October 2019

FUNCTION AND DISSIPATION IN FINITE STATE AUTOMATA - FROM COMPUTING TO INTELLIGENCE AND BACK

Natesh Ganesh

FUNCTION AND DISSIPATION IN FINITE STATE AUTOMATA - FROM COMPUTING TO INTELLIGENCE AND BACK

A Dissertation Presented

by

NATESH GANESH

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

September 2019

Electrical and Computer Engineering

FUNCTION AND DISSIPATION IN FINITE STATE AUTOMATA - FROM COMPUTING TO INTELLIGENCE AND BACK

A Dissertation Presented

by

NATESH GANESH

Approved as to style and content by:

Neal G Anderson, Chair

Marco Duarte, Member

Weibo Gong, Member

David Moorman, Member

Robert Jackson, Department Chair
Electrical and Computer Engineering

To all of my grandparents who would have loved to read this.

ACKNOWLEDGMENTS

This has been a long, arduous and joyful journey (do I know of any other way?), but here I am at the precipice of completing this stage of my PhD career looking forward to whatever adventure is up next. At this moment, I am reflecting back on all those who have helped complete this dream of mine. I would like to start by thanking my parents who have been extremely supportive of my goals and ambitions, believing in my work at times when I did not myself. I would next thank my advisor Prof. Anderson who has been both a great mentor and a good friend, giving me a good mix of the intellectual freedom that I crave and the scientific rigor that I need. I want to take this moment to also thank my committee for agreeing to help me complete this journey. Last but not the least, I want to thank my partner Ashley who has been my rock over the last few years, helping me (kicking my butt) over the finish line. I am forever grateful for all your help, patience, belief and sacrifice. There have been so many friends over the years - ones whose names I remember and those that I do not, who have played a crucial role in shaping me and my thinking over the course of my PhD - to all of you, a big thank you. It takes a village to raise a child and help me get my PhD. So for that, I am indebted to all of you. Thank you everyone!!

ABSTRACT

FUNCTION AND DISSIPATION IN FINITE STATE AUTOMATA - FROM COMPUTING TO INTELLIGENCE AND BACK

SEPTEMBER 2019

NATESH GANESH

B.Tech., NATIONAL INSTITUTE OF TECHNOLOGY, TRICHY, INDIA

M.Sc., UNIVERSITY OF MASSACHUSETTS, AMHERST

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Neal G Anderson

Society has benefited greatly by the technological revolution and the tremendous growth in computing powered by Moore's law. However, we are fast approaching the ultimate physical limits in terms of both device sizes and the associated energy dissipation. It is important to characterize these limits in a physically grounded and implementation-agnostic manner, in order to capture the fundamental energy dissipation costs associated with performing computing operations with classical information in nanoscale quantum systems. It is also necessary to identify and understand the effect of quantum indistinguishability, noise, and device variability on these dissipation limits. Identifying these parameters is crucial to designing more energy efficient computing systems moving forward. In this dissertation, we will provide a physical description a finite state automata, an abstract tool commonly used to describe computational operations under the Referential Approach to physical information theory.

We will derive the fundamental limits of dissipation associated with a state transition in deterministic and probabilistic finite state automata, and propose efficacy measures to capture how well a particular state transition has been physically realized. We will use these dissipation bounds to understand the limits of dissipation during learning during training and testing phases in feedforward and recurrent neural networks. This study of dissipation in neural network provides key hints at how dissipation is fundamentally intertwined with learning in physical systems. These ideas connecting energy dissipation, entropy and physical information provide the perfect toolkit to critically analyze the very foundations of computing, and our computational approaches to artificial intelligence. In the second part of this dissertation, we derive the non-equilibrium *reliable low dissipation* condition for predictive inference in self-organized systems. This brings together the central ideas of homeostasis, prediction and energy efficiency under a single non-equilibrium constraint. The work was further extended to study the relationship between adaptive learning and the reliable high dissipation conditions, and the exploitation-exploration trade-offs in active agents. Using these results, we will discuss the differences between observer dependent and independent computing, and propose an alternative novel descriptive framework of *intelligence* in physical systems using thermodynamics. This framework is called *thermodynamic intelligence* and will be used to guide the engineering methodologies (devices and architectures) required to implement these descriptions.

TABLE OF CONTENTS

LIST OF FIGURES

Figure **Page**

CHAPTER 1

INTRODUCTION

The tremendous success of the computing industry over the last seven decades has been primarily driven by Moore's law [1]. This self-fulfilling prophecy predicted the doubling of computing power approximately every 18 months through the shrinking of transistor devices, and enabled the first technology revolution. However, as we fast approach the fundamental physical limits to scaling, things have begun to slow down with thermal noise, quantum effects and device variability making further shrinking unfeasible - *death by size* [150]. Since 2004, Dennard scaling [5] has slowed down and clock frequencies have started to flatten out, as the industry have embraced the use of multiple cores that can operate in a parallel fashion. However, continued addition of extra cores is not viable and does not guarantee improved speed in computing - *death by parallelism* [3], [4]. Last, but not the least is the problem of energy consumption and dissipation, which will be the main focus of this thesis. As we approach the fundamental limits to dissipation, the challenge of controlling the heat dissipated is going to affect the density of devices that are packed in our chips, as well as how fast they are run - *death by heat* [6]. It is estimated that the amount of energy required for computing systems account for about 5% of the total energy expenditure of the United States, and this number is expected to grow. It is predicted that at this current rate, computing will be unsustainable by 2040, since it will require more energy than the world can produce - *death by starvation*. There are important economic factors to consider with approximately 50% of the lifetime budget of a modern high-performance computing center being used to pay the energy bill. In order to sustain the growing

demands for computing and sustainably usher in the next technological revolution, we need to work to overcome these major barriers.

There are two possible paths for computing moving forward - *Evolutionary* and *Revolutionary* [150]. The evolutionary path deals with continuing scaling of CMOS devices as well as new FET devices, changes to architecture with better pipelines, 3D design, improved parallelism through multicore systems and using application specific integrated chips (ASICs). While these approaches should continue to improve our computing systems, the return on investment is expected to plateau quickly. It is hard to imagine a second technological revolution built on the back of the *More Moore* evolutionary approach. The revolutionary path (i.e. *More than Moore*) on the other hand involves the study of novel devices like carbon nanotube transistors, memristors, spintronics, photonics, quantum and molecular devices, as well as non von-Neumann acrchitectures - crossbar, neuromorphic, computation-in-memory. There is a need for significant work in these new technologies across the stack - to improve device reliability and signal to noise ratio (SNR), identify optimal architectures and computing paradigms under which these new devices can be utilized well, and fabrication techniques to minimize defect rates when scaling up production. It is also important to identify the problem/application space in which these more than Moore strategies will provide clear benefits over existing technology. Many of these problems have hindered their progress, with systems based on CMOS devices and the von Neumann architecture continue to be the industry standard for performing large number of mathematical and logical operations efficiently.

In addition to the fast approaching end to Moore's law is the shift in the focus of the industry away from performing traditional mathematical operations, and towards building intelligent systems that learn from large amounts of data. These two factors together provide for an unique opportunity to explore these novel unconventional devices, architectures and theoretical frameworks for new tasks at hand. Artificial

intelligence (AI) achieved through powerful machine learning techniques have led to remarkable success in the fields of computer vision, healthcare, natural language translation, recommendation systems and the potential to power future technologies like self-driving cars and lifelong learning robots. However the current state of the systems are severely limited by the amount of data and compute available to learn the necessary task. The demand for compute power by machine learning systems have been doubling every 3.5 months, at a super-Moore rate [8]. Even with availability of large amounts of cheap compute, this is not sustainable and it is paramount that we understand how to build AI systems in an energy efficient manner.

In order to achieve the goals of this dissertation, the following list of objectives need to be met -

(a) Devise a technology-agnostic description of a general finite-state automaton (FSA) instantiated in a physical system, and utilize it to obtain the fundamental lower bound on the average energy dissipated per FSA state transition operated in steady state. Apply these bounds to simple finite-state automata to understand the relationship between the energy dissipation, inputs, state transitions and use them to define irreversibility in FSAs.

(b) Characterize feed-forward and recurrent neural networks as finite-state automata and use the bounds on dissipation from above to understand the ultimate limits to learning and inference in these networks, as well as the effect of different network parameters on the dissipation.

(c) Without assuming learning apriori, describe the non-equilibrium thermodynamic conditions under which physical systems realize adaptive learning and predictive inference.

(d) Propose a new physical framework of thermodynamic intelligence, and describe the changes in devices, architectures, programmability and design methodologies that is necessary to realize it.

The thesis is organized as follows -

Chapter 2: In this first of two review chapters, the basics of classical information theory is introduced. Important ideas in classical information theory like entropy, conditional entropy and relative entropy are discussed. These concepts are then used to build into concepts like rate-distortion theory and computational channels that are necessary later in this dissertation. The chapter ends with a discussion on Landauer's Principle which introduced the vital physical consequences of information processing,

Chapter 3: In this second review chapter, we will continue to build upon the important concept of classical information as a physical quantity and focus on quantum systems specifically. The density matrix formalism used to characterize a quantum systems is used to define the quantum mechanical equivalent of classical information theory concepts like entropy and mutual information. The referential approach to physical information theory - the framework that the rest of the dissertation will utilize is used to provide physical descriptions of abstract logical operations, as well recast Landauer's principle under this framework. Important computational efficacy measures for logical operations are discussed and their connection to the dissipation is elucidated.

Chapter 4: Finite state automata are a powerful computational model that is very commonly used. In this chapter, we discuss an abstract deterministic irreducible FSA driven by independent identically distributed (IID) inputs, and introduce the equivalent physical description under the referential approach. These are used to derive the fundamental lower bound on dissipation in steady state for a

deterministic irreducible FSA. These dissipation bounds were applied to simple illustrative cases, and studied to derive the condition for physical irreversibility in FSA. The approach was further extended to include more general FSA like deterministic (but not necessarily irrducible) and probabilistic FSA, as well FSA driven by time-correlated inputs. The chapter ends with an extension of the efficacy measures (from the previous chapter for logical transformation) to capture how well a FSA state transition has been physically instantiated.

Chapter 5: In this chapter, we introduce neural networks which form the basis of many modern techniques in AI. Feed-forward networks - their training and use in classification is discussed and characterized as a FSA, consisting of both the neural nodes and network weights. The results from the previous chapter are used to determine the fundamental dissipation limits in these networks for training and testing. The effect of input data probabilities and training parameters such as learning rate on the dissipation is also studied. The chapter continues with the study of dissipation limits in recurrent Hopfield and Boltzmann networks. The fundamental dissipation limit for the simulated annealing technique and a dissipation complexity measure for optimization problems are introduced.

Chapter 6: This will be first of two chapters in which we will build towards a physically grounded theory of thermodynamic intelligence. The chapter begins with an introduction to both equilibrium and non-equilibrium thermodynamics, as well as complexity and self-organization. This is followed by deriving the non-equilibrium reliable low dissipation condition under which predictive inference capabilities emerges in complex self-organized systems. The relationship between this condition and the *stability-plasticity problem* (which captures the trade-off between a system's prediciton capability for a finite amount of memory) is briefly explored. These conditions are further extended to active agents

that can act on their environment to study exploitation-exploration trade-offs in such systems. The chapter ends with discussion on incorporating information theoretic measures into recent published results on dissipation driven adaptation.

Chapter 7: In this chapter, we start by examining the fundamental philosophical underpinnings of current computational approaches to intelligence. The difference between observer dependent versus independent computation and intelligence is discussed, and while we desire the latter, we explore why our current approaches will only produce the first. The results from the previous chapter are visited as path forward to explain observer-independent intelligence. The chapter concludes with a discussion of this new framework of thermodynamic intelligence as a path towards building energy efficient AI systems, and the engineering challenges that needs to be overcome in order to reboot computing for intelligence [7].

Chapter 8: In this last chapter of the dissertation, the results of the work presented in the previous chapters are summarized. This is followed by a brief reflection of this unique time in the computing field that we find ourselves in, important ideas to think about moving forward, as well as future work.

This dissertation will be the story of my own intellectual journey, beginning at characterizing the ultimate dissipation bounds for finite state automata, extending those results to simple neural networks to understand the limits to learning and finally using these results to questions the very fundamentals of computing as we explore the non-equilibrium thermodynamic conditions under which predictive intelligence emerges in systems. Since the work will pull topics from different areas like information theory, quantum mechanics, non-equilibrium thermodynamics and finite-state automata, we will start by discussing the necessary ideas needed for this dissertation

over the next two chapters. The results in Chapters 4,5 and 6 on FSAs, neural networks and thermodynamic conditions for predictive inference have been rigorously derived, while Chapter 7 is more speculative as we discuss the foundational principles of computing and a new engineering paradigm to build energy efficient artificial intelligence systems.

CHAPTER 2

INTRODUCTION TO CLASSICAL INFORMATION THEORY

This will be the first of two chapters that will introduce technical content important for the research completed in this dissertation. We shall start by first introducing the idea of a communication channel and various components associated with it. Classical information theory was invented by Claude Shannon, in order to mathematically quantify the amount of information that is transmitted over a communication channel. The source coding and channel coding theorems based on information theory forms the basis of modern communication theory. We will explain a variety of important concepts under information theory including self-entropy, joint entropy and mutual information, and explore their properties. These sections of this chapter have been adapted from the lecture notes of ECE697PT Physical Information Theory [1] [21]. We will build on these ideas and introduce both Rate Distortion theory and the Information Bottleneck, the latter which will be utilized in later chapters. Following this, we will explore the idea of logical operations as computational channels, and how those can be described using information theoretic-measures. The chapter will conclude with a discussion of Landauer's principle that provides an important connection between the abstract and physical notions of information.

[1]The notes are not publicly available or peer-reviewed, but some of the important information theory concepts are also available in [10].

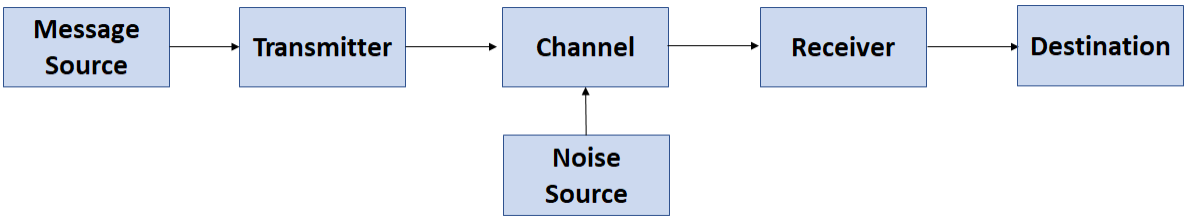

Figure 2.1. The communication channel framework used by Shannon in [11], and adapted from [21].

2.1 Communication Process & Channel

Classical information theory was originally formulated to be used in the field of digital communication. Shannon described the "fundamental problem of communication" in his seminal paper, *A Mathematical Theory of Communication* [11] as *"...reproducing at one point either exactly or approximately a message selected at another point."* This would require a mathematical description of this process and a "...represent(ation of) the various elements involved as mathematical entities, suitably idealized from their physical counterparts." The framework used by Shannon for analyzing this process is given by the Fig.(2.1) [11].

The various elements from the communication process are

- Information Source - The process of generating a string of symbols from a set of available symbols. This string of symbols is the message that needs to be transmitted. In this dissertation, we will be mainly focused on what are referred to as *discrete information sources* - a process generating messages from a finite alphabet of symbols. A popular type of discrete information source that is often used in the study of communication theory are discrete IID sources. A discrete IID source is a process that generates sequences $\mathbf{X} = X^{(0)}X^{(1)}X^{(2)}...$, of independently identically distributed (IID) discrete random variables $X^{(n)}$, drawn from a d-ary $\{x\} = \{x_1, x_2, x_3, ..., x_d\}$ given by a probability mass function $\{p\} = \{p(x_1), p(x_2), p(x_3), ..., p(x_d)\}$.

- *Transmitter* - A process mapping messages from the source into signals, or physical disturbances in the channel. This process also includes *encoding* - translating the strings of symbols from an IID information source into a more efficient form before transmitting it. There are advantages to coding strings (or blocks) of source symbols instead of individual symbols, and this is known as *block coding*. In "fixed-length" block coding, strings of source output are divided into length-N message blocks $x^N = x^{(0)}x^{(1)}...x^{(k)}...x^{(N-1)}$, and generates a length-M codeword sequence $\mathcal{C}(x^N) = c^{(0)}c^{(1)}c^{(2)}...c^{(M-1)}$ for each message block according to a codebook, which maps each of the d^N possible message blocks x^N into a corresponding codeword.

