the index i ranges over some set of integer values. In terms of the state vector, the Born rule states that the probability $P(q_i)$ of obtaining a given outcome upon measurement is given by $P(q_i) = \langle\psi|\hat{P}_i|\psi\rangle$, where $\hat{P}_i$ is the projection operator onto the space of eigenvectors of $\hat{Q}$ with eigenvalue q_i. We can express this probability in terms of the density operator by taking the operator trace, denoted with the symbol Tr, of $P(i)$. Given any orthonormal basis $\{|j\rangle\} = \{|1\rangle, |2\rangle ...\}$ for the Hilbert space of the system under consideration, where the index j ranges over a set of integer values, the trace operation on a operator $\hat{O}$ takes the form

$$\mathrm{Tr}\big[\hat{O}\big] = \sum_j \langle j|\hat{O}|j\rangle. \tag{66}$$

This operation leaves $P(q_i)$ unchanged, since the trace of a single number is just that number itself, and so $P(q_i) = \mathrm{Tr}\Big[\langle\psi|\hat{P}_i|\psi\rangle\Big]$. Now, by noting that the trace of a product such as $\langle\psi|\hat{P}_i|\psi\rangle$ is invariant upon cyclic permutations of the factors, we can write $P(q_i)$ as

$$P(q_i) = \mathrm{Tr}\Big[\langle\psi|\hat{P}_i|\psi\rangle\Big] = \mathrm{Tr}\Big[\hat{P}_i|\psi\rangle\langle\psi|\Big] = \mathrm{Tr}\big[\hat{P}_i\hat{\rho}\big]. \tag{67}$$

As a result of such a measurement, the state vector will collapse to an eigenvector of $\hat{Q}$ corresponding to the outcome of the measurement, q_i. Specifically, the state vector after measurement is obtained by applying the projection operator $\hat{P}_i$ to the pre-measurement state vector, and then dividing by the factor

$\sqrt{P(q_i)} = \sqrt{\langle\psi|\hat{P}_i|\psi\rangle}$ for normalization. Symbolically, the state vector undergoes the evolution $|\psi\rangle \rightarrow |\psi'\rangle = \hat{P}_i|\psi\rangle/\sqrt{\langle\psi|\hat{P}_i|\psi\rangle}$. The corresponding evolution for the density operator, from $\hat{\rho}$ to $\hat{\rho}'$, is then given by

$$\hat{\rho} = |\psi\rangle\langle\psi| \rightarrow \hat{\rho}' = |\psi'\rangle\langle\psi'| = \frac{\hat{P}_i|\psi\rangle}{\sqrt{P(q_i)}}\frac{\langle\psi|\hat{P}_i^\dagger}{\sqrt{P(q_i)}} = \frac{1}{\mathrm{Tr}\big[\hat{P}_i\hat{\rho}\big]}\hat{P}_i\hat{\rho}\hat{P}_i, \tag{68}$$

where $\hat{P}_i^\dagger$ denotes the Hermitian conjugate of $\hat{P}_i$. We have used the fact that $\hat{P}_i$ is Hermitian, or that $\hat{P}_i = \hat{P}_i^\dagger$. We can also consider how the density operator will evolve in isolation, in the absence of measurement. In this case, the state vector evolves from $|\psi\rangle$ to $|\psi'\rangle = \hat{U}|\psi\rangle$, where $\hat{U}$ is some unitary operator determined by the Hamiltonian of the system. Under this evolution, the adjoint of $|\psi\rangle$ evolves into $\langle\psi|\hat{U}^\dagger$, and so we see that corresponding evolution of the density operator takes the form

$$\hat{\rho} = |\psi\rangle\langle\psi| \rightarrow \hat{\rho}' = |\psi'\rangle\langle\psi'| = \hat{U}|\psi\rangle\langle\psi|\hat{U}^\dagger = \hat{U}\hat{\rho}\hat{U}^\dagger. \tag{69}$$