- *Receiver* - A process detecting disturbances in the channel, decoding and mapping them back into messages for delivery to their ultimate destination. The coded messages are decoded using a decoding rule that assigns a length-N sequence $\mathcal{D}(\mathcal{C}(x^N))$, constructed from the symbols in the source alphabet, to each M-digit codeword $\mathcal{C}(x^N)$. A code is *lossless* if there is a decoding scheme such that $\mathcal{D}(\mathcal{C}(x^N)) = x^N$ for all possible source messages x^N. This means that the decoding scheme can perfectly reconstruct every possible message from the source without ambiguity in a noiseless channel. On the other hand in *lossy* coding, $\mathcal{D}(\mathcal{C}(x^N)) \neq x^N$ for one or messages that is generated by the source and hence the decoder cannot successfully reconstruct every source message.

- *Channel* - The part of the physical world that can be used for propagating transmitted signals through space and time to the receiver. The most common type of channels that are studied in communication systems are *discrete memoryless channels*. The memoryless characteristic ensures that the transmission of each symbol is independent of which symbols were previously sent or received. Such a channel is characterized by an input variable X with finite alphabet

$\{x_i\}$, output Y with alphabet $\{y_j\}$ and the channel matrix - a set of conditional probabilities $\{q_{j|i}\} = Prob\{Y = y_j|X = x_i\}$, which is the probability that symbol y_j is received when the input symbol x_i is transmitted.

- *Noise Source* - A process that corrupts the message signal, and inhibits the receivers ability to distinguish the received signal from the transmitted one and to decode the messages. If this disturbance is produced by the channel, it is referred to as *channel noise*, but in general refers to any physical factor that limits the ability of the receiver to distinguish between different messages. Noise sources are often modeled as Gaussian sources.

In the next section, classical information theory will be introduced as an important tool to quantify the amount of information in the messages that will be sent through the communication channel. These tools are exceptionally powerful through their use in the noiseless and noisy coding theorems [10]. Our focus however will be in the use of these tools in computational channels.

2.2 Classical Information Theory

Information theory seeks to obtain the fundamental limits on the reliability of compressing and exchanging data. The theory, originally used in the communication field, has since developed and found applications in a wide variety of disciplines [10]. Application of this theory to nanoelectronic circuits is necessary as their intended purpose includes communication, computation and information storage. It can also be used to develop many important performance metrics which will help in the exploration of future devices. Information theory allows us to connect such performance measures directly with related thermodynamic quantities such as thermodynamic entropy and energy dissipation, and hence provide us with important knowledge on the capabilities of these nanoelectronic circuits.

2.2.1 Entropy

Entropy is a measure of the disorder of the system. It was originally a thermodynamic concept, but since been adapted to other fields of study, including information theory, complex systems and machine learning. It is a central concept that links these multiple fields and plays a crucial role in understanding physical conceptions of information. It can be expressed as

A state of high order $\rightarrow$ low probability

A state of low order $\rightarrow$ high probability

The information entropy introduced by Claude Shannon is often eponymously called *Shannon entropy* or *Shannon self-information*. Shannon entropy is a measure of the uncertainty associated with a random variable. For an event X with n outcomes, (x_i, $i = 1, 2, 3, ..., n$) the information entropy, denoted by $H(X)$ or $H(\{p(x_i)\})$, is defined as

$$H(X) = \sum_{i=1}^{n} p(x_i) \log_b p(x_i). \tag{2.1}$$

where $p(x_i)$ is the probability mass function of the outcome x_i, b is the base of the logarithm used. The unit of the information entropy H depends on the value of base b, and is expressed in bits when $b = 2$ and in nats when $b = e$. For our purposes we shall use $b = 2$. The above definition was evolved for Shannon entropy by imposing three reasonable conditions on the quantitative measure of the "information content of an event". These are

(i) Information is non-negative.

(ii) Least probable events provide the most information.

(iii) Information is additive for independent events.

This relationship between the probability of an event and the associated entropy can be understood using the following example. Let us say a coin with known prob-

abilities of coming up heads or tails is tossed. The entropy of the unknown result of a toss of the coin is maximized, in the situation of maximum uncertainty with heads and tails equally probable. The amount of information associated with each toss can be quantified to be one bit of information. However if we the coin is not fair, and one side is more likely to come up than the other, this reduced uncertainty is reflected in a lower entropy, and each toss of the coin delivers less than one bit of information. A double-headed coin which never comes up tails is an extreme case in which there is no uncertainty, the entropy is zero and each toss of the coin delivers zero bits of information.

The use of the logarithm function allows the entropy function as defined above, to follow the first and third condition (assuming entropy vanishes for $p(x_i) = 0$). If there is a set of n mutually exclusive events ($a_j,\ j = 1, 2, ..., n$) each with equal probability $p(a_j) = \frac{1}{n}$, the Shannon entropy of this set of events is equal to $\log_b n$ units. Furthermore, consider a set of m mutually exclusive events which are independent from the previous set of events, with the probability of each event given as $\frac{1}{m}$. The Shannon entropy associated with this set would then be $\log_b m$ units. If both sets are considered together, i.e. for the set of mn possible events each with a probability of $\frac{1}{mn}$, the Shannon entropy is $\log_b(mn) = \log_b m + \log_b n$ units which is the sum of the Shannon entropies of the two independent sets of events.

2.2.2 Joint Entropy

Using the definition of the joint probability of the event $X = x_i$ and $Y = y_j$ as

$$p_{ij} = Prob\{X = x_i, Y = y_j\}$$

the joint Shannon entropy of the probability mass function $\{p_{ij}\}$ is given as

$$H(X, Y) = -\sum_i \sum_j p_{ij} log_2 p_{ij}$$

The joint entropy is symmetric with $H(X, Y) = H(Y, X)$ and is bounded as

$$H^* \leq H(X,Y) \leq H(X) + H(Y)$$

where H^* is the larger of the self-entropies $H(X)$ and $H(Y)$. Equality is achieved in the lower bound when X can be completely inferred from Y or vice versa, and equality in the upper bound is achieved when X and Y are independent random variables

2.2.3 Conditional Entropy

For an input $X = x_i$ that is transmitted through a channel to produce output Y, distributed with conditional probabilities $q_{j|i}$, the conditional entropy of Y for the fixed input $X = x_i$ is given as

$$H(Y|x_i) = -\sum_j q_{j|i} log_2 q_{j|i}$$

The conditional entropy is the expectation of this quantity over all inputs, and is the average entropy of Y given that X is known.

$$\begin{aligned} H(Y|X) &= \sum_i p_i H(Y|x_i) \\ &= -\sum_i \sum_j p_i q_{j|i} \log_2 q_{j|i} \end{aligned}$$

Unlike joint entropy, the conditional entropy is not symmetric and is bounded as

$$0 \leq H(Y|X) \leq H(Y)$$

where equality is achieved in the lower bound when each x_i uniquely maps onto a single output y_j, with $q_{j|i} = 1$ for only one j for every i and zero for other j's. Equality is achieved in the upper bound when X and Y are independent of each other. The conditional entropy is related to the self and joint entropies as below

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$$

2.2.4 Relative Entropy and Shannon Mutual Information

The relative entropy or *Kullback-Liebler divergence* is a distance measure used to determine the similarities between the two distributions $\{p_k\}$ and $\{q_k\}$, and defined as

$$H(P||Q) = H(\{p_k\}||\{q_k\}) = \sum_k p_k[\log_2 p_k - \log_2 q_k]$$

The relative entropy is bounded as $0 \leq H(P||Q) \leq \infty$, with equality achieved in the lower bound if and only if $\{p_k\} = \{q_k\}$.

The Shannon mutual information between two variables X and Y is used to measure the dependence of one variable on another i.e., the amount of correlation between X and Y. For two discrete random variables X and Y with joint pmf $\{p_{ij}\}$ and marginal pmfs $\{p_i\}$, the mutual information is defined as

$$\begin{aligned} I(Y;X) &= H(Y) + H(X) - H(Y,X) \\ &= H(\{p_{ij}\}||\{p_i q_j\}) \end{aligned} \tag{2.2}$$

where $H(\{p_{ij}\}||\{p_i q_j\})$ is the relative entropy between the $\{p_{ij}\}$ and $\{p_i q_j\}$ distributions.

$$H(\{p_{ij}\}||\{p_i q_j\}) = \sum_i \sum_j p_{ij} \left[\log_2(p_{ij}) - \log_2(p_i q_j)\right]$$

Using the relationship between self, joint and conditional entropies from before

$$H(Y,X) = H(X) + H(Y|X) = H(Y) + H(X|Y) = H(X,Y)$$

the mutual information $I(X;Y)$ can be rewritten as

$$\begin{aligned} I(X;Y) &= H(X) + H(Y) - H(X,Y) \\ &= H(Y) - H(Y|X) \\ &= H(X) - H(X|Y) \end{aligned}$$

Mutual information is symmetric and thus $I(X;Y) = I(Y;X)$. It is also bounded as

$$0 \leq I(Y;X) \leq H^{**}$$

where H^{**} is the lesser of $H(X)$ and $H(Y)$, and equality in the lower bound is achieved when X is independent of Y. Equality in the upper bound with $I(X;Y) = H(X)$, is achieved when the output Y completely determines the input X for all j, and there is one and only one i for each j with $q_{j|i} > 0$.

2.3 Rate Distortion Theory

The noiseless coding theorem gives us the minimum rate at which we can code at asymptotically small decoding error. However if we are willing to tolerate a certain amount of error in order to code at rates below this limit, we want to know how the code rate and error probability are related. The branch of information theory that considers these trade os is called rate-distortion theory. It gives an analytical expression for how much compression can be achieved using lossy compression methods, and created by Claude Shannon in his foundational work on information theory. The rate is usually understood as the number of bits per data sample to be stored or transmitted, and distortion can be defined in a number of ways but the most common way is to use the mean squared error - the expected value of the square of the difference between input and output signal.

The rate and distortion can be related using the following optimization problem -

$$\min_{q_{Y|X}(y|x)} I_q(Y;X) \text{ subject to } D_q \leq D^*$$

where $q_{Y|X}(y|x)$ is the conditional probability distribution of the compressed signal Y for a given input signal X, and $I_q(Y;X)$ is the mutual information between Y and X. D_q and D^* are the distortion between X and Y for $q_{Y|X}(y|x)$, and the prescribed maximum distortion respectively. If we use the mean squared error for the distortion measure, we can define it as

$$D_q = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} q_{Y|X}(y|x) p_X(x) (x-y)^2 dxdy$$

Calculating a rate distortion function requires the stochastic description of the input X in terms of it's $p_X(x)$, and then aims at finding the conditional pdf $q_{Y|X}(y|x)$ that minimizes the rate $I_q(y;X)$ for a given distortion D^*. These definitions can be formulated measure-theoretically to account for discrete and mixed random variables. An analytical solution to this minimization problem is often difficult to obtain except in some instances. The Blahut - Arimoto algorithm [13] is an elegant iterative technique for numerically obtaining rate distortion functions of arbitrary finite input/output alphabet sources and much work has been done to extend it to more general problem instances.

2.3.1 Information Bottleneck

The information bottleneck method [140] is a technique in information theory for finding the best trade-off between accuracy and complexity when summarizing or compressing a random variable X, given a joint probability distribution $p(X,Y)$ between X and an observed relevant variable Y. The information bottleneck can also be viewed as a rate distortion problem, with a distortion function that measures how well Y is predicted from a compressed representation Z, compared to its direct prediction from X. For the compressed variable Z, the bottleneck can be represented as the following constraint optimization problem

$$\min_{p(z|x)} I(X;Z) - \beta I(Z;Y)$$

where $I(Z;Y)$ and $I(X;Z)$ the mutual information between Z and Y, and X and Z respectively. We can view $I(Z;Y)$ and $I(X;Z)$ as representing accuracy and complexity respectively. β is the Lagrange trade-off parameter. Solving this equation for $p(z|x)$, we get the solution

$$p(z|x) = \frac{p(z)}{Z(x,\beta)} \exp\left(-\beta D_{KL}[p(y|x)|p(y|z)]\right)$$

with

$$Z(x,\beta) = \sum_z p(z) \exp\left(-\beta D_{KL}[p(y|x)|p(y|z)]\right)$$

where $Z(x,\beta)$ is the normalization partition function. The detailed derivation is available in appendix (A.1).

It is important to note that the Kullback-Leibler divergence, $D_{KL}[p(y|x)|p(y|z)]$, emerged as the relevant "effective distortion measure" from our variational principle and is not assumed otherwise. It is therefore natural to consider it as the "correct" distortion $D(x,z) = D_{KL}[p(y|x)|p(y|z)]$ for quantization in the information bottleneck setting. The following three equations are solved self-consistently in an iterative manner to obtain the desired distributions for $p(z)$ and $p(z|x)$

$$p(y|z) = \sum_x p(y|x)p(x|z)$$

$$p(z) = \sum_x p(z|x)p(x)$$

$$p(z|x) = \frac{p(z)}{Z(x,\beta)} exp\left(-\beta D_{KL}[p(y|x)|p(y|z)]\right)$$

The information bottleneck approach discussed here has been used in wide variety of scenarios including in machine learning, signal processing and dynamical systems. One of the important applications of the information bottleneck is the *past-future*

information bottleneck, and it's use in predictive inference. Consider a compression of the past trajectories $\overleftarrow{x}_t$ onto the current state of a system s_t. These states are then used to map onto or make predictions of future trajectories $\overrightarrow{x}_t$. Both of these functions can be probabilistic and characterized by the probability distributions $p(s_t|\overleftarrow{x}_t)$ and $p(\overrightarrow{x}_t|s_t)$. The complexity of a model (in this case here on how the past trajectories are mapped onto the state of the system), is captured by $I(\overleftarrow{x}_t; s_t)$ the amount of information about the past trajectories that the state contains. Thus if all the past trajectories are mapped onto a single state or if each trajectory is mapped onto all the states, we have complexity $I(\overleftarrow{x}_t; s_t) = 0$. The predictive power of the model is captured by $I(s_t; \overrightarrow{x}_t)$ the amount of information the current state s_t contains about the future trajectories $\overrightarrow{x}_t$. We are thus looking for an assignment of these past trajectories onto the states that produce maximal predictive power at fixed memory. In order to do so, we solve the past-future information bottleneck as a constraint optimization problem

$$max_{p(s_t|\overleftarrow{x}_t)} \left(I(s_t; \overrightarrow{x}_t) - \lambda I(\overleftarrow{x}_t; s_t)\right)$$

where λ is the Langrange parameter controlling the tradeoff between model complexity and predictive power. For each value of λ, this optimization results in an optimal probabilistic assignment of past trajectories to model states, i.e., the information bottleneck finds a family of optimal models that are parameterized by λ. Each of these solutions satisfies the self-consistent equations

$$p(s_t|\overleftarrow{x}_t) = \frac{p(s_t)}{Z(\overleftarrow{x}_t)} exp\left(-\frac{1}{\lambda} D_{KL}[p(\overrightarrow{x}_t|\overleftarrow{x}_t)||p(\overrightarrow{x}_t|s_t)]\right)$$

$$p(\overrightarrow{x}_t|s_t) = \sum_{\overleftarrow{x}_t} p(\overleftarrow{x}_t, \overrightarrow{x}_t) \frac{p(s_t|\overleftarrow{x}_t)}{p(s_t)}$$

$$p(s_t) = \sum_{\overleftarrow{x}_t} p(s_t|\overleftarrow{x}_t) p(\overleftarrow{x}_t)$$

with the normalization constant $Z(\overleftarrow{x}_t)$

$$Z(\overleftarrow{x}_t) = \sum_{s_t} p(s_t) exp\left(-\frac{1}{\lambda} D_{KL}[p(\overrightarrow{x}_t|\overleftarrow{x}_t)||p(\overrightarrow{x}_t|s_t)]\right)$$

The solutions can be compared to the Gibbs-Boltzmann distribution and the Lagrange parameter λ has been identified as a *pseudo-temperature*, and is not to be confused with physical temperature. In the limit of large λ, fluctuations prevent any structure from being resolved. We can see clearly that $p(\overrightarrow{x}_t|s_t)$ is higher when $D_{KL}[p(\overrightarrow{x}_t|\overleftarrow{x}_t)||p(\overrightarrow{x}_t|s_t)]$ is lower, and that entails the prediction of future trajectories made from the state of the system described through $p(\overrightarrow{x}_t|s_t)$ be similar to the actual conditional distribution of the future trajectories $p(\overrightarrow{x}_t|\overleftarrow{x}_t)$ and explains the information bottleneck approach to predictive inference. This approach to predictive learning and its emergence in physical systems will be explored further in later chapters.

2.4 Computational Channels

Comptutation is the deliberate process of converting inputs into outputs using a specific model (like Turing machines, finite state automaton, Lambda calculus) and a series of steps (like an algorithm). Winograd and Cowan [14] adapted Shannon's conception of the noisy communication channel to the information theoretic characterization of noisy computation as a memoryless "computation channel." Like Shannon's discrete communication channel, it consists of an input alphabet, an output alphabet, and a set of conditional probabilities that characterize the statistical properties of the channel. A logical transformation $\mathcal{L} : x_i \rightarrow \mathcal{L}(x_i)$ maps d inputs $x_i \in \{x_i\}$ into r outputs $y_j \in \{y_j\}$ as

$$y_j = \mathcal{L}(x_i) \ \forall i \in S_j$$

where S_j is the set of indices labeling all of the inputs x_i that map into the same output y_j , as specified by the truth table for $\mathcal{L}$. For a *deterministic* computation channel implementing $\mathcal{L}$,

$$q_{j|i} = 1 \ \ \forall i \in S_j$$

$$q_{j|i} = 0 \ \ \forall i \notin S_j$$

The quantity $H(X|Y)$ serves as a measure of the average uncertainty in the channel input given the channel output, with $H(X|Y) > 0$ indicating that, for at least one output, information is lost in the mapping from input X to output Y. Information loss is undesirable in communication channels, where the goal is to infer every one of the channel inputs from the channel outputs without ambiguity ($H(X|Y) = 0$), but is completely natural in the computation channels that correspond to many logical transformations. Thus for a channel that implements a logically irreversible transformation, it follows that the information loss $\Delta I = H(X) - I(Y;X) > 0$ (with the transformation being logically reversible when $-\Delta I = 0$). Winograd and Cowan identified this connection and stated that "the destruction of information" as the defining feature of computation - *"We say that computation occurs if $H(X|Y)$ greater than 0 i.e, if the output symbols do not completely specify the input configurations; and we say that communication occurs if $H(X|Y) = 0$, i.e. if the output symbols completely specify the input configurations...It follows...that computation occurs if $H(X)$ is greater than $H(Y)$, i.e. if information is lost going from X and Y"* [14].

Even in ideal computation channels, information is necessarily lost while going from input to output that directly implements a logically irreversible operation. Such irreversible operations include AND, OR, NAND, NOR, etc., which form the cornerstones of logical operations that are performed in all general purpose computing (these operations have the property that the number of inputs d is greater than the number of outputs r). Consider the computation channel that directly implements the AND operation. If the input pmf is uniform ($p_i = \frac{1}{4} \forall i$), the input entropy is $H(X) = 2$ bits and the mutual information is $I(X;Y) = H(Y) = 0.81$ bits, indicating a loss of $H(X|Y) = 1.19$ bits of information in the channel. This (selective) destruction of information, which is required for direct implementation of the AND

operation, would render it a very poor communication channel. Ultimately information in both communication and computation channels have to be implemented in physical systems, and thus necessary to understand the physical consequences of information processing. In order to do that, we will briefly review statistical mechanics and equilibrium thermodynamics needed to do this.

2.5 Statistical Mechanics

Statistical mechanics is the branch of physics of that deals with studying physical systems with a large number of degrees of freedom [17]. The approach is based on statistical methods, probability theory and the microscopic physical laws, and can be used to explain the thermodynamic properties of large systems - both in and out of equilibrium. While classical mechanics deals with a single state, statistical mechanics introduces the *statistical ensemble*, which is a large collection of virtual, independent copies of the system in various states. The statistical ensemble is a probability distribution over all possible states of the system. In classical statistical mechanics, the ensemble is a probability distribution over phase points (points in the space of position and momentum vectors), as opposed to a single phase point in ordinary mechanics. In quantum statistical mechanics, the ensemble is a probability distribution over pure states, and can be compactly summarized as a density matrix. Both pure states and density matrices will be discussed in the next chapter.

2.5.1 Microstates & Macrostates

Microstates and macrostates provide a statistical description of physical systems in mechanics. A microstate is a specific microscopic configuration of a thermodynamic system that it may occupy with a certain probability in the course of its evolution. In a classical system of point particles, for example, a microstate defines the position and momentum of every particle. For most systems of interest, the number of microstates

is astronomically large. The macrostate of a system on the other hand is specified by the value of macroscopic variables, such as temperature, pressure, volume and density. There may be a huge number of microstates, all corresponding to the same macrostate [134]. For example, suppose if one were to measure the total energy and volume of a box of gas, there would be an enormous number of arrangements of the individual gas molecules that all add up to that energy and volume. A macrostate Ω is thus characterized by a probability distribution $p(i|\Omega)$, that describes the probability of finding the system in a certain microstate $i \in \{i\}_\Omega$ (the set of all microstates that correspond to the same macrostate Ω), corresponding to that macrostate. Statistical mechanics shows how the concepts from macroscopic observations (such as temperature and pressure) are related to the description of microscopic state that fluctuates around an average state.

2.6 Thermodynamics

Thermodynamics is the study of heat, and its relationship to energy and work [17]. These quantities are governed by the laws of thermodynamics, applicable irrespective of the composition or specific properties of the material or system in question. The strength of thermodynamics lies in its universal applicability. Equilibrium thermodynamics is the study of transfers of matter and/or energy in systems as they pass from one state of thermodynamic equilibrium to another, where 'thermodynamic equilibrium' indicates a state of balance. In an equilibrium state there are no unbalanced potentials, or driving forces, between macroscopically distinct parts of the system. An important goal of equilibrium thermodynamics is to determine for a given system in a well-defined initial equilibrium state and its surroundings, what will be the final equilibrium state of the system after a specified thermodynamic operation has changed its surroundings. The system and the surroundings are separated by a cleared defined boundary - allowing one to clearly say whether a given part of the world is in the

system or in the surroundings. If matter is not able to pass across the boundary, then the system is said to be closed; otherwise, it is open. A closed system may still exchange energy with the surroundings unless the system is an isolated one, in which case neither matter nor energy can pass across the boundary.

The entropy is a state variable and a measure of the number of possible microscopic configurations or microstates, which comply with the macroscopic state of the system. For the macrostate Ω with probability distribution $p(i|\Omega)$, we can view the entropy as

$$S_\Omega = -\sum_i p(i|\Omega) \ln p(i|\Omega)$$

as the entropy of the system in macrostate Ω. Entropy plays a very important role in the physical sciences across various disciplines. We can already start to see a possible connection between thermodynamic entropy and Shannon entropy discussed earlier in the chapter. In fact when Shannon came up with the formula for his entropy measure, he was initially unaware of thermodynamic entropy and was encouraged by von Neumann (who introduced quantum entropy) to call his measure entropy [16].

It is important to understand the relationship between these two quantities as it is significant for topics discussed later in this dissertation. Consider a system in a thermodynamic macrostate (of our choice) Ω with underlying microstate distribution $\{p_i^\Omega\}$. As before the thermodynamic entropy $S_\Omega = -\sum_i p_i^\Omega \ln p_i^\Omega$. Now consider that there is computational variable of interest $\mathcal{X}$ that is obtained by measuring some observable of the system such that it can take values from the set $\{x_k\}$. Let π_k be the probability of finding the system in the computational state x_k, where we have $\pi_k = \sum_{i \in M_k} p_i^\Omega$ where M_k is the set of all microstates of Ω that map to the k-th computational state. We have the Shannon entropy measure in bits to be $H_\mathcal{X} = -\sum_k \pi_k \log_2 \pi_k$. The relationship between the thermodynamic S_Ω and Shannon $H_\mathcal{X}$ entropy can be seen as simple extension of the Shannon grouping rule.

$$S_\Omega = k_B \ln(2) \left[H_\mathcal{X} + \sum_k \pi_k H(\{\Omega | x_k\}) \right]$$

where $H(\{\Omega|x_k\}) = -\sum_{i \in M_k} p_i^{(k)} \log_2 p_i^{(k)}$ and $p_i^{(k)} = \frac{p_i^\Omega}{\pi_k}$. There is also a $k_B \ln(2)$ factor up front in order to convert from Shannon entropy units to thermodynamic entropy units. $\ln(2)$ is to account for the that thermodynamic entropy is calculated to base e and Shannon entropy to base 2. k_B is the Boltzmann constant and it factors given the relationship thermodynamic entropy has with energy dissipation and temperature. The value of $k_B = 1.380649 \times 10^{-23}$ J/K and the definition of the Kelvin is based on this value of k_B (It is thus entirely possible to redefine all physical variables so that the value of $k_B = 1$). Thus if we chose the computationally relevant state and the thermodynamic macrostate such that there is no uncertainty in the microstate of Ω given the computationally relevant state x_k i.e. $H(\{\Omega|x_k\}) = 0$ for all k, and if we measure $H_\mathcal{X}$ in base e and rescale all physical constants such that $k_B = 1$, we would have $S_\Omega = H_\mathcal{X}$.

The three laws of thermodynamics are

1. The first law is called the Law of Conservation of Energy, states that energy cannot be created or destroyed in an isolated system and energy is transformed from one form to another. It can also be stated as the total energy in the universe is constant. Any change in the internal energy (ΔE) of a system is given by the sum of the heat (Q) that flows across its boundaries, and the work (W) done on the system by the surroundings: $\Delta E = Q + W$. For example, the transformation of stored body energy to kinetic energy (W) of the pushed car plus the heat generated (Q) by the action of pushing. The entropy can also be seen as measure of the heat loss.

2. The second law of thermodynamics states that the entropy of any isolated system always increases. It is also alternatively stated as the entropy of the universe (an isolated system) only increases and never decreases. For example,

consider a room containing a glass of melting ice as one system. The difference in temperature between the warmer room and the colder glass of ice and water is equalized as heat from the room is transferred to the ice-water mixture. The temperature of the glass and its contents and the temperature of the room achieve balance after some period of time. While the entropy of the room has decreased, the entropy of the ice and water in the glass has increased more than the entropy of the room has decreased. The second law defines the "arrow of time," in that it proves there are processes that cannot be reversed in time. It is the only fundamental physical law that distinguishes past from the future, since all microscopic dynamics are reversible in time whereas the macroscopic world is irreversible [17].

3. The third law of thermodynamics states that the entropy of a system approaches a constant minimum value as the temperature approaches absolute zero.

The entropy of a system is maximized at thermal equilibrium, with the probability distribution of microstates i and j with energies E_i and E_j is given by the Boltzmann distribution

$$\frac{p(i)}{p(j)} = exp\left[\frac{-(E_i - E_j)}{k_B T}\right]$$

or alternatively

$$p(i) = \frac{exp\left(\frac{-E_i}{k_B T}\right)}{Z}$$

where k_B is the Boltzmann constant again and T is the temperature of the thermal bath that the system is in contact with. Z is the partition function given by $Z = \sum_k exp\left(\frac{-E_k}{k_B T}\right)$. In the example above, when the system of the room and ice water system has reached temperature equilibrium, there is no further entropy change as the entropy of the final state is at its maximum. The entropy of the thermodynamic system is a measure of how far such an equalization has progressed. A detailed review

of modern equilibrium thermodynamics is available at [17]. In the next section, we will briefly extend this connection between information theory and thermodynamics by discussing the entropic and energy consequences of irreversible information processing in physical systems.

2.7 Landauer's Principle

The principle first put forward by Rolf Landauer, pertains to the lower theoretical limits of energy dissipation associated with logical computation, and provides an inextricable link between the abstract and physical notions of computation [60]. It plays a central role in resolving an important paradox associated with the second law of thermodynamics called Maxwell's Demon. It is best restated by Bennett in [19] as "*any logically irreversible manipulation of information, such as the erasure of a bit or the merging of two computation paths, must be accompanied by a corresponding entropy increase in non-information bearing degrees of freedom of the information processing apparatus or its environment.*" It allows us to relate thermodynamical quantities to the amount of information associated with the system. The *entropic* and *energetic* forms of the principle are given as

$$\Delta S \geq -k_B \ln(2) \Delta I$$

$$\Delta E \geq -k_B T \ln(2) \Delta I$$

where k_B is the Boltzmanns constant, T is the absolute temperature of the environment and $-\Delta I$ is the amount of information lost in an operation. The entropic form of Landauer's principle indicates that the entropy increase (ΔS) (in thermodynamic units) is lower bounded by $k_B \ln(2)$ per bit of information lost ($-\Delta I$) in the information-processing operation. The energetic form associates a minimum energy increase (Δ) of $k_B T \ln(2)$ per bit of information lost. It is commonly assumed in these

inequalities that the loss of information from a physical system is a state transformation that reduces uncertainty in the system state, as quantified by a self-referential information measure, defined in the terms of the state of the system undergoing the information loss.

Landauer's principle remains a topic that is widely studied given its connection to fundamental concepts of information, with continued debate on its validity. Questions arise out of confusion over how the systems and their boundaries are specified, how entropies, energies and information are defined in these systems of interest. The lack of rigor in Landauer's original paper to arrive at the inequalities only compounded the problem further. There is a rich history of work seeking to clarify the wide range of issues surrounding Landauer's Principle [20]. In the next chapter, we shall briefly review the *Referential framework* to physical information, where information in a system is a correlational measure that is described with respect to a referent. This referent remains unchanged during the process of information loss, and provides a suitable framework for studying information processing in computing systems.

2.8 Summary

In this chapter, we reviewed some of the critical ideas that are necessary for this dissertation moving forward. We began with the discussion of a communication channel, and how the need to quantify the amount of information transmitted across a channel inspired the birth of the important field of information theory. Definitions for self-entropy, joint and conditional entropy and Shannon mutual information were provided. We then continued with the concept of a computation channel and characterized computation as a necessary loss of information. Definitions of micro and macrostates, thermodynamic entropy as well as the laws of thermodynamics were briefly discussed. The chapter ended with the introduction of Landauer's principle, an important principle pertaining to the physical consequences associated with the

manipulation of information in a system. In the next chapter, we will continue down this path of the importance of the physicality of information, by studying quantum systems and how classical information is instantiated in the states of these systems.

CHAPTER 3

INTRODUCTION TO PHYSICAL INFORMATION THEORY AND THE REFERENTIAL APPROACH

In the previous chapter, we provided a brief introduction to important concepts in classical information theory, communication and computation channels. Over the last 60 years of Moore's law scaling, the size of the device implementing computation has become extremely small, and these devices now have to be described using quantum mechanics (as opposed to classical mechanics) and quantum effects play an extremely crucial role. We note that this is still a study of classical information (distinguishable 1's and 0's) in quantum systems and not quantum information (that uses qubits and forms the basis of quantum computing).

Similar to the statistical description of systems in classical mechanics, we will introduce the density matrix formalism for quantum mechanics using notes from [21], and also discuss the realization of a computation channel in a physical system. The central concept in this framework is the density matrix of the system, and it will be used to define a quantum equivalent of classical information theoretic quantities like entropy and mutual information. There are some important distinctions between classical and quantum systems, and these differences are reflected in the important results related to these information theoretic quantities. This chapter will also contain a review of the *Referential Approach* to physical information theory [21], [22], [23], [24], [25],[26], [27]. We will define a *referent*, setup the physical system for information processing and describe its use in quantifying information loss in computing systems. This part of the chapter will conclude with revisiting Landauer's principle

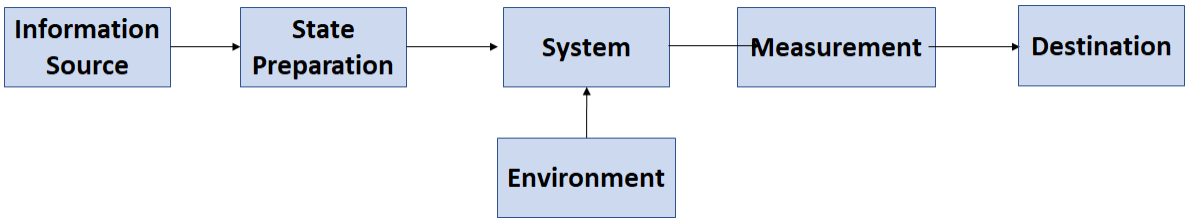

Figure 3.1. Physical view of the communication channel, adapted from [21].

for information processing in physical systems under this framework. Physical implementation of computations at the quantum scale are often noisy due to quantum indistinguishability and device variability. The chapter will conclude with a discussion on this phenomenon, and a review of information theoretic measures that are used to quantify "how well" a logical operation has been carried out in these channels. The relationship between these computational efficacy measures and the lower bound on dissipation is explored and the trade-offs between the energy efficiency and efficacy of realization is explored.

3.1 Information Processing in Physical Systems

Analogous to the discussion of the components of an abstract communication system in the previous chapter, we will briefly discuss the realization of those components in physical systems required for both communication and computation, as indicated in Fig.(3.1).

- **Source** - The source is the process generating messages and the physical system in which these messages are registered. For a classical discrete information source, the messages are realized in distinguishable source system states drawn from a fixed alphabet.