Taken together, equations (67)-(69) describe the dynamics of a pure state, and the probabilities which it encodes, in terms of operations on the density operator. To understand how these properties generalize to the case of statistical mixtures, it is useful to first consider the measurement probabilities associated with a mixed state. As described previously, mixed states occur when a quantum system is in one of some set of definite states, but it is not known which one. For some given mixed state, we denote this set of possible states as $\{|\psi^n\rangle\} = \{|\psi^1\rangle, |\psi^2\rangle, |\psi^3\rangle ...\}$, where the index n ranges over some set of integer values. To describe a mixed state, we assign a set of probabilities $\{p_n\} = \{p_1, p_2, p_3 ...\}$ over the set of possible states, such that p_n gives the probability that the system is actually in the state $|\psi_n\rangle$. The best way to assign these probabilities is contingent upon what is known about the preparation procedure which produced the quantum state. At this stage, we will just take this set of probabilities as a given.

Now, a straightforward application of the rules of probability can tell us the probabilities of obtaining the various values $\{q_i\}$ of the observable $\hat{Q}$. Suppose that upon measurement, the value q_i is observed. This result could have emerged in a number of ways. For example, the true state of the system prior to measurement might have been given by $|\psi^1\rangle$, and the measurement could have induced a collapse to a new state $\hat{P}_i|\psi^1\rangle/\sqrt{\langle\psi^1|\hat{P}_i|\psi^1\rangle}$ with probability $\langle\psi^1|\hat{P}_i|\psi^1\rangle$, as given by the Born rule and (67). Alternately, the system could have collapsed to a new state $\hat{P}_i|\psi^2\rangle/\sqrt{\langle\psi^2|\hat{P}_i|\psi^2\rangle}$ from $|\psi_2\rangle$, with probability $\langle\psi^2|\hat{P}_i|\psi^2\rangle$. In general, there will be as many different possible "paths" by which this measurement can occur as there are states in $\{|\psi^n\rangle\}$, the set of possible initial states. Moreover, each one of these possibilities corresponds to a

mutually exclusive event: The system is only ever truly in one quantum state, and only one of these possible collapses occurs. The probability of obtaining some result from a set of mutually exclusive outcomes is simply the sum of the individual probabilities of each event in the set. Therefore, if we define $P(|\psi^n\rangle \text{ and } q_i)$ as the probability that the system began in state $|\psi^n\rangle$ and then the value q_i is obtained, then the total probability $P(q_i)$ of obtaining q_i is the sum of these probabilities,

$$P(q_i) = \sum_n P(|\psi^n\rangle \text{ and } q_i). \tag{70}$$

Since each term in this sum is a joint probability of two separate events, we can express each term as a product of probabilities, in accordance with (62) in Appendix A. Specifically, for each term, the joint probability $P(|\psi^n\rangle \text{ and } q_i)$ is equal to the probability $P(|\psi^n\rangle)$ that the system begins in the state $|\psi^n\rangle$, multiplied by the *conditional* probability $P(q_i||\psi^n\rangle)$ that the value q_i is obtained, *given* the initial state $|\psi^n\rangle$. That is, we have

$$P(|\psi^n\rangle \text{ and } q_i) = P(|\psi^n\rangle)P(q_i||\psi^n\rangle). \tag{71}$$

However, we have already seen the factors in this product. The first factor $P(|\psi^n\rangle)$ is, by definition, equal to the probability p_n, since we defined each member p_n of the set $\{p_n\}$ as the probability that the system was prepared in the initial state $|\psi^n\rangle$. The second factor is given by the Born rule and (67): If the system begins in the state $|\psi^n\rangle$, then a value q_i will be obtained upon measurement with probability $\langle\psi^n|\hat{P}_i|\psi^n\rangle$. So, after substituting these results into (71), and then rewriting the sum in (70) in these terms, we find the result

$$P(q_i) = \sum_n P(|\psi^n\rangle)P(q_i||\psi^n\rangle) = \sum_n p_n \langle\psi^n|\hat{P}_i|\psi^n\rangle. \tag{72}$$