- **System** - The part of the physical universe that will function as the channel and propagate the information through via evolution of its physical state.

- **State preparation** - This the process of mapping messages into physical channel states, with transmission of a given symbol being equivalent to preparation of the channel in a corresponding signal state.

- **Measurement** - A physical interaction with the channel system that produces a measurement outcome, correlated to the pre-measurement channel state. The measurement outcome is registered in the state of another physical system that acts as the measurement apparatus.

- **Environment** - Physical systems that are not directly controlled by the state preparation process and can interact with the channel and/or measurement apparatus to affect the measurement outcomes.

In the next section, we will formalize these components with respect to classical information in quantum systems.

3.2 Physical Information in Quantum Systems

Information is encoded in the states of classical and quantum physical systems. In quantum systems, the encoding of information is done by using the quantum state vectors of the system of interest. In most scenarios, the state vector of a quantum system is either not defined, or only probabilities for various state vectors are available. In such situations, the density matrix formalism is used, and will be discussed in detail in the following section. We will introduce the concept of von Neumann entropy, quantum equivalents of joint, conditional, relative and mutual information under the density matrix formalism. These are all necessary concepts for understanding information processing in quantum systems. A detailed treatment of quantum mechanics, entropy and information in quantum systems is available in [28],[51].

3.2.1 Density Matrix Formalism

We want a description of physical systems in quantum mechanics that answer the addresses the following question -

(1) Specify the state of the isolated physical system.

(2) Describe the dynamical evolution of the system state.

(3) Be able to predict measurement outcomes.

(4) Handle description of multi-component systems.

In quantum mechanics, the state of an isolated physical system is represented by a normalized state vector in a complex Hilbert state space of the system. The density matrix formalism is used in the case where the state vector for a system is not defined or the state vector is not known; and only the probabilities of various vectors are known.

Definition: *The density matrix operator is a positive operator with unit trace defined on the complex Hilbert state space of the system, and represents a statistical description of a quantum system.*

Since it is positive operator, it is Hermitian and normal and the trace of the operator $Tr[\hat{\rho}] = 1$. Now consider a quantum system that is known to be in some state vector $\psi_i\rangle$ from the fixed set $\{|\psi_i\rangle\}$, where the $|\psi_i\rangle$ are normalized but need not be orthogonal. Let p_i indicate the probability that the system is with state vector $|\psi_i\rangle$. The possible states of the system as indicated by the state vector, together with their corresponding probabilities, constitutes an *ensemble* of states denoted as $\epsilon^{\mathcal{S}} = \{p_i, ||\psi_i\rangle\langle\psi_i|\}$. $|\psi_i\rangle\langle\psi_i|\}$ represents the outer product of the vector $|\psi_i\rangle$ with itself. The density operator associated with this ensemble of the system is given by

$$\hat{\rho} = \sum_i p_i |\psi_i\rangle\langle\psi_i|$$

In a more general case, we construct an ensemble of mixed states given by $\{p_i, \hat{\rho}_i\}$ where we have p_i to be the probability of the mixed state $\hat{\rho}_i$. The mixed in turn is given as an expectation over the outer-product of the state vectors $|\psi_n^{(i)}\rangle\langle\psi_n^{(i)}|$. Thus we have $\hat{\rho}_i = \sum_{n=1}^{N} p_n^{(i)} |\psi_n^{(i)}\rangle\langle\psi_n^{(i)}|$ and $\sum_n p_i^{(n)} = 1$. The density operator of the ensemble of mixed states is given by

$$\hat{\rho} = \sum_i p_i \hat{\rho}_i$$

While the trace of the density operator $\hat{\rho}$ is equal to one for both pure and mixed state, we have that for pure states $Tr[\hat{\rho}^2] = 1$ and for mixed states $Tr[\hat{\rho}^2] < 1$.

The density operator for an isolated quantum system evolves in time according to the Louiville equation, which can be viewed as a version of the time-dependent Schroedinger equation for density operators. The change in the system density operator $\hat{\rho}(t)$ with respect to time t is given by the equation below

$$i\bar{h}\frac{d\hat{\rho}(t)}{dt} = [\hat{H}, \hat{\rho}(t)] = \hat{H}\hat{\rho}(t) - \hat{\rho}(t)\hat{H}$$

where $\hat{H}$ is the Hamiltonian operator for the system. $[A, B]$ is the commutation operation over two operators and given as $[A, B] = AB - BA$. This can also be written for an isolated system in state $\hat{\rho}(t_1)$ at time t_1, and it will evolve to state $\hat{\rho}(t_2)$ at time t_2 according to the equation

$$\hat{\rho}(t_2) = \hat{U}(t_1, t_2)\hat{\rho}(t_1)\hat{U}(t_1, t_2)^{\dagger}$$

where $\hat{U}(t_1, t_2)$ is the time development operator given by

$$\hat{U}(t_1, t_2) = exp\left[-\tfrac{i}{\bar{h}}\hat{H}(t_2 - t_1)\right]$$

i is the complex number such that $i^2 = -1$ and $\bar{h} = \frac{h}{2\pi}$ with h being the Planck constant.

Measurements are an important aspect of quantum mechanics, since they differ significantly from measurements in classical mechanics (A measurement can often be

simply thought of taking a stick, and poking a system in order to learn something about it). In classical systems, measurement of a system does not change the system state. This is not the case in the quantum systems, and the post-measurement state of the quantum system is different from the pre-measurement state. Measurements of quantum systems are characterized by a set of measurement operators $\{\hat{\mathcal{M}}_j\}$ defined on the Hilbert space of the system, each associated with one possible measurement outcome.

If measurement M is to be performed on a system in state $\hat{\rho}$, then the a priori probability for realization of the j-th outcome is

$$q_j = Tr[\hat{\mathcal{M}}_j^\dagger \hat{\mathcal{M}}_j \hat{\rho}]$$

and the post-measurement state of the system after a measurement $\mathcal{M}$ is performed and the j-th outcome is obtained

$$\hat{\rho}'_j = \tfrac{1}{q_j} \hat{\mathcal{M}}_j \hat{\rho} \hat{\mathcal{M}}_j^\dagger$$

The completeness condition for the measurement operators is given as $\sum_j \hat{\mathcal{M}}_j^\dagger \hat{\mathcal{M}}_j = 1$, where the sum is over all measurement outcomes. In the case of *blind* measurements, where we know a measurement $\mathcal{M}$ has been performed but do not know the measurement outcome, the post-measurement state of the system is a mixture given by

$$\hat{\rho}' = \sum_j q_j \hat{\rho}'_j = \sum_j \hat{\mathcal{M}}_j \hat{\rho} \hat{\mathcal{M}}_j^\dagger$$

We will now move from discussion of density operators of individual systems to composite multi-partite systems. The density operators of composite systems can be defined on the composite Hilbert space, which is a direct product of the Hilbert spaces for the component systems. The reduced density operator which provides the state of an individual subsystem can be obtained from a partial trace of the composite-system density operator excluding the subsystem of interest. For a bipartite system,

the density operator for simply separable states on Hilbert space $\mathcal{V} \otimes \mathcal{W}$ is given by the tensor product on the density operators of the individual subsystems.

$$\hat{\rho}^{\mathcal{VW}} = \hat{\rho}^{\mathcal{V}} \otimes \hat{\rho}^{\mathcal{W}}$$

where $\hat{\rho}^{\mathcal{V}}$ and $\hat{\rho}^{\mathcal{W}}$ are the density operators on $\mathcal{V}$ and $\mathcal{W}$ respectively and are fully uncorrelated. In the more general case, the states of the subsystem are not separable and are mixtures as given below.

$$\hat{\rho}^{\mathcal{VW}} = \sum_k \pi_k \hat{\rho}_k^{\mathcal{VW}}$$

where the reduced density operators for the subsystems are achieved using a partial trace operation.

$$\hat{\rho}^{[\mathcal{V}]} = Tr_{\mathcal{W}}[\hat{\rho}^{\mathcal{VW}}] \quad \hat{\rho}^{[\mathcal{W}]} = Tr_{\mathcal{W}}[\hat{\rho}^{\mathcal{VW}}]$$

The partial trace operation over $\mathcal{W}$ of an outer product $|v \otimes w\rangle\langle v' \otimes w'|$ over $\mathcal{V} \otimes \mathcal{W}$ is define as

$$Tr_{\mathcal{W}}[|v \otimes w\rangle\langle v' \otimes w'| = |v\rangle\langle v'|Tr_{\mathcal{W}}|w\rangle\langle w'|$$

The reduced density operator can be viewed as the 'apparent' local state of a subsystem when a state vector or density operator cannot be properly defined for the subsystem. Information about the state of a system is available only through measurements and the reduced density operator $\hat{\rho}^{[\mathcal{V}]}$ is equivalent to the post-measurement states of the system by performing a local measurement on the subsystem $\mathcal{V}$ of the composite system $\mathcal{VW}$.

We can summarize the postulates of quantum mechanics using the density matrix formalism discussed as follows -

(1) The state of an isolated physical system is described by a density operator defined on a complex Hilbert space, which represents the state space for the system.

(2) The density operator for an isolated quantum system evolves in time according to the Liouville equation

$$i\bar{h}\frac{d\hat{\rho}(t)}{dt} = [\hat{H}, \hat{\rho}(t)] = \hat{H}\hat{\rho}(t) - \hat{\rho}(t)\hat{H}$$

where $\hat{H}$ is the Hamiltonian operator for the system.

(3) Every measurement $\mathcal{M}$ that can possibly be performed on a quantum system is characterized by a complete set $\{\hat{\mathcal{M}}_j\}$ of measurement operators defined on the Hilbert space of the system, each associated with one possible outcome of the measurement.

(4) The state of a composite physical system is described by a density operator defined on a composite-system Hilbert state space, which is the direct product of the Hilbert spaces for the component systems. The state of an individual subsystem of a composite system is described by a reduced density operator, defined as the density operator obtained from a partial trace of the composite-system density operator excluding the subsystem of interest.

Using the density matrix formalism from this section, we will now build into the definitions of von Neumann entropy and quantum mutual information for physical quantum systems as seen in [21].

3.2.2 Von Neumann Entropy

The von Neumann entropy (or quantum entropy) associated with a density operator $\hat{\rho}$ is

$$S(\hat{\rho}) = -Tr[\hat{\rho}\log\hat{\rho}].$$

If the density operator $\hat{\rho}$ can be written in the form of a spectral decomposition as $\hat{\rho} = \sum_i \lambda_i |\phi_i\rangle\langle\phi_i|$, where λ_i and $|\phi_i\rangle$ are the eigenvalues and eigenvectors respectively. Then $\log\hat{\rho}$ is an operator given as

$$\log \hat{\rho} = \sum_i \log(\lambda_i) |\phi_i\rangle\langle\phi_i|$$

$S(\hat{\rho})$ maps the density operator into a real number, much as the Shannon entropy $H(\{p_i\})$ maps a probability distribution $\{p_i\}$ into a real number. $S(\hat{\rho})$ can be most conveniently calculated by solving for the eigenvalues $\{\lambda_i\}$ of $\hat{\rho}$, and applying the result: *the von Neumann entropy of $\hat{\rho}$ is the Shannon entropy of its eigenvalue spectrum.*

$$S(\hat{\rho}) = -\sum_i \lambda_i \log \lambda_i$$

The von Neumann entropy is non-negative and $S(\hat{\rho} = |\psi_i\rangle\langle\psi_i|) = 0$ for pure state $|\psi\rangle$ as the density operator for any pure state has identically one eigenvalue which is $\lambda = 1$. In general for an ensemble of mixed states $\{p_i, \hat{\rho}_i\}$ it is bounded as

$$\sum_i p_i S(\hat{\rho}_i) \leq S(\hat{\rho}) \leq H(\{p_i\}) + \sum_i p_i S(\hat{\rho}_i)$$

Equality is achieved in the upper bound when the density operators $\hat{\rho}_i$ have support on orthogonal spaces i.e., $\hat{\rho}_i \hat{\rho}_{i'} = \delta_{ii'} \hat{\rho}_i$ for all i, i'. For an ensemble of quantum signal states $\epsilon = \{p_i, \hat{\rho}_i\}$, $S(\hat{\rho})$ can be thought of as the "entropy of the average signal state," while the quantity $\sum_i p_i S(\hat{\rho}_i)$ represents the "averaged entropy of signal states," and $H(\{p_i\})$ is the preparation entropy, which is the Shannon entropy of the information source driving the state preparation process. The bounds now say that the entropy of the average channel state is never less than the average entropy of the channel state and never greater than the average of the channel state plus the preparation entropy. For pure signal states this bound reduces to $0 \leq S(\hat{\rho}) \leq H(\{p_i\})$.

3.2.3 Quantum Joint Entropy

The joint entropy of a composite system is simply the von-Neumann entropy of the composite system density operator.

$$S(\hat{\rho}^{\mathcal{VW}}) = -Tr[\hat{\rho}^{\mathcal{VW}} log_2 \hat{\rho}^{\mathcal{VW}}]$$

Joint entropy is bounded above as

$$S(\hat{\rho}^{\mathcal{VW}}) \leq S(\hat{\rho}^{[\mathcal{V}]}) + S(\hat{\rho}^{[\mathcal{W}]})$$

where the equality condition is achieved for the case of simply separable states when $\hat{\rho}^{\mathcal{VW}} = \hat{\rho}^{[\mathcal{V}]} \otimes \hat{\rho}^{[\mathcal{W}]}$. This inequality condition is called the *subadditivity condition*.

It is important to understand the *joint entropy* theorem [30] while studying joint entropy of quantum systems. For a density operator on $\mathcal{U} \otimes \mathcal{V}$ of the form

$$\hat{\rho}^{\mathcal{UV}} = \sum_i p_i \hat{\rho}_i^{\mathcal{UV}} = \sum_i p_i \left(|u_i\rangle\langle u_i| \otimes \hat{\rho}_i^{\mathcal{V}} \right)$$

where $|u_i\rangle$ are orthogonal pure states on $\mathcal{U}$ and $\hat{\rho}_i^{\mathcal{V}}$ are arbitrary density operators on $\mathcal{V}$. Since the operator $\hat{\rho}_i^{\mathcal{UV}}$ has orthogonal support on $\mathcal{U} \otimes \mathcal{V}$, using the relationship $S(\hat{\rho}) = H(\{p_i\}) + \sum_i p_i S(\hat{\rho}_i)$ when the $\hat{\rho}_i$'s are orthogonal i.e.$\hat{\rho}_i \hat{\rho}_j = \delta_{ij}$ (from the previous section), we obtain the joint entropy theorem

$$S(\hat{\rho}^{\mathcal{UV}}) = H(\{p_i\}) + \sum_i p_i S(\hat{\rho}_i^{\mathcal{UV}})$$

for states of the form given above.

3.2.4 Quantum Conditional Entropy

The quantum conditional entropies for the composite system $\mathcal{VW}$ are defined as

$$S(\hat{\rho}^{[\mathcal{V}]}|\hat{\rho}^{[\mathcal{W}]}) = S(\hat{\rho}^{\mathcal{VW}}) - S(\hat{\rho}^{[\mathcal{W}]})$$

$$S(\hat{\rho}^{[\mathcal{W}]}|\hat{\rho}^{[\mathcal{V}]}) = S(\hat{\rho}^{\mathcal{VW}}) - S(\hat{\rho}^{[\mathcal{V}]})$$

Like the classical case, the quantum conditional is not symmetric $S(\hat{\rho}^{[\mathcal{V}]}|\hat{\rho}^{[\mathcal{W}]}) \neq S(\hat{\rho}^{[\mathcal{W}]}|\hat{\rho}^{[\mathcal{V}]})$. However unlike the classical case, the quantum conditional entropy can be negative.

3.2.5 Quantum Relative Entropy & Quantum Mutual Information

The quantum relative entropy of two density operators denoted by $\hat{\rho}$ and $\hat{\sigma}$ defined on the same vector space are

$$S(\hat{\rho}||\hat{\sigma}) = Tr[\hat{\rho}log_2\hat{\rho}] - Tr[\hat{\rho}log_2\hat{\sigma}]$$

The relative entropy acts as distance measure capturing the differences between the two density operators. It is bounded as $0 \leq S(\hat{\rho}||\hat{\sigma}) \leq \infty$, with equality in the lower bound obtained if and only if the two density operators are the same $\hat{\sigma} = \hat{\rho}$. The relative entropy is not symmetric i.e. $S(\hat{\rho}||\hat{\sigma}) \neq S(\hat{\sigma}||\hat{\rho})$.