That is, the total probability of obtaining the result q_i in a measurement of $\hat{Q}$ is the sum of the Born probabilities of observing this result for each possible initial state $|\psi^n\rangle$, weighted by the probabilities p_n. Given this expression for $P(q_i)$, to understand how the density operator might be defined for the case of mixed states, we can ask: How should our original definition of $\hat{\rho}$ in (65) be generalized to statistical mixtures, so that the probabilities $P(q_i)$ take the same form $\mathrm{Tr}\big[\hat{P}_i\hat{\rho}\big]$ as in the pure state case? We can answer this question by manipulating our expression for $P(q_i)$ in (72) as follows:

$$P(q_i) = \sum_n p_n \mathrm{Tr}\Big[\langle\psi^n|\hat{P}_i|\psi^n\rangle\Big] = \sum_n p_n \mathrm{Tr}\Big[\hat{P}_i|\psi^n\rangle\langle\psi^n|\Big] = \mathrm{Tr}\left[\hat{P}_i\left(\sum_n p_n|\psi^n\rangle\langle\psi^n|\right)\right]. \tag{73}$$

So, we see that if we define that density operator as $\hat{\rho} = \sum_n p_n|\psi^n\rangle\langle\psi^n|$ for mixed states, then the probabilities $P(q_i)$ are still given as $\mathrm{Tr}\big[\hat{P}_i\hat{\rho}\big]$, just as for pure states. Therefore, we take this as the definition of the density operator in the general case, for a statistical mixture with possible states $\{|\psi^n\rangle\}$ and associated probabilities $\{p_n\}$:

$$\hat{\rho} \equiv \sum_n p_n|\psi^n\rangle\langle\psi^n| = \sum_n p_n\hat{\rho}_n. \tag{74}$$

Here, we have introduced the notation $\hat{\rho}_n \equiv |\psi^n\rangle\langle\psi^n|$ to denote the density matrix associated with the pure state $|\psi^n\rangle$. The density operator for a mixed state is then simply a weighted sum of the density matrices for the pure states in $\{|\psi^n\rangle\}$. For a pure state, all of the probabilities in the set $\{p_n\}$ vanish except for one, and this general definition reduces to the original definition given in (65).

After defining the density operator in this way, we find that the evolution of mixed states also takes the same form as the pure state case, for both evolution due to measurement and unitary evolution when the system is isolated. We have already noted that the system will, in the course of a measurement with value q_i, evolve into some state of the form $\hat{P}_i|\psi^n\rangle/\sqrt{\langle\psi^n|\hat{P}_i|\psi^n\rangle}$, with the value of n dependent on the initial pre-measurement state:

$$|\psi\rangle \to |\psi'\rangle \stackrel{?}{=} \frac{P_i|\psi^1\rangle}{\sqrt{\langle\psi^1|P_i|\psi^1\rangle}},\ \frac{P_i|\psi^2\rangle}{\sqrt{\langle\psi^2|P_i|\psi^2\rangle}}\ ... \tag{75}$$

Therefore, the post-measurement density operator, in accordance with the definition given in (74), will be given by a weighted sum of the density operators for these possible states. In this case, the appropriate weight for each term $|\psi'\rangle\langle\psi'| = \hat{P}_i\hat{\rho}_n\hat{P}_i/\mathrm{Tr}\big[\hat{P}_i\hat{\rho}_n\big]$ is the probability that the system collapsed into this state, *given* that a measurement of q_i was obtained. But this is just the probability that the system started in $|\psi^n\rangle$, given the measurement result q_i, since there is only one state in the mixture that could evolve into $\hat{P}_i|\psi^n\rangle/\sqrt{\langle\psi^n|\hat{P}_i|\psi^n\rangle}$. If we denote this probability as $P(|\psi^n\rangle|q_i)$, then the post-measurement density operator $\hat{\rho}'$ is given by

$$\hat{\rho} \to \hat{\rho}' = \sum_n P(|\psi^n\rangle|q_i)\frac{\hat{P}_i\hat{\rho}_n\hat{P}_i}{\mathrm{Tr}\big[\hat{P}_i\hat{\rho}_n\big]} = \sum_n \frac{P(|\psi^n\rangle|q_i)}{P(q_i||\psi^n\rangle)}\hat{P}_i\hat{\rho}_n\hat{P}_i. \tag{76}$$

On the right side of this expression, we have noted that the expression $\mathrm{Tr}\big[\hat{P}_i\hat{\rho}_n\big]$ is just $P(q_i||\psi_n\rangle)$, the probability of the measurement outcome q_i, given the initial state $|\psi^n\rangle$. By invoking Bayes' Theorem, given in (63) in Appendix A, we can rewrite the ratio of probabilities in each term as follows:

$$P(|\psi^n\rangle|q_i) = \frac{P(q_i||\psi^n\rangle)}{P(q_i)}P(|\psi^n\rangle) \implies \frac{P(|\psi^n\rangle|q_i)}{P(q_i||\psi^n\rangle)} = \frac{P(|\psi^n\rangle)}{P(q_i)} = \frac{p_n}{\mathrm{Tr}\big[\hat{P}_i\hat{\rho}\big]}. \tag{77}$$

Substituting this into (76), we find that

$$\hat{\rho} \to \hat{\rho}' = \sum_n \frac{p_n}{\mathrm{Tr}\big[\hat{P}_i\hat{\rho}\big]}\hat{P}_i\hat{\rho}_n\hat{P}_i = \frac{1}{\mathrm{Tr}\big[\hat{P}_i\hat{\rho}\big]}\hat{P}_i\left(\sum_n p_n\hat{\rho}_n\right)\hat{P}_i = \frac{1}{\mathrm{Tr}\big[\hat{P}_i\hat{\rho}\big]}\hat{P}_i\hat{\rho}\hat{P}_i. \tag{78}$$

This is exactly the same form as the evolution of a pure state under measurement, as given in (68). Last, we can look how a density operator for a mixed state evolves when the system under consideration is isolated. Given a mixed state with a density operator $\hat{\rho} = \sum_n p_n|\psi^n\rangle\langle\psi^n| = \sum_n p_n\hat{\rho}_n$, there are a number of different ways that the system might evolve, depending on the true initial state $|\psi^n\rangle$ that the system occupies. Specifically, the state $|\psi^1\rangle$ will evolve to $\hat{U}|\psi^1\rangle$, $|\psi^2\rangle$ will evolve to $\hat{U}|\psi^2\rangle$, and so on. Therefore, in general, after the evolution the system will be in the state $\hat{U}|\psi^n\rangle$ with probability p_n. This means that the corresponding evolution of the density operator is given by

$$\hat{\rho} \to \hat{\rho}' = \sum_n p_n\hat{U}|\psi^n\rangle\langle\psi^n|\hat{U}^\dagger = \hat{U}\left(\sum_n p_n|\psi^n\rangle\langle\psi^n|\right)\hat{U}^\dagger = \hat{U}\hat{\rho}\hat{U}^\dagger. \tag{79}$$

So again, this evolution takes the same form as in the pure state case. We now see that the formulas (67)-(69) governing the probabilities and dynamics associated with the density operator are equally valid for both pure states and statistical mixtures, given that the density operator is defined as in (74). This is what makes the density operator formalism useful in the context of statistical mixtures: The density operator alone can fully encode both the set of possible states $\{|\psi^n\rangle\}$ in a statistical ensemble, and the corresponding set of probabilities $\{p_n\}$.

The other useful application of the density matrix formalism that we will discuss here occurs when a quantum system cannot be described by a single state vector $|\psi\rangle$, not because of a preparation which generates a statistical ensemble, but because the system of interest has become entangled with another system. In this case, although a state vector cannot be assigned to the state of the individual systems, we will find that a density operator can be associated with each system that captures all the relevant probabilities and dynamics for that system.

To develop the density matrix formalism for entangled systems, let us consider the general form of a density operator for a composite system. Consider two subsystems A and B, which form a composite system $A \times B$. Suppose we can find some orthonormal basis $\{|i\rangle_A\} = \{|1\rangle_A, |2\rangle_A, |3\rangle_A ...\}$ for the Hilbert space associated with A, and another basis $\{|j\rangle_B\} = \{|1\rangle_B, |2\rangle_B, |3\rangle_B ...\}$ for the Hilbert space of B. For both bases, the indices i and j range over some set of integers, possibly infinite. In this case, a basis for the Hilbert space of the composite system is given by $\{|i\rangle_A \otimes |j\rangle_B\}$, where the indices i and j range over all possible combinations of their values in the bases $\{|i\rangle_A\}$ and $\{|j\rangle_B\}$. Here, the expression $|i\rangle_A \otimes |j\rangle_B$ denotes the tensor product of $|i\rangle_A$ and $|j\rangle_B$, which we will write more concisely as $|i\rangle_A|j\rangle_B$ or $|i, j\rangle$.