As in the classical case, the quantum relative entropy can be used to define the quantum mutual information. The quantum mutual information "between" two subsystems $\mathcal{V}$ and $\mathcal{W}$ can be defined as the quantum relative entropy between the composite system density operator $\hat{\rho}^{\mathcal{VW}}$ and the density operator implied by assuming that the two subsystems are simply separable using the reduced density operators, i.e. $\hat{\rho}^{[\mathcal{V}]} \otimes \hat{\rho}^{[\mathcal{W}]}$. The quantum mutual information or the correlation entropy between $\mathcal{V}$ and $\mathcal{W}$ is given as

$$\mathcal{I}(\hat{\rho}^{[\mathcal{V}]}; \hat{\rho}^{[\mathcal{W}]}) = S(\hat{\rho}^{[\mathcal{V}]}) + S(\hat{\rho}^{[\mathcal{W}]}) - S(\hat{\rho}^{\mathcal{VW}})$$

The quantum mutual information is symmetric with $\mathcal{I}(\hat{\rho}^{[\mathcal{V}]}; \hat{\rho}^{[\mathcal{W}]}) = \mathcal{I}(\hat{\rho}^{[\mathcal{W}]}; \hat{\rho}^{[\mathcal{V}]})$, and non-negative $\mathcal{I}(\hat{\rho}^{[\mathcal{V}]}; \hat{\rho}^{[\mathcal{W}]}) \geq 0$. Equality in the lower bound $\mathcal{I}(\hat{\rho}^{[\mathcal{V}]}; \hat{\rho}^{[\mathcal{W}]}) = 0$, is obtained for uncorrelated mixtures with $\hat{\rho}^{\mathcal{VW}} = \hat{\rho}^{\mathcal{V}} \otimes \hat{\rho}^{\mathcal{W}}$. For perfectly correlated mixtures of the form

$$\hat{\rho}^{\mathcal{VW}} = \sum_i p_i |v_i \otimes w_i\rangle\langle v_i \otimes w_i|$$

with $\hat{\rho}^{[\mathcal{V}]} = \sum_i p_i |v_i\rangle\langle v_i|$ and $\hat{\rho}^{[\mathcal{W}]} = \sum_i p_i |w_i\rangle\langle w_i|$. Thus we get $S(\hat{\rho}^{\mathcal{VW}}) = S(\hat{\rho}^{[\mathcal{V}]}) = S(\hat{\rho}^{[\mathcal{W}]}) = H\{p_i\}$. This would make the quantum mutual information $\mathcal{I}(\hat{\rho}^{[\mathcal{V}]}; \hat{\rho}^{[\mathcal{W}]}) = H(\{p_i\})$ for perfectly correlated mixtures.

We also have the following inequalities that are very useful while dealing with the quantum mutual information in composite systems [21]

(a) Quantum relative entropy is monotonic for any two density operators $\hat{\rho}^{\mathcal{VW}}$ and $\hat{\sigma}^{\mathcal{VW}}$ on the space $\mathcal{V} \otimes \mathcal{W}$. This can also be viewed as the fact that one cannot increase the relative entropy by performing the partial trace on a system.

$$S(\hat{\rho}^{\mathcal{VW}}||\hat{\sigma}^{\mathcal{VW}}) \geq S(\hat{\rho}^{\mathcal{W}}||\hat{\sigma}^{\mathcal{W}})$$

(b) Quantum relative entropy also exhibits the property of strong sub-additivity. For a tripartite system $\mathcal{XYZ}$, we have

$$S(\hat{\rho}^{[\mathcal{X}]}) + S(\hat{\rho}^{[\mathcal{Y}]}) \geq S(\hat{\rho}^{[\mathcal{XZ}]}) + S(\hat{\rho}^{[\mathcal{YZ}]})$$

which can be equivalently written as

$$S(\hat{\rho}^{\mathcal{XYZ}}) + S(\hat{\rho}^{[\mathcal{Y}]}) \geq S(\hat{\rho}^{[\mathcal{XY}]}) + S(\hat{\rho}^{[\mathcal{YZ}]})$$

3.2.6 Holevo Information and Accessible Information

The bound on von-Neumann entropy from the section above can be written as

$$0 \leq S(\hat{\rho}) - \sum_i p_i S(\hat{\rho}_i) \leq H(\{p_i\}) + \sum_i p_i S(\hat{\rho}_i) - \sum_i p_i S(\hat{\rho}_i)$$

$$0 \leq \chi(\epsilon) \leq H(\{p_i\})$$

with $\chi(\epsilon) = S(\hat{\rho}) - \sum_i p_i S(\hat{\rho}_i)$, which is called the Holevo information or sometimes the entropy defect for the ensemble $\epsilon = \{p_i, \hat{\rho}_i\}$. The Holevo information defined for an ensemble is very important in understanding the amount of *accessible information* $\mathcal{I}_{acc}$ for that ensemble. The accessible information is the maximum mutual information between the random variables associated with the source output X and the channel output Y, that can be from an optimum measurement M of an ensemble ϵ .

$$\mathcal{I}_{acc} = max_M I(Y:X)$$

It is very hard to calculate the $\mathcal{I}_{acc}$ for a general signal ensemble but there is very useful upper bound provided by the Levitin-Holevo theorem [31]. *Theorem:* For a quantum channel $\mathcal{S}$ and signal ensemble $\epsilon^{\mathcal{S}} = \{p_i, \hat{\rho}_i^{\mathcal{S}}\}$ with an associated density operator $\hat{\rho} = \sum_i p_i \hat{\rho}_i$, the accessible information is upper bounded as

$$\mathcal{I}_{acc} \leq \chi(\epsilon^{\mathcal{S}})$$

where $\chi(\epsilon^{\mathcal{S}})$ is the Holevo information of the ensemble $\chi(\epsilon^{\mathcal{S}})$ and given by

$$\chi(\epsilon^{\mathcal{S}}) = S(\hat{\rho}^{\mathcal{S}}) - \sum_i p_i S(\hat{\rho}_i^{\mathcal{S}})$$

$\chi(\epsilon)$ and $I(X;Y)$ have the same bounds. They are both non-negative and upper bounded by the Shannon entropy of the source. The upper bound is achieved when the signal states have support on orthogonal subspaces, and the accessible information is less than or equal to the preparation entropy, and can be achieved using an appropriate measurement. While the Holevo information places a fundamental physical limit on the mutual information for the channel, it does not identify measurements that would actually achieve equality in the bound or even tell us if such measurements exist. The Holevo bound can also be seen as a corollary of the Schumacher, Westmoreland and Wootters bound (SWW) bound [32].

$$\mathcal{I}_{acc} \leq \chi(\epsilon^{\mathcal{S}}) - \sum_j q_j \chi(\epsilon_j^{\mathcal{S}})$$

This is a tighter bound than the Holevo bound, and says that the accessible information can never exceed the Holevo information of the as-prepared signal ensemble less the average Holevo information "left in" the system after measurement.

3.3 Referential Approach to Physical Information Theory

The referential approach is based on the fundamental premise that *information is always about something else* - and quantified as a measure of correlation between

the system and a referent [24], [23], [25], [26]. The approach has many significant advantages, one being the clear divide between the entropic self information of a system and the mutual information of a system with a referent. Since quantum mutual information is defined between two different systems, it cannot be defined between the density operators of the same system at two different time instants. However the referential approach allows the calculation of information loss in the system over time with respect to an referent that remains unchanged during the course of the information processing operation. Furthermore, in terms of engineering applications in computing systems, the approach proves to be very beneficial, as the information we manipulate and perform operations upon are usually physical encodings of input information which is present in another location, for example the memory which can act as our referent. The physical states of the memory are perfect for providing a referent, as they remain unchanged during the course of the computation in the arithmetic and logic unit (ALU) and allow us to measure the correlational information between the memory and the ALU before and after the computation. Since such memory elements like flip-flops and latches that provide storage capabilities are used in abundance in the intermediate stages of multiple cycle calculations, analyzing such processes using the referential approach can provide crucial insight. Thus the referential approach to physical information theory must be explored in detail to reap its full benefits. This approach has been explored in detail in [22], [23], [24], [25],[26], [27]. In the next section, we will setup the physical system under this approach in order to analyze the information loss, entropic and energetic costs associated with implementing a logical transformation.

3.4 Logical Transformations under Referential Approach

In this section, we will review the description of logical transformations under the referential approach from [22]. Logical transformations are an integral component

of computing systems and it is necessary to have a physically grounded description of these operations that allow us to determine the ultimate limits of dissipation associated with their realization. These logical transformations will be realized as $\mathcal{L}$ machines, a concept introduced by Ladyman, Presnell, Short and Groisman for idealized computation [50],[48]. They define this $\mathcal{L}$ machine as a *"hybrid physical - logical entity that combines a physical device, a specification of which physical states of that device correspond to various logical states, and an evolution of that device which corresponds to the logical transformation."*.

The original description assumed idealized computation, not taking into the effect of noise on the physical realization. This was accounted for and a more general description of logical $\mathcal{L}$ machines was provided under the referential approach in [22]. We will adopt the same description from [21], [22] to describe the input and output ensembles of a d-input r-output logical transformation (with $d \leq r$).

3.4.1 Input and Output Ensembles

In order to consider the implementation of a d-input r-output logical transformation $\mathcal{L}$ via evolution of the system $\mathcal{S}$, we will start by defining a $\mathcal{L}$-referent $\mathcal{R}_{\mathcal{L}}$ associated with a d-input r-output logical transformation $\mathcal{L}$. The $\mathcal{L}$-referent consists of

- A bipartite quantum system $\mathcal{R}_{\mathcal{L}} = \mathcal{R}_{in}\mathcal{R}_{out}$.
- A set $\{\hat{r}_i^{\mathcal{R}_{in}}\}$ of d distinguishable pure states of $\mathcal{R}_{in}$.
- A set $\{\hat{r}_j^{\mathcal{R}_{out}}\}$ of r distinguishable pure states of $\mathcal{R}_{out}$.
- A set $\{\hat{r}_i^{\mathcal{R}_{\mathcal{L}}}\}$ of d product states of $\mathcal{R}_{\mathcal{L}}$ - $\hat{r}_i^{\mathcal{R}_{\mathcal{L}}} = \hat{r}_i^{\mathcal{R}_{in}} \otimes \hat{r}_j^{\mathcal{R}_{out}}$ for all $i \in \{i\}_j = \{i|\mathcal{L}(x_i) = y_j\}$.

where $\mathcal{L}$ is logical transformation that maps d logical input states $x_i \in \{x_i\}$ into r logical output states $y_j \in \{y_j\}$ via $x_i \leftarrow \mathcal{L}(x_i) = y_j$. The input referent in most

applications will be a real physical system which contains a physical instantiation of the logical input that will remain unchanged until the process of computation is complete. These include the cache, latches and flip-flops in the intermediate stages of a multi-staged logical computation. The output referent is a perfect physical instantiation of the logical outputs of a perfect logical transformation. It need not exist and as the name suggests, it provides a reference to which we can compare our actual physical outputs of the logical transformation.

The input ensemble is $\chi(\epsilon_X^{\mathcal{R}_\mathcal{L}\mathcal{S}}) = \{p_i, \hat{\rho}_i^{\mathcal{R}_\mathcal{L}\mathcal{S}}\}$ where p_i is the probability that $\mathcal{R}_\mathcal{L}\mathcal{S}$ is initially prepared in the state $\hat{\rho}_i^{\mathcal{R}_\mathcal{L}\mathcal{S}} = \hat{r}_i^{\mathcal{R}_\mathcal{L}} \otimes \hat{\rho}_i^{\mathcal{S}}$ corresponding to the i-th logical input x_i. The density operator describing the statistical state of this ensemble is

$$\hat{\rho}^{\mathcal{R}_\mathcal{L}\mathcal{S}} = \sum_{i=1}^{d} p_i \hat{\rho}_i^{\mathcal{R}_\mathcal{L}\mathcal{S}}$$

In order to obtain the output ensemble, all the members of the input ensemble must be evolved via $\mathcal{C}$, a quantum operation (which is a linear, completely positive map from the set of density operators into itself) to obtain the evolved input ensemble $\chi(\epsilon_X^{\mathcal{R}_\mathcal{L}\mathcal{S}'}) = \{p_i, \hat{\rho}_i^{\mathcal{R}_\mathcal{L}\mathcal{S}'}\}$, where $\hat{\rho}_i^{\mathcal{R}_\mathcal{L}\mathcal{S}'} = \hat{r}_i^{\mathcal{R}_\mathcal{L}} \otimes \mathcal{C}(\hat{\rho}_i^{\mathcal{S}})$. The elements of the output ensemble are $\epsilon_Y^{\mathcal{R}_\mathcal{L}\mathcal{S}'} = \{q_j, \hat{\rho}_j^{\mathcal{R}_\mathcal{L}\mathcal{S}'}\}$, can then be projected out of the statistical state $\hat{\rho}^{\mathcal{R}_\mathcal{L}\mathcal{S}'} = \sum_{i=1}^{d} p_i \hat{\rho}_i^{\mathcal{R}_\mathcal{L}\mathcal{S}'}$ out of the evolved input ensemble as

$$\hat{\rho}_j^{\mathcal{R}_\mathcal{L}\mathcal{S}'} = \tfrac{1}{q_j} \hat{\Pi}_j^{\mathcal{R}_\mathcal{L}\mathcal{S}} \hat{\rho}^{\mathcal{R}_\mathcal{L}\mathcal{S}} \hat{\Pi}_j^{\mathcal{R}_\mathcal{L}\mathcal{S}} = \sum_{i \in S_j} p_i^{(j)} \hat{\rho}_i^{\mathcal{R}_\mathcal{L}\mathcal{S}'}$$

where $\hat{\Pi}_j^{\mathcal{R}_\mathcal{L}\mathcal{S}}$ is the projector associated with the j-th logical output and is given by $\hat{\Pi}_j^{\mathcal{R}_\mathcal{L}\mathcal{S}} = \sum_{i \in \{i\}_j} \hat{\pi}_i^{\mathcal{R}_\mathcal{L}\mathcal{S}}$ [1]

---

[1] The general form for this projector is of the form

$$\hat{\prod}_j^{\mathcal{R}_\mathcal{L}\mathcal{S}} = \hat{I}^{\mathcal{R}_\mathcal{L}\mathcal{S}} - \prod_{i \in S_j} \left(\hat{I}^{\mathcal{R}_\mathcal{L}\mathcal{S}} - \hat{\pi}_i^{\mathcal{R}_\mathcal{L}\mathcal{S}} \right)$$

where $\hat{I}^{\mathcal{R}_\mathcal{L}\mathcal{S}}$ is the identity of $\hat{\mathcal{H}}^{\mathcal{R}_\mathcal{L}\mathcal{S}}$. The $\hat{\pi}_i^{\mathcal{R}_\mathcal{L}\mathcal{S}}$ are mutually orthogonal and the product terms on the right reduces to $\hat{I}^{\mathcal{R}_\mathcal{L}\mathcal{S}} - \sum_{i \in S_j} \hat{\pi}_i^{\mathcal{R}_\mathcal{L}\mathcal{S}}$ and we get the reduced expression for the identity $\hat{\Pi}_j^{\mathcal{R}_\mathcal{L}\mathcal{S}}$.

,on $H^{\mathcal{R}_\mathcal{L}} \otimes H^{\mathcal{S}}$ with $\hat{\pi}_i^{\mathcal{R}_\mathcal{L}\mathcal{S}} = \hat{r}_i^{\mathcal{R}_\mathcal{L}} \otimes \hat{\pi}_i^{\mathcal{S}}$. $\hat{\pi}_i^{\mathcal{S}}$ is the identity of the support of $\mathcal{C}(\hat{\rho}_i^{\mathcal{S}})$, and $\hat{\Pi}_j^{\mathcal{R}_\mathcal{L}\mathcal{S}}$ is the identity for the support subspace associated with the j-th output.

We have the probability of the j-th output $q_j = Tr[\hat{\Pi}_j^{\mathcal{R}_\mathcal{L}\mathcal{S}} \hat{\rho}^{\mathcal{R}_\mathcal{L}\mathcal{S}'}] = \sum_{i \in S_j} p_i$ and $p_i^{(j)} = \frac{p_i}{q_j}$. We define N-output ensembles $\epsilon_j^{\mathcal{R}_\mathcal{L}\mathcal{S}'} = \{p_i^{(j)}, \hat{\rho}_i^{\mathcal{R}_\mathcal{L}\mathcal{S}'}\} | i \in S_j\}$, associated with the r logical outputs, and the j-th reduced density operator is given by

$$\hat{\rho}_j^{\mathcal{S}'} = Tr_{\mathcal{R}_\mathcal{L}}[\hat{\rho}_j^{\mathcal{R}_\mathcal{L}\mathcal{S}'}] = \sum_{i \in S_j} p_i^{(j)} \mathcal{C}(\hat{\rho}_i^{\mathcal{S}})$$

$\hat{\rho}_j^{\mathcal{S}'}$ is the physical representation of the j-th output stage or y_j, and provides a statistical representation of the outputs for input distribution $\{p_i\}$ in the state of device $\mathcal{S}$ alone.

3.4.2 Revisiting Landauer's Principle - Entropic & Energy Cost of Information Processing

We will now restate the version of Landauer's Principle [60] from [21] and [22] under the referential approach. Consider a closed composite system consisting of an "information bearing" subsystem $\mathcal{RS}$ and environment $\mathcal{B}$. Let the states of $\mathcal{R}$ and $\mathcal{S}$ be initially correlated and assume that $\mathcal{RS}$ is initially isolated from $\mathcal{B}$. An operation processing information about $\mathcal{R}$ which is encoded in $\mathcal{S}$ is given as an unitary evolution $\hat{\mathcal{U}}$ of $\mathcal{RSB}$ that involves only interactions between $\mathcal{S}$ and $\mathcal{B}$. The entropic form of Landauer's principle is given in Eq.(3.1) - the entropy increase is lower bounded at $k_B \ln(2)$ per bit of information that is lost during the information processing operation. The detailed derivation of this bound is avialable in appendix (A.2).

$$\Delta S \geq -k_B \ln(2) \Delta I \tag{3.1}$$

where ΔS is the increase in total entropy of the system and bath combined and $-\Delta I$ is the loss of quantum mutual information between the system $\mathcal{S}$ and the referent $\mathcal{R}$ over the information processing operation.

In order to study the energy costs of operations that discard information, like irreversible logical operations, we assume that the environment is initially a thermal bath at temperature T. Using the entropic derivation of Landauers Principle, we can obtain the energetic version of Landauer's Principle -

$$\Delta\langle E^{\mathcal{B}}\rangle \geq -k_B T \ln(2) \Delta S^{\mathcal{S}}$$

where $\Delta\langle E^{\mathcal{B}}\rangle$ is the expected energy increase in the environment and $-\Delta S^{\mathcal{S}}$ is the loss in von Neumann entropy of the system $\mathcal{S}$. A detailed derivation of the bound from [21] is available in appendix (A.2).

This inequality implies that there is a minimum environmental energy increase of $k_B T \ln(2)$ associated with every operation that reduces the system entropy $\Delta S^{\mathcal{S}}$ by 1 bit, regardless of how much information is lost. The bound thus accommodates scenarios in which entropy of $\mathcal{S}$ is increased and energy is transferred out of the environment during processes that cause loss of information. This stands in contrast with the traditional form of Landauers Principle which associates a energy transfer into the environment with loss of information.

We will now illustrate this with a simple thermal reset example. Consider a simple system $\mathcal{RS}$ which is initially perfectly correlated given by the density operator

$$\hat{\rho}^{\mathcal{RS}} = \sum_{i=1}^{2} p_i (|r_i\rangle\langle r_i| \otimes |s_i\rangle\langle s_i|)$$

where $\{|r_i\rangle\}$ and $\{|s_i\rangle\}$ are orthonormal sets spannings the spaces $\mathcal{H}^{\mathcal{R}}$ and $\mathcal{H}^{\mathcal{S}}$ respectively. The quantum mutual information between $\mathcal{R}$ and $\mathcal{S}$, and the entropy of $\mathcal{S}$ is given as

$$I(\hat{\rho}^{\mathcal{R}}; \hat{\rho}^{\mathcal{S}}) = H(\{p_i\})$$

$$S(\hat{\rho}^{\mathcal{S}}) = H(\{p_i\})$$

Let the system $\mathcal{S}$ interact with a large thermal bath $\mathcal{B}$ at temperature T - that completely thermalizes the state of $\mathcal{S}$ such that

$$\hat{\rho}^{\mathcal{RS}'} = \sum_{i=1}^{2} p_i(|r_i\rangle\langle r_i| \otimes \hat{\rho}_{th}^{\mathcal{S}'}$$

where $\hat{\rho}_{th}^{\mathcal{S}'} = Z^{-1} \exp -\frac{-H^{\mathcal{S}}}{k_B T}$. $H^{\mathcal{S}}$ is the Hamiltonian of the system with eigenvalues E_1 and E_2. Since the nal states of $\mathcal{R}$ and $\mathcal{S}$ are completely uncorrelated, the final quantum mutual information is

$$I(\hat{\rho}^{\mathcal{R}}; \hat{\rho}^{\mathcal{S}'}) = 0$$

and all information about $\mathcal{R}$ in $\mathcal{S}$ is erased. For sufficiently low temperatures we have, $\hat{\rho}^{\mathcal{S}'} = |E_1\rangle\langle E_1|$ and $S(\hat{\rho}^{\mathcal{S}'}) = 0$. In this limit

$$\Delta I = I(\hat{\rho}^{\mathcal{R}}; \hat{\rho}^{\mathcal{S}'}) - I(\hat{\rho}^{\mathcal{R}}; \hat{\rho}^{\mathcal{S}}) = -H(\{p_i\})$$

$$\Delta S = S(\hat{\rho}^{\mathcal{S}'}) - S(\hat{\rho}^{\mathcal{S}}) = -H(\{p_i\})$$

$$\Delta S = \Delta I$$

Thus we have the general bound on energy flow to be reduced to

$$\Delta\langle E^{\mathcal{B}}\rangle \geq -k_B T \ln(2) \Delta I$$

This says that for this type of erasure operation - resetting of the system $\mathcal{S}$ to a standard pure state - at least $k_B T \ln(2)$ of energy is dissipated into the environment per bit of information erased.

3.5 Noisy Computational Channels and Efficacy Measures

In this section, we will discuss the computational efficacy measures developed in [22] to quantify noisy computational channel. A d-input, r-output discrete channel with $0 < q_{j|i} < 1$ for at least one of the outputs y_j, cannot be associated with the implementation of any logical transformation, since direct implementation requires that each x_i map into one and only one output y_j and this requirement is not met if

$0 < q_{j|i} < 1$ for any $q_{j|i}$. Thus rather than trying to answer the question "what logical transformation L is implemented by the noisy channel?", we should try and answer the question "How well does the noisy computational channel implement the logical transformation L?" These measures will be further extended in the next chapter for finite state automata.

3.5.1 Representational Faithfulness

For the computational channel to "complete the work" of implementing a logical transformation $\mathcal{L}$, then all device input states "belonging to" the same logical output of $\mathcal{L}$ must evolve into the same device output state $\mathcal{U}_{\mathcal{L}}(\mathcal{S}_i^{(in)}) = \mathcal{S}_j^{(out)} \forall i \in S_j = \{i|\mathcal{L}(x_i) = y_j\}$. This condition requires that the evolved states should contain no information that could help identify the state $\mathcal{S}_i^{(in)} \in \{\mathcal{S}_i^{(in)}\}_j$ which it is evolved. $\{\mathcal{S}_i^{(in)}\}_j$ is the set of input states $\mathcal{S}_i^{(in)}$ that map to the same j-th output state. This implies

$$I(\hat{\rho}_j^{\mathcal{R}_{in}}; \hat{\rho}_j^{\mathcal{S}'}) = \chi(\epsilon_j^{\mathcal{S}'}) = 0$$

From this, the following definition of representational faithfulness can be developed [22], [21].

Definition $\equiv$ For a quantum machine that implements a logical transformation $\mathcal{L}$ and input distribution $\{p_i\}$, the representational faithfulness is

$$f_{\mathcal{L}} \equiv 1 - \frac{1}{H_{\mathcal{L}}(X|Y)} \sum_{j=1}^{N} q_j \chi(\epsilon_j^{\mathcal{S}'})$$

where q_j and $H_{\mathcal{L}}(X|Y)$ are the j-th output probability and the conditional entropy associated with the logical transformation $\mathcal{L}$ for input distribution $\{p_i\}$ and $\chi(\epsilon_j^{\mathcal{S}'})$ is the Holevo information associated with the ensemble $\epsilon_j^{\mathcal{S}'} = \{p_i^{(j)}, \hat{\rho}_i^{\mathcal{S}'} | i \in \{i\}_j\}$ of the final reduced device states $\hat{\rho}_i^{\mathcal{S}'}$ representing the logical output states y_j of $\mathcal{L}$.

$f_{\mathcal{L}} H_{\mathcal{L}}(X|Y)$ is the average over all logical outputs, of information about the logical input that is lost in producing the physical representations of the logical outputs. It is bounded as $0 \leq f_{\mathcal{L}} \leq 1$.

3.5.2 Computational Fidelity

This efficacy measure is concerned with the distinguishability of the output states independent of their faithfulness. It is related to the amount of information about the correct logical output-encoded in output referent states- that is reflected in the final physical state of $\mathcal{S}$, i.e by the quantum mutual information

$$I(\hat{\rho}^{\mathcal{R}_{out}}; \hat{\rho}^{\mathcal{S}\prime}) = S(\hat{\rho}^{\mathcal{R}_{out}}) + S(\hat{\rho}^{\mathcal{S}\prime}) - S(\hat{\rho}^{\mathcal{R}_{out}\mathcal{S}\prime}) = \chi(\epsilon_Y^{\mathcal{S}'})$$

We obtain the following definition from [22], [21]

Definition: For a quantum machine implementing the logical transformation $\mathcal{L}$ and input distribution $\{p_i\}$, the computational fidelity is

$$F_{\mathcal{L}} = \frac{1}{H_{\mathcal{L}}(Y)}\chi(\epsilon_Y^{\mathcal{S}'})$$

where $H_{\mathcal{L}}(Y)$ is the entropy associated with the logical transformation $\mathcal{L}$ for input distribution $\{p_i\}$, and $\chi(\epsilon_Y^{\mathcal{S}'})$ is the Holevo information associated with the ensemble $\epsilon_Y^{\mathcal{S}'} = \{q_j, \hat{\rho}_j^{\mathcal{S}'}\}$ of final device states representing the logical outputs y_j of $\mathcal{L}$.

$F_{\mathcal{L}}H_{\mathcal{L}}(Y)$ indicates the amount of information about the logical output that is present in the final device state. Computational fidelity is bounded as $0 \leq F_{\mathcal{L}} \leq 1$.

3.5.3 Information Loss in Terms of Computational Fidelity and Representational Faithfulness

Using mutual information, the information about the logical input X that is lost as the system $\mathcal{S}$ evolves from its initial to final state to implement the logical transformation is

$$-\Delta I = I(\hat{\rho}^{\mathcal{R}_{in}}; \hat{\rho}^{\mathcal{S}}) - I(\hat{\rho}^{\mathcal{R}_{in}}); \hat{\rho}^{\mathcal{S}'}) \tag{3.2}$$

If $\mathcal{S}$ initially holds all the information about X, since the x_i are encoded in distinguishable input states of $\mathcal{S}$, then $I(\hat{\rho}^{\mathcal{R}_{in}}); \hat{\rho}^{\mathcal{S}}) = H(X)$ and the information loss

$$-\Delta I = H(X) - I(\hat{\rho}^{\mathcal{R}_{in}}); \hat{\rho}^{\mathcal{S}'})$$

where $I(\hat{\rho}^{\mathcal{R}_{in}}); \hat{\rho}^{\mathcal{S}'}) = \chi(\epsilon_X^{\mathcal{S}'})$ and since $\chi(\epsilon_X^{\mathcal{S}'}) \leq H(X)$, $-\Delta I \geq 0$. The information loss can be restated in terms of computational fidelity and representational faithfulness [22]

$$-\Delta I = (1 - F_{\mathcal{L}})H_{\mathcal{L}}(Y) + f_{\mathcal{L}}H_{\mathcal{L}}(X|Y) \tag{3.3}$$

The first term in Eq.(3.3) indicates the necessary desirable information loss that is required to produce faithful representations of logical output states in channels implementing the logical transformation. The second term accounts for the undesirable information loss associated with the indistinguishability of the output states. From the equation (3.3), we can clearly see that if the channel flawlessly implements the logical transformation $\mathcal{L}$ i.e. $F_{\mathcal{L}} = 1$, $f_{\mathcal{L}} = 1$, $-\Delta I = H(X|Y)$, and if a channel that produce unfaithful ($f_{\mathcal{L}} = 0$) yet perfectly distinguishable outputs ($F_{\mathcal{L}} = 1$), information loss $\Delta I = 0$ which is what is expected in a perfect communication channel.

3.5.4 Lower Bounds on Energy Dissipation in Terms of Efficacy Measures

In Eq. (3.3), we have related the information loss in a logical transformation $\mathcal{L}$ with the efficacy which indicated *how well the logical transformation L was achieved*. Since the information loss is directly related with the heat dissipation to the environment, substituting Eq. (3.3) in Eq. (A.2), we get

$$\Delta\langle E^{\mathcal{B}}\rangle \geq k_B T \ln(2)[(1 - F_{\mathcal{L}})H_{\mathcal{L}}(Y) + f_{\mathcal{L}}H_{\mathcal{L}}(X|Y) + \langle \Delta S_i^{\mathcal{S}}\rangle] \tag{3.4}$$

where $\langle \Delta S_i^{\mathcal{S}}\rangle$ is the average change in the von Neumann entropies of the system state during the logical transformation $\mathcal{L}$. We thus have a very important relation between the lower bound on the physical cost the user must pay to achieve a logical transformation, in terms of the performance metrics fidelity and faithfulness which indicate how well the logical transformation was performed.

3.6 Summary

In this chapter of technical background, we first explored the density matrix formalism of quantum mechanics including definitions for pure states, mixed states and ensembles. The concepts of von Neumann entropy, joint and conditional entropy, quantum mutual information and Holevo information were introduced and their properties discussed. This laid the groundwork to move onto the discussion of information as a physical quantity under the Referential approach, in which information is always discussed as the amount of correlation between two physical systems. The approach is highly suitable to discuss information processing in computing systems as physical processes. We set up the framework with descriptions of input and output ensembles of a logical transformation, and derived the entropy and energy forms of Landauer's principle for physical system instantiating a logical transformation under the referential approach. The chapter concluded with the introduction of two efficacy measures to capture how well a logical transformation is achieved in a noisy computational channel. The two measures - computational fidelity and representational faithfulness capture the distinguishability of the output states, and whether the physical output states contain more information than what is allowed by the abstract logical map respectively. The chapter concludes with substituting the measures in to the lower bound on energy dissipation to obtain the relationship between how well a logical transformation is physically instantiated with the minimum dissipation cost of that instantiation.

CHAPTER 4

DISSIPATION IN FINITE STATE AUTOMATA

There is a fundamental connection between irreversible information loss and energy dissipation in computational processes, as has been recognized since Landauer's work in the early 1960s [36]. An important implication of this connection is that the minimum *physical* costs required for implementation of a specified computing process can be determined from an *abstract* description of its computational behavior, provided that the abstract description appropriately captures and quantifies irreversible information loss [38],[37]. Abstract descriptions of digital computing processes based on finite-state automata (FSA) [39, 40], which are very general, powerful, and widely used, are obvious candidates. If irreversible information loss in finite automata can be quantified and tied to a sufficiently general physical description, then fundamental lower bounds on dissipation can be obtained for particular FSAs and used to explore the inherent dissipative costs of a very wide variety of computing processes and schemes.

While previous studies have addressed some aspects of irreversibility and dissipation in determinstic FSAs (e.g. [41]) and stochastic FSAs (e.g. [42]), we are aware of no general, physically grounded approach for quantifying irreversible information loss in specified FSAs and obtaining fundamental, implementation-independent lower bounds on the resulting dissipation. Such an approach is provided in the present chapter, starting with a large and important class of FSAs - deterministic, irreducible FSAs - driven by a classical information source. Park of the work discussed in this chapter has been published in [43], [44] and [45].

We will start the chapter with a discussion on reversibility and other aspects of abstract FSAs that are relevant to this work, and establish our fundamental physical description of deterministic FSAs. Following that we will state and prove a physical-information-theoretic lower bound on the average amount of energy dissipated into a thermal environment per state transition for a deterministic FSA driven by a random input source with specified statistics. This "Landauer-like" bound is proportional to the average information about past inputs that is lost from the FSA state on each transition, which we propose as a measure of the computational irreversibility for the FSA and illustrate the application of our approach to a simple FSA system. We then compare the dissipation cost of generating outputs between two types of FSAs - Mealy and Moore machines which differ in the input-output dependency. We then extend our description from deterministic FSAs to probabilistic FSAs and derive the lower bound on dissipation in terms of the information loss associated with a state transition. We then introduce FSA computational efficacy measures, similar to the ones from the previous chapter and describe the lower bound on dissipation with how well the FSA has been physically implemented. We will finally discuss the dissipation bounds of FSAs with temporally correlated inputs and see how these are different from the FSAs studied earlier in the chapter, with a simple learning system as an example. We will end this chapter summarizing the results.