Once we have chosen this basis, a general density operator $\hat{\rho}$ which describes the composite system $A \times B$ can be expanded as a sum of outer products of the vectors in this basis,

$$\hat{\rho} = \sum_{i,i',j,j'} \rho_{ij,i'j'} |i\rangle_A |j\rangle_B \langle i'|_A \langle j'|_B = \sum_{i,i',j,j'} \rho_{ij,i'j'} |i, j\rangle\langle i', j'|. \tag{80}$$

Here, the coefficients $\rho_{ij,i'j'}$ of this expansion are given simply as $\rho_{ij,i'j'} = \langle i, j|\hat{\rho}|i', j'\rangle$, since we have expanded $\hat{\rho}$ in terms of an orthonormal basis. These are the matrix elements of $\hat{\rho}$ in the basis $\{|i, j\rangle\}$.

However, there are many cases where we would like to focus our analysis on one subsystem or the other, without handling the full density matrix $\hat{\rho}$. For example, we may be considering measurements of the system A alone, or measurements or observables which are *local* to A. A local observable for A has an operator $\hat{Q}$ of the form

$$\hat{Q} = \hat{Q}_A \otimes \hat{\mathbb{1}}_B = \hat{Q}_A \otimes \left(\sum_j |j\rangle_B \langle j|_B \right), \tag{81}$$

where $\hat{Q}_A$ is the operator for this observable in the Hilbert space for the system A. To construct the corresponding operator $\hat{Q}$ for this observable in the complete Hilbert space for the composite system $A \times B$, we must take the tensor product of $\hat{Q}_A$ with the identity operator $\hat{\mathbb{1}}_B = \sum_j |j\rangle_B \langle j|_B$ for the Hilbert space of B. The form of local operators implies that they will have the same set of eigenvalues as the lower-dimensional operator from which they were constructed. That is, $\hat{Q}$ has the same eigenvalues as $\hat{Q}_A$, and these eigenvalues correspond to the set of possible measurement outcomes associated with these operators. In addition, the projection operator $\hat{P}_i$, which projects onto the space of eigenvectors of $\hat{Q}$ with the eigenvalue q_i, takes the simple form

$$\hat{P}_i = \hat{P}_i^A \otimes \hat{\mathbb{1}}_B = \hat{P}_i^A \otimes \left(\sum_j |j\rangle_B \langle j|_B \right). \tag{82}$$

Here, $\hat{P}_i^A$ is the projection operator for $\hat{Q}_A$ in the Hilbert space for the system A alone. So, we construct the projection operator $\hat{P}_i$ from $\hat{P}_i^A$ in the same way that we construct $\hat{Q}$ from $\hat{Q}_A$. Given this form for local operators and their associated projection operators, we can now examine how measurements of local observables are carried out, and compute the corresponding probabilities. This will lead us to a method for constructing a density operator associated with the subsystem A alone.

Like any observable, the probabilities associated with the possible measured values of $\hat{Q}$ are given in (67) as

$$P(q_i) = \mathrm{Tr}\big[\hat{P}_i \hat{\rho}\big] = \mathrm{Tr}\Big[(\hat{P}_i^A \otimes \hat{\mathbb{1}}_B)\hat{\rho}\Big]. \tag{83}$$

However, since the projection operator $\hat{P}_i = \hat{P}_i^A \otimes \hat{\mathbb{1}}_B$ in this expression factors due to the locality of the observable $\hat{Q}$, we can express this probability in a special form, in terms of a density operator for A alone. To do this, we first reexamine the trace operator, as defined earlier in (66). In the orthonormal basis $\{|i,j\rangle\}$ that we have selected for the composite system, the trace of some operator $\hat{O}$ is given by

$$\mathrm{Tr}[\hat{O}] = \sum_{i,j} \langle i,j|\hat{O}|i,j\rangle = \sum_i \langle i|_A \left(\sum_j \langle j|_B \hat{O}|j\rangle_B \right) |i\rangle_A. \tag{84}$$