4.1 Description of Finite-State Automata

4.1.1 Abstract Finite-State Automata

We begin with a brief discussion of abstract finite-state automata that emphasizes the concepts, definitions, terminology, and notation used in this work. An abstract FSA $\mathcal{F}_A \triangleq \{\{\sigma\}, \{x\}, \{\mathcal{L}\}\}$ consists of a finite set $\{\sigma\}$ of states, a set $\{x\}$ of input symbols that induce transitions between states, and a set $\{\mathcal{L}\}$ of transition rules - one for each input - that govern the state transitions. Specifically, for every input x_j

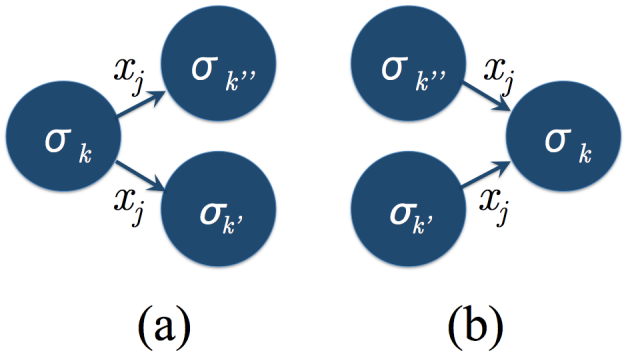

Figure 4.1. (a) State mapping disallowed in a deterministic FSA; no state σ_k can map into two different states $\sigma_{k'}$ and $\sigma_{k''}$ for any input x_j. (b) State mapping disallowed in a codeterministic FSA; no two states $\sigma_{k'}$ and $\sigma_{k''}$ can map into the same state σ_k for any input x_j. (Adapted from [46].)

there is an "input transition rule" $\mathcal{L}_j$ that maps every "current state" σ_k to a "next state" $\sigma'_{kj} = \mathcal{L}_j(\sigma_k) \in \{\sigma\}$. Outputs are generally defined for FSAs as well, but we do not consider them in this work.

The reversibility of an abstract FSA is tied to two properties of the transition rules. An FSA is *deterministic* if every input transition rule $\mathcal{L}_j$ assigns one and only one next state to every current state, and is *codeterministic* if no transition rule $\mathcal{L}_j$ maps more than one current state into any given next state (see Fig. 4.1). A deterministic FSA is *reversible* if it is also codeterministic [46], and is *irreversible* if it is not codeterministic.

In this work we will be concerned exclusively with deterministic FSAs, both reversible and irreversible. All input transition rules are necessarily bijective in reversible FSAs, whereas one or more of the $\mathcal{L}_j$ are non-injective in irreversible FSAs (as in Fig. 1(b)). We also limit our consideration to *irreducible* FSAs, in which the set $\{\mathcal{L}_j\}$ is such that every state is reachable from every other state in a finite number of transitions via some sequence of inputs.

The statistical properties of deterministic FSAs driven by random sequences – including the amount of irreversible information loss – depend directly on the input statistics. For input sequences generated by an IID classical information source, i.e. random sequences $\vec{X} = X^{(1)}X^{(2)}...$ of identically distributed discrete random variables $X^{(n)} = \{q, x\}$ (each with the same symbol set $\{x\}$ and corresponding probability mass function $\{q\}$), the conditional probability that an FSA in state σ_k will transition to state $\sigma_{k'}$ on any given step is

$$p_{k\to k'} = \sum_{j\in\{j\}_{k\to k'}} q_j$$

where $\Big(\{j\}_{k\to k'} = \{j|\mathcal{L}_j(\sigma_k) = \sigma_{k'}\}\Big)$. With this, a statistical transition matrix P with elements $p_{k\to k'}$ can be constructed. P has two properties of interest here. First, for the n-th input in the sequence $\vec{X}$, P relates the "current state" probability vector $\pi^{(n-1)}$ to the "next state" probability vector $\pi^{(n)}$ simply as $\pi^{(n)} = P\pi^{(n-1)}$. (Here $\pi^{(n)}$ is the vector with elements $\pi_k^{(n)}$, where $\pi_k^{(n)}$ is the probability that the FSA is in the state σ_k after the n-th transition. $\pi^{(n-1)}$ is defined similarly.) Second, since $\pi_k^{(n-1)} = \pi_k^{(n)}$ for all k in steady state, the "steady-state" occupation probabilities for the FSA states are just the elements of the eigenvector π of P with eigenvalue 1 ($\pi = P\pi$). This eigenvector is unique for an irreducible FSA.

4.1.2 Physical Finite-State Automata

We now construct a very general *physical* description of a deterministic FSA, defined abstractly as above. We formalize this description after identifying the physical realizations of FSA states, inputs, and transitions.

- **States:** Abstract FSA states σ_k are faithfully represented in distinguishable physical states[1] $\hat{\sigma}_k^{\mathcal{S}}$ of a quantum-mechanical register system $\mathcal{S}$, which interacts

[1]The $\hat{\sigma}_k^{\mathcal{S}}$ are quantum mechanical density operators. States $\hat{\sigma}_k^{\mathcal{S}}$ and $\hat{\sigma}_{k'}^{\mathcal{S}}$ are distinguishable if they have orthogonal support, so $Tr[\hat{\sigma}_k^{\mathcal{S}}\hat{\sigma}_{k'}^{\mathcal{S}}] = 0$. $\hat{\sigma}_k^{\mathcal{S}}$ is a faithful physical representation of FSA state

with its local environment $\mathcal{B}$. Here $\mathcal{B}$ is taken to be a (finite) heat bath nominally in a thermal state $\hat{\rho}_{th}^{\mathcal{B}}$ at temperature T.

- **Inputs:** Random length-n input strings $\vec{X}$ are physically instantiated in the state of an n-partite "referent" system $\mathcal{R} = \mathcal{R}_0\mathcal{R}_1$, which can be regarded as a physical "input tape" that holds the output of a classical information source. Subsets $\vec{X}_k = X^{(1)}X^{(2)}...X^{(n-1)}$ of strings leading to FSA state $\hat{\sigma}_k^{\mathcal{S}}$ are represented by distinguishable mixed states $\hat{\rho}_k^{\mathcal{R}_0}$ of $\mathcal{R}_0$, and $X^{(n)}$ is represented by a mixture of pure distinguishable states $\hat{x}_j^{\mathcal{R}_1}$ of $\mathcal{R}_1$.

- **State Transitions:** The n-th state transition is realized by dynamical evolution of the state of $\mathcal{S}$, conditioned on the state of $\mathcal{R}_1$ (i.e. the n-th input) and in interaction with $\mathcal{B}$ (to the next state of $\mathcal{S}$ which is referred to as $\mathcal{S}'$). Global evolution of the interacting joint $\mathcal{R}_1\mathcal{S}\mathcal{B}$ producing this transition is assumed to be governed by the time-dependent Schrodinger equation to ensure consistency with physical law (implying unitary evolution of the state of $\mathcal{R}_1\mathcal{S}\mathcal{B}$). The n-th input remains encoded in $\mathcal{R}_1$ at the conclusion of the FSA state transition.

To complete the "physical universe" relevant to description of the FSA, the FSA's local environment $\mathcal{B}$ is embedded in a "greater environment" $\bar{\mathcal{B}}$ which acts to "rethermalize" $\mathcal{B}$ whenever it is driven from equilibrium by interaction with $\mathcal{S}$ during state transitions. $\bar{\mathcal{B}}$ is also taken to include all other subsystems required for global closure of the composite system $\mathcal{R}\mathcal{S}\mathcal{B}\bar{\mathcal{B}}$.

Consider the n-th state transition, which is depicted schematically in Fig. 4.2. Prior to this transition, the "current" FSA state encoded in the physical state of $\mathcal{S}$ is correlated with $\mathcal{R}_0$ (i.e. the first $n-1$ inputs) but not yet with $\mathcal{R}_1$ (i.e. the n-th input). At the completion of the n-th state transition, correlations will have been created

---

σ_k if it encodes no more information about the FSA history than is present in the corresponding abstract FSA description.

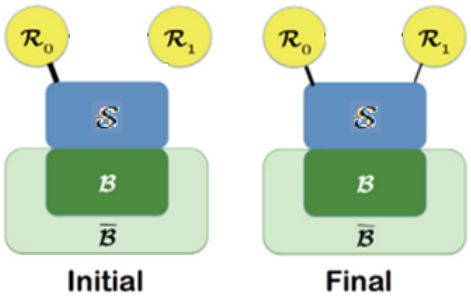

Figure 4.2. Physical description of an FSA undergoing a state transition. The system $\mathcal{S}$, which registers the FSA state, is initially correlated with previous inputs physically encoded in $\mathcal{R}_0$ (Initial). On the state transition, $\mathcal{S}$ becomes correlated with a new input encoded in $\mathcal{R}_1$ (Final). This generally weakens the preexisting correlations between $\mathcal{R}_0$ and $\mathcal{S}$, inducing dissipation into the FSA's local environment (a heat bath $\mathcal{B}$).

between the state of $\mathcal{S}$ and the state of $\mathcal{R}_1$. This weakens the correlation between $\mathcal{S}$ and $\mathcal{R}_0$ in an irreversible FSA, which amounts to an irreversible loss of information from the FSA state about its own history. We will quantify this information loss in the next section and show that it necessarily results in dissipation of energy to $\mathcal{B}$, but first provide the formal description of FSA state transitions upon which proof of the dissipation bound is based.

Initial State: Prior to the n-th input, the statistical state of the composite $\mathcal{RSB}$ is given by the density operator

$$\hat{\rho}^{\mathcal{RSB}} = \hat{\rho}^{\mathcal{R}_0\mathcal{S}} \otimes \hat{\rho}^{\mathcal{R}_1} \otimes \hat{\rho}^{\mathcal{B}}_{th}$$

or

$$\begin{aligned} \hat{\rho}^{\mathcal{RSB}} = \left(\sum_k \pi_k^{(n-1)} \{ \hat{\rho}_k^{\mathcal{R}_0} \otimes \hat{\sigma}_k^{\mathcal{S}} \} \right) \otimes \left(\sum_j q_j \hat{x}_j^{\mathcal{R}_1} \right) \otimes \hat{\rho}^{\mathcal{B}}_{th} \\ = \sum_k \sum_j \pi_k^{(n-1)} q_j \hat{\rho}_{kj}^{\mathcal{RSB}}. \end{aligned} \tag{4.1}$$

Here $\hat{\rho}_k^{\mathcal{R}_0}$ is a statistical mixture of all states of $\mathcal{R}_0$ instantiating input strings that map the initial state of the machine to the k-th FSA state $\hat{\sigma}_k^{\mathcal{S}}$, $\hat{x}_j^{\mathcal{R}_1}$ is the state encoding the n-th FSA input $x_j^{(n)}$ in $\mathcal{R}_1$, and

$$\hat{\rho}_{kj}^{\mathcal{RSB}} = \hat{\rho}_k^{\mathcal{R}_0} \otimes \hat{\sigma}_k^{\mathcal{S}} \otimes \hat{x}_j^{\mathcal{R}_1} \otimes \hat{\rho}_{th}^{\mathcal{B}}.$$

The FSA state is correlated only to the first $n-1$ inputs.

State Transition: The n-th state transition is a unitary transformation

$$\hat{\rho}^{\mathcal{RSB}'} = \hat{U}\hat{\rho}^{\mathcal{RSB}}\hat{U}^{\dagger}$$

of $\mathcal{RSB}$ involving interactions only between $\mathcal{R}_1$, $\mathcal{S}$, and $\mathcal{B}$:

$$\hat{U} = \hat{I}^{\mathcal{R}_0} \otimes \hat{U}^{\mathcal{R}_1\mathcal{SB}}.$$

If the process is to physically implement an FSA defined by abstract states $\sigma_k \in \{\sigma\}$ (physically encoded in register states $\hat{\sigma}_k^{\mathcal{S}} \in \{\hat{\sigma}^{\mathcal{S}}\}$) and by input transition rules $\mathcal{L}_j \in \{\mathcal{L}\}$, then $\hat{U}$ must be such that

$$\hat{\rho}_{kj}^{\mathcal{S}'} = Tr_{\mathcal{RB}}[\hat{U}\hat{\rho}_{kj}^{\mathcal{RSB}}\hat{U}^{\dagger}] \in \{\hat{\sigma}^{\mathcal{S}}\}.$$

This condition can be written as

$$\hat{\rho}_{kj}^{\mathcal{S}'} = \tilde{\mathcal{L}}_j(\hat{\sigma}_k^{\mathcal{S}}) \in \{\hat{\sigma}^{\mathcal{S}}\}$$

to highlight connection to the abstract FSA description, where the $\tilde{\mathcal{L}}_j \in \{\tilde{\mathcal{L}}\}$ are local, nonunitary input transition superoperators that act on $\mathcal{S}$ alone to induce the required state transitions.

Final State: At the conclusion of the n-th state transition, the state of $\mathcal{RSB}$ is:

$$\hat{\rho}^{\mathcal{RSB}'} = \sum_j q_j \left(\hat{x}_j^{\mathcal{R}_1} \otimes \sum_k \pi_k^{(n-1)} \left(\hat{\rho}_k^{\mathcal{R}_0} \otimes \hat{\rho}_{kj}^{\mathcal{SB}'} \right) \right) \tag{4.2}$$
$$= \sum_k \sum_j \pi_k^{(n-1)} q_j \hat{\rho}_{kj}^{\mathcal{RSB}'}$$

where

$$\hat{\rho}_{kj}^{\mathcal{RSB}'} = \hat{\rho}_k^{\mathcal{R}_0} \otimes \tilde{\mathcal{L}}_j(\hat{\sigma}_k^{\mathcal{S}}) \otimes \hat{x}_j^{\mathcal{R}_1} \otimes \hat{\rho}_{kj}^{\mathcal{B}'}.$$

The new register state is correlated with the previous register state (thus the first $n-1$ inputs) *and* the n-th input, as is must be at this stage, and the bath has become correlated with the n-th input and the previous register state.

We conclude this section by noting that each FSA transition is implicitly followed by a spontaneous "rethermalization" of $\mathcal{B}$ by the greater environment $\bar{\mathcal{B}}$. This process "resets" $\mathcal{B}$ to a thermal state before the next FSA transition, washing all information about the history of the FSA from the register's immediate surroundings into the greater environment. Since the "universe" $\mathcal{RSB}\bar{\mathcal{B}}$ is globally closed, this amounts to destruction of correlations between $\mathcal{RS}$ and $\mathcal{B}$ and creation of correlations between $\mathcal{RS}$ and $\bar{\mathcal{B}}$. This rethermalization of the FSA's immediate surroundings by a "greater" environment is a realistic process; it accommodates treatment of a *finite* local environment as an ordinary heat bath - with no memory of past interactions with the FSA - at the beginning of every state transition. (See [37] for further discussion of this heterogeneous environment model.)

4.2 Dissipation and Irreversibility in FSAs

We now state a fundamental dissipation bound for a physical FSA, defined as above and denoted $\mathcal{F}_P \triangleq \left\{ \mathcal{S}, \mathcal{R}, \{\hat{\sigma}^{\mathcal{S}}\}, \{\hat{x}^{\mathcal{R}}\}, \{\tilde{\mathcal{L}}\} \right\}$, which we prove using a "referential approach" to physical information theory. This approach has been used to obtain lower bounds on dissipation resulting from irreversible information loss in overwriting

[37], erasure [47], implementation of logical transformations [38, 49]), and instruction execution in a simple processor [45]. We preface statement and proof of this bound with a note on the essential feature that distinguishes FSAs from physical implementations of logical transformations in "L-machines" [38, 48]. In an L-machine, physical representations of the *input* and *output* of a logical transformation $\mathcal{L}$ are encoded in the initial and final states of the "machine," respectively. In a physical FSA, however, successive FSA states are internally encoded in the physical state of $\mathcal{S}$ whereas the inputs x_j are physically instantiated in an *external* referent system $\mathcal{R}$. These inputs influence transformation of the FSA state, selecting the transformation $\mathcal{L}_j$ that the FSA state will undergo on each step, but inputs need not ever be directly encoded in the physical state of the state register $\mathcal{S}$ in an FSA.