On the right hand side of this expression, we see that the trace of an operator in the Hilbert space of the composite system $A \times B$ can be thought of as two separate operations: A sum over the basis states associated with the B subsystem, followed by a sum over the basis states for the A subsystem, or visa versa. These operations are defined as the partial traces over the degrees of freedom of B and A, respectively. That is, the partial trace $\mathrm{Tr}_A\big[\hat{O}\big]$ of $\hat{O}$ over the degrees of freedom of A is defined as

$$\mathrm{Tr}_A\big[\hat{O}\big] \equiv \sum_i \langle i|_A \hat{O}|i\rangle_A. \tag{85}$$

$\mathrm{Tr}_B[\hat{O}]$ is defined in an analogous way. We can now take the complete trace of $\hat{O}$ by applying the partial traces over the individual subsystems in either order, so that $\mathrm{Tr}\big[\hat{O}\big] = \mathrm{Tr}_A\Big[\mathrm{Tr}_B\big[\hat{O}\big]\Big] = \mathrm{Tr}_B\Big[\mathrm{Tr}_A\big[\hat{O}\big]\Big]$.

We can use this decomposition of the trace operator to rewrite our expression for $P(q_i)$ in (83) as follows,

$$P(q_i) = \mathrm{Tr}_A\Big[\mathrm{Tr}_B\big[(\hat{P}_i^A \otimes \hat{\mathbb{1}}_B)\hat{\rho}\big]\Big] = \mathrm{Tr}_A\left[\sum_j \langle j|_B (\hat{P}_i^A \otimes \hat{\mathbb{1}}_B)\hat{\rho}|j\rangle_B\right]. \tag{86}$$

Now, note that since $\hat{P}_i^A$ is an operator on the Hilbert space associated with A alone, it is unaffected by the products with the basis vectors $|j\rangle_B$ of the Hilbert space of B, and therefore can be taken out of the sum over j. This allows for the simplification

$$P(q_i) = \mathrm{Tr}_A \left[\hat{P}_i^A \left(\sum_j \langle j|_B \hat{\mathbb{1}}_B \hat{\rho} |j\rangle_B \right) \right] = \mathrm{Tr}_A \left[\hat{P}_i^A \left(\sum_j \langle j|_B \hat{\rho} |j\rangle_B \right) \right] = \mathrm{Tr}_A \Big[\hat{P}_i^A \mathrm{Tr}_B \big[\hat{\rho} \big] \Big]. \tag{87}$$

In this manipulation, the tensor product symbol $\otimes$ is dropped after the projection operator $\hat{P}_i^A$ is taken outside the sum, since the sum over j is not an operator on the Hilbert space of B, but simply a number. In addition, we drop the identity operator $\hat{\mathbb{1}}_B$, since by the definition of the identity operator we have $\langle j|_B \hat{\mathbb{1}}_B = \langle j|_B$. Given this expression for $P(q_i)$, we can now usefully define the *reduced* density operator $\hat{\rho}^A$ associated with the subsystem A as

$$\hat{\rho}^A \equiv \mathrm{Tr}_B \big[\hat{\rho} \big]. \tag{88}$$

After substituting this definition into (87), the probabilities $P(q_i)$ take the form

$$P(q_i) = \mathrm{Tr}_A \Big[\hat{P}_i^A \hat{\rho}^A \Big]. \tag{89}$$

So, for observables which are local to the subsystem A, the reduced density matrix $\hat{\rho}^A$ encodes the probabilities associated with measurement outcomes in exactly the same way as density matrices for a lone quantum system. That is, for a given $\hat{\rho}^A$, to compute the probability of a measurement result q_i, we apply the appropriate projection operator P_i^A, and then take the trace of the result. Only now, after tracing over the degrees of freedom of B to obtain $\hat{\rho}^A$, this computation is done entirely within the Hilbert space associated with A. In addition, during a local measurement on the subsystem A, the reduced density operator $\hat{\rho}^A$ evolves in the same way as a density operator of a full system. We can see this by tracing over the degrees of freedom of B in equation (68), in the particular case of a local observable where $\hat{P}_i = \hat{P}_i^A \otimes \hat{\mathbb{1}}_B$:

$$\mathrm{Tr}_B \big[\hat{\rho} \big] = \hat{\rho}^A \to \hat{\rho}^{A\prime} = \mathrm{Tr}_B \left[\frac{1}{\mathrm{Tr}\big[\hat{P}_i \hat{\rho} \big]} \hat{P}_i \hat{\rho} \hat{P}_i \right] = \mathrm{Tr}_B \left[\frac{1}{\mathrm{Tr}_A \big[\hat{P}_i^A \hat{\rho}^A \big]} (\hat{P}_i^A \otimes \hat{\mathbb{1}}_B) \hat{\rho} (\hat{P}_i^A \otimes \hat{\mathbb{1}}_B) \right], \tag{90}$$

$$\hat{\rho}^A \to \hat{\rho}^{A\prime} = \frac{1}{\mathrm{Tr}_A \big[\hat{P}_i^A \hat{\rho}^A \big]} \hat{P}_i^A \mathrm{Tr}_B \big[\hat{\rho} \big] \hat{P}_i^A = \frac{1}{\mathrm{Tr}_A \big[\hat{P}_i^A \hat{\rho}^A \big]} \hat{P}_i^A \hat{\rho}^A \hat{P}_i^A. \tag{91}$$

Again, this is just as in the case of an single isolated system: To obtain the post-measurement density operator, we apply the projection operator $\hat{P}_i^A$ to either side of the initial density operator $\hat{\rho}^A$, and then normalize by dividing by the factor $P(q_i) = \mathrm{Tr}_A\left[\hat{P}_i^A \hat{\rho}^A\right]$.

Of course, for non-local measurements, and for the evolution of A in the absence of any interaction external to $A \times B$, the evolution of $\hat{\rho}^A$ will be more complicated. This is because these processes depend on the state of the subsystem B and its interactions with A in a nontrivial way. However, working with $\hat{\rho}^A$ can still be a useful tool for understanding the information encoded in the subsystem A alone, in terms of the probabilities $P(q_i)$ associated with making various local measurements on A. In particular, the reduced density operator allows us to understand how interaction with B produces an effective statistical mixture of the possible states of A. This is because although the subsystems A and B might begin in individual pure states $\hat{\rho}^A$ and $\hat{\rho}^B$, in which case the density operator factorizes as $\hat{\rho} = \hat{\rho}^A \otimes \hat{\rho}^B$, a nontrivial interaction between A and B will generally evolve the system so that the new reduced density operators $\hat{\rho}^{A\prime}$ and $\hat{\rho}^{B\prime}$ each correspond to mixed states. This is the content of entanglement: After interaction, neither reduced density operator can be associated with a single state vector, but only an ensemble of possibilities.

References

[1] J. Audretsch. *Entangled Systems: New Directions in Quantum Physics.* Wiley-VCH Verlag, 1st edition, 2007.

[2] L. Boltzmann. Further Studies on the Thermal Equilibrium of Gas Molecules. *Sitzungsberichte Akad. Wiss.*, 66:275370, 1872.

[3] R. T. Cox. *The Algebra of Probable Inference.* John Hopkins Press, 1st edition, 1961.

[4] E. T. Jaynes. Information Theory and Statistical Mechanics. II. *Phys. Rev.*, 108(2):171–190, 1957.

[5] E. T. Jaynes. Gibbs vs Boltzmann Entropies. *Am. J. Phys.*, 33(5):391–398, 1965.

[6] E. T. Jaynes. *Probability Theory: The Logic of Science.* Cambridge University Press, 1st edition, 2003.

[7] R. Landauer. Irreversibility and Heat Generation in the Computing Process. *IBM J. Res. Develop.*, 5(3):183–191, 1961.

[8] L. Maccone. Quantum Solution to the Arrow-of-Time Dilemma. *Phys. Rev. Lett.*, 103.8(080401), 2009.

[9] L. Mlodinow and T. A. Brun. Relation between the psychological and thermodynamic arrows of time. *Phys. Rev. E*, 89(052102), 2014.

[10] R. Smith. Do Brains Have an Arrow of Time? *Philos. Sci.*, 81(2):265–275.

[11] J. von Neumann. *Mathematical Foundations of Quantum Mechanics*. Princeton University Press, 1955.