4.2.1 Dissipation Bound for FSAs in Steady State

Theorem-1: For physical FSA $\mathcal{F}_P = \left\{\mathcal{S}, \mathcal{R}, \{\hat{\sigma}^{\mathcal{S}}\}, \{\hat{x}^{\mathcal{R}}\}, \{\tilde{\mathcal{L}}\}\right\}$ and input pmf $\{q\}$, the input-averaged amount of energy dissipated to a thermal environment $\mathcal{B}$ on each state transition is lower bounded in steady state as

$$\Delta\langle E^{\mathcal{B}}\rangle \geq k_B T \ln(2) \sum_j q_j \left(\mathcal{I}^{\mathcal{R}_0\mathcal{S}} - \mathcal{I}_j^{\mathcal{R}_0\mathcal{S}'}\right) \tag{4.3}$$

where k_B is the Boltzmann constant, T is the environment temperature, and $\mathcal{I}^{\mathcal{R}_0\mathcal{S}} - \mathcal{I}_j^{\mathcal{R}_0\mathcal{S}'}$ is, for a state transition induced by input $\hat{x}_j^{\mathcal{R}_1}$, the reduction in quantum mutual information between the state of the register system $\mathcal{S}$ and the sequence of all past inputs physically instantiated in referent system $\mathcal{R}_0$. The bound is rigorously derived in [138] and is available in the Appendix B.1 here.

4.2.2 Discussion: Irreversibility and Information Loss in FSA

We have shown above that the average amount of energy dissipated into the bath per FSA transition is proportional to the quantity

$$-\langle \Delta \mathcal{I}_j^{\mathcal{R}_0 \mathcal{S}} \rangle = \sum_j q_j \left(\mathcal{I}^{\mathcal{R}_0 \mathcal{S}} - \mathcal{I}_j^{\mathcal{R}_0 \mathcal{S}'} \right)$$

i.e. to the input-averaged amount of physical information about $\mathcal{R}_0$ that is lost from $\mathcal{S}$ on each state transition. We argue below that this physical-information-theoretic quantity can be heuristically interpreted as the average amount of information that the machine state loses about its own history on each transition, and propose it as a measure irreversibility for irreducible FSA. This discussion is adapted from the author's paper on this very topic [43].

Prior to the state transition induced by the n-th input, the entire history of a deterministic FSA's evolution is reflected in the first $n-1$ inputs and the initial machine state. Since the first $n-1$ inputs are taken to be physically encoded in $\mathcal{R}_0$, and since the FSA state is encoded in $\mathcal{S}$, it is natural to associate the mutual information (or correlation entropy) $\mathcal{I}^{\mathcal{R}_0 \mathcal{S}}$ with the "amount of information the FSA state holds about its own dynamical history."

In the n-th FSA state transition, transformation of the physical state of $\mathcal{S}$ - and its correlation with the (fixed) physical state of $\mathcal{R}_0$ - depends on the n-th input via the input-selective application of the mapping $\tilde{\mathcal{L}}_j$. If the state mapping induced by an input $\hat{x}_j^{\mathcal{R}_1}$ is bijective, then there is no loss of correlation between $\mathcal{R}_0$ and $\mathcal{S}$ and $\mathcal{I}^{\mathcal{R}_0 \mathcal{S}} = \mathcal{I}_j^{\mathcal{R}_0 \mathcal{S}'}$. If, however, the state mapping induced by an input $\hat{x}_j^{\mathcal{R}_1}$ is non-injective, then $\mathcal{I}^{\mathcal{R}_0 \mathcal{S}} > \mathcal{I}_j^{\mathcal{R}_0 \mathcal{S}'}$.

It follows that for reversible FSAs - where, by definition, all inputs induce bijective state transformations - the input-averaged reduction in mutual information is

$$-\langle \Delta \mathcal{I}_j^{\mathcal{R}_0 \mathcal{S}} \rangle = 0$$

for any distribution $\{q\}$ of input probabilities. It also follows that for irreversible FSAs - where, by definition, at least one input induces a non-injective state transformation - we have

$$-\langle \Delta \mathcal{I}_j^{\mathcal{R}_0 \mathcal{S}} \rangle > 0$$

which is upper bounded at $H(\{\pi\}) = -\sum_k \pi_k \log_2 \pi_k$ – the Shannon entropy associated with the state-occupation probabilities at steady-state – for FSAs that "forget their entire history" on each state transition.

One can further show that

$$-\langle \Delta \mathcal{I}_j^{\mathcal{R}_0 \mathcal{S}} \rangle = \sum_j q_j H_j(S^{(n-1)}|S^{(n)})$$

where $H_j(S^{(n-1)}|S^{(n)})$ is the classical conditional Shannon entropy of the $(n-1)$-th state (represented by the random variable $S^{(n-1)} = \{k^{(n-1)}, \pi_k^{(n-1)}\}$) -*conditioned on* specification of the final state $S^{(n)}$ - for transitions induced by input x_j. This quantity, which can heuristically be interpreted as the statistical uncertainty in the initial state given the final state (averaged over final states, and for the j-th input), is upper bounded in steady state by the Shannon entropy $H(\{\pi\})$ of the steady-state occupation probabilities $\{\pi\}$. The input-averaged information loss per step is thus bounded in steady state as

$$0 \leq -\langle \Delta \mathcal{I}_j^{\mathcal{R}_0 \mathcal{S}} \rangle \leq H(\{\pi\})$$

with equality in the lower bound for reversible FSAs and equality in the upper bound for FSAs that irreversibly lose all information about prior inputs on each state transition.

These considerations recommend the quantity $-\langle \Delta \mathcal{I}_j^{\mathcal{R}_0 \mathcal{S}} \rangle$ as a quantitative measure of FSA irreversibility, and support its heuristic interpretation as the average per-step loss of information from the the FSA state about its own history. It can be evaluated directly as a classical[2] conditional entropy defined for an abstract FSA $\mathcal{F}_A$

[2]Note that the "informationally classical" nature of the irreversibility measure and corresponding dissipation bound do not in any way require that the physical machine states are themselves classical.

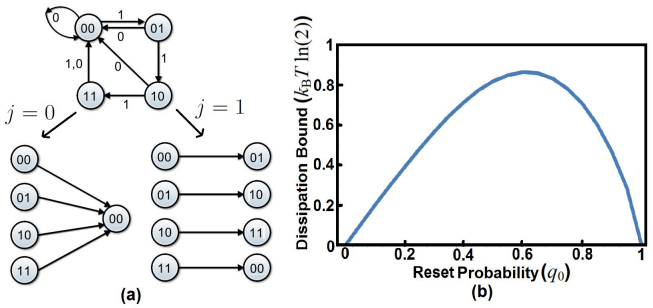

Figure 4.3. (a) State diagram of a simple four-state, two-input up counter FSA (top) that resets for input 0 and increments for input 1, together with the associated input-specific state mappings (bottom). (b) Lower bound on the average per-step amount of energy dissipated into the FSA's local environment as a function of the reset probability q_0. This bound reflects the component of dissipation resulting solely from irreversible information loss.

with input pmf $\{q\}$, and used with (B.1) to lower bound the energy dissipation per step for any physical FSA $\mathcal{F}_P$ that realizes $\mathcal{F}_A$ and is driven by an input source with these same statistics.

4.2.3 Illustrative Examples

We now illustrate application of our FSA dissipation bound to a simple example FSA. We consider a 2-bit binary up-counter with reset, which has four states and two inputs. On each step, the counter increments if the input is $x_1 = 1$ and resets if the input is $x_0 = 0$. The state diagram for this FSA is shown in Fig. (4.3), together with the individual input state mappings $\mathcal{L}_1$ and $\mathcal{L}_0$ implemented by the FSA for inputs x_1 and x_0.

This example, albeit simple, provides a good illustration of the input dependence of reversibility and dissipation in FSA, as up counting is reversible ($\mathcal{L}_1$ is bijective) whereas resetting is irreversible ($\mathcal{L}_0$ is non-injective). Since there is no information

---

We have taken the states of $\mathcal{S}$ to be generally quantum states that – together with the quantum dynamics governing state transitions – satisfy the stated requirements for faithful realization of $\mathcal{F}_A$ by $\mathcal{F}_P$.

loss for $x_j = x_1 = 1$ and total information loss for $x_j = x_0 = 0$, we have $\mathcal{I}_1^{\mathcal{R}_0\mathcal{S}'} = \mathcal{I}^{\mathcal{R}_0\mathcal{S}}$ and $\mathcal{I}_0^{\mathcal{R}_0\mathcal{S}'} = 0$) and the computational irreversibility is

$$-\langle \Delta\mathcal{I}_j^{\mathcal{R}_0\mathcal{S}} \rangle = \sum_j q_j \left(\mathcal{I}^{\mathcal{R}_0\mathcal{S}} - \mathcal{I}_j^{\mathcal{R}_0\mathcal{S}'} \right) = q_0 \mathcal{I}^{\mathcal{R}_0\mathcal{S}} = q_0 H(\{\pi\}).$$

where $\{\pi\}$ is the distribution of states in steady state. The corresponding lower bound on the average amount of energy dissipated into the bath on each step is simply

$$\Delta\langle E^{\mathcal{B}} \rangle \geq k_B T \ln(2) q_0 H(\{\pi\}).$$

This dissipation bound is plotted as a function of the reset probability q_0 in Fig. 4.3(b). In the small q_0 limit, where the FSA is reversibly cycling through its four states, the lower bound approaches zero. Dissipation increases with increasing q_0, as the FSA state evolution becomes random and the probability of information loss from reset increases, up to $q_0 \approx 0.6$. The bound decreases with further increase in q_0, as skewing of the state occupation probabilities toward $\pi_0 = 1$ reduces $H(\{\pi\})$, until the bound finally vanishes at $q_0 = 1$ where the counter is reset in every step (and $H(\{\pi\} = 0)$. In this limit there is no fundamental minimum dissipation even though the reset mapping is irreversible, since the FSA remains frozen in the reset state and there is no nontrivial history about which the FSA state can lose information on any step.

4.3 Dissipation in Moore Machine

Moore machines are a type of deterministic FSA in which the output is a function of the current state only, as shown in Fig. (4.4). This is the major difference with Mealy machines, where the output of the FSA depends upon the current state of the machine and the input (and will be discussed in the next section). The state transition

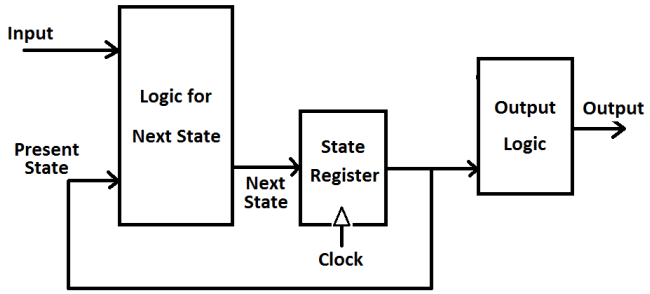

Figure 4.4. General block diagram of a Moore machine. The output state is only dependent on the current state of the FSA.

in a FSA is synchronous i.e associated with a time event e.g. a rising clock edge. Hence the output in a Moore machine changes synchronously with the state of the FSA. We will be dealing with irreducible Moore machines only. Like in the previous section the abstract Moore machine is defined by $\mathcal{F}_M^A \triangleq \{\{\sigma\}, \{x\}, \{\mathcal{L}\}, \{\omega\}, \{\mathcal{J}\}\}$ consists of a finite set $\{\sigma\}$ and $\{\omega\}$ of automaton and output states respectively, a set $\{x\}$ of input symbols that induce transitions between the automata states, a set $\{\mathcal{L}\}$ of deterministic transition rules that govern the automata state transitions and a set $\{\mathcal{J}\}$ of deterministic transition rules that map the automaton states to their respective outputs.

4.3.1 Description of Physical Moore FSA

In this section we will provide the physical description of the abstract Moore finite state automaton $\mathcal{F}_A$. We will formalize the physical realization of it's states, inputs, outputs and state transition. This description will be very similar to the description of the deterministic FSA from the previous section.

Automata and Output States: Since the output is only dependent on the current state of the system, the k-th automata state σ_k and it's corresponding output state $\omega_k = \mathcal{J}(\sigma_k)$ of the abstract Moore machine can be recast as a larger irreducible FSA without any outputs from the previous section in which the k-th state of this new FSA is given

as $\delta_k = \{\sigma_k, \omega_k\}$. Let the state of this larger FSA be instantiated in distinguishable orthogonal states $\{\hat{\delta}_k^{\mathcal{S}_M}\}$ of generally quantum mechanical register system $\mathcal{S}_M$. The state probability of δ_k is equal to the state probability of σ_k. The system interacts with a local bath $\mathcal{B}$, a finite heat bath in a thermal state $\hat{\rho}^{\mathcal{B}}$ at temperature T at the start of a cycle.

Inputs: Random length-*(n+1)* strings $\vec{X}$ are once again physically implemented in the state of a *(n+1)*-partite referent system $\mathcal{R} = \mathcal{R}_0\mathcal{R}_1\mathcal{R}_2$. Subsets $\vec{X}_k = X^{(0)}X^{(1)}...X^{(n-1)}$ of strings that map to machine state $\sigma_k^{\mathcal{S}}$ are grouped into the mixed states $\hat{\rho}_k^{\mathcal{R}_0}$ of the referent $\mathcal{R}_0$, and the new incoming inputs $X^{(n)}$ and $X^{(n+1)}$ are instantiated as a mixture of distinguishable pure states $\hat{x}_j^{\mathcal{R}_1}$ of $\mathcal{R}_1$, and $\hat{x}_j^{\mathcal{R}_2}$ of $\mathcal{R}_2$ respectively.

State Transition and Output Generation: The n-th cycle is realized by the dynamical evolution of the state of $\mathcal{S}_M$ interacting with the heat bath. Global evolution of $\mathcal{R}\mathcal{S}_M\mathcal{B}$ producing the state transition and output generation is assumed to be governed by the time-dependent Schrodinger equation to ensure consistency with physical law. A different evolution operator is required for the synchronous and asynchronous transformations.

This physical universe is completed with the local bath $\mathcal{B}$ being embedded in a larger environment $\bar{\mathcal{B}}$. After every evolution, $\mathcal{B}$ is driven from equilibrium due to its interactions with $\mathcal{S}_M$, and the larger environment $\bar{\mathcal{B}}$ rethermalizes $\mathcal{B}$. This resets the local bath back to the thermal state before another transition, removing all the information about the referents present in the the bath, into the greater environment. The rethermalization process allows us to treat the local heat bath as an ordinary bath with no prior information about the FSA or the output register, at the start of every operation.

4.3.2 Dissipation Bound for Moore Machines

Theorem-2: For an abstract Moore machine $\mathcal{F}_M^A \triangleq \{\{\sigma\},\{x\},\{\mathcal{L}\},\{\omega\},\{\mathcal{J}\}\}$, implemented as a physical FSA $\mathcal{F}_M = \left\{\mathcal{S}_M, \mathcal{R}, \{\hat{\delta}^{\mathcal{S}_M}\}, \{\hat{x}^{\mathcal{R}}\}, \{\tilde{\mathcal{L}_M}\}\right\}$ and input pmf $\{q\}$, the input-averaged amount of energy dissipated to a thermal environment $\mathcal{B}$ on each state transition is lower bounded in steady state as

$$\Delta\langle E^{\mathcal{B}}\rangle \geq k_B T \ln(2) \sum_j q_j \left(\mathcal{I}^{\mathcal{R}_0\mathcal{S}_M} - \mathcal{I}_j^{\mathcal{R}_0\mathcal{S}'_M}\right) \tag{4.4}$$

where k_B is the Boltzmann constant, T is the environment temperature, and $\mathcal{I}^{\mathcal{R}_0\mathcal{S}_M} - \mathcal{I}_j^{\mathcal{R}_0\mathcal{S}'_M}$ is, for a state transition induced by input $\hat{x}_j^{\mathcal{R}_1}$, the reduction in quantum mutual information between the state of the automata plus output system $\mathcal{S}_M$, and the sequence of all past inputs physically instantiated in referent system $\mathcal{R}_0$. From the previous section, we know that this bound of an irreducible FSA with no outputs is only dependent on the steady state probability distribution of the automaton states. Since the state distributions of the original Moore machine with output and this larger FSA without the outputs are the same, the lower bound on dissipation for the Moore machine with an output is simply equal to the lower bound on dissipation of the same FSA without the output. We will demonstrate with a simple example in the next subsection.

4.3.3 Dissipation Bound for Moore Machine with Separate Output Register

Earlier in this section, we analyzed the lower bound on dissipation in Moore machines in which the output state was absorbed into the automaton state to form new state machine. Since the outputs are uniquely determined by each FSA state, this new FSA has the same number of states as the original FSA and will have no lower bound on dissipation for generating the Moore machine output. The lower bound on dissipation for this realization of the Moore machine is identical to a FSA

without an output. However it might be the case that the output is realized as a separate register system. The lower bound on dissipation in this scenario would be the sum of the lower bound of the outputless FSA in steady state plus the cost of generating the output at every clock signal (once per every state transition).

From section 4.2.2, the lower bound on dissipation for a physical FSA in steady state is given by

$$\Delta\langle E^{\mathcal{B}}_{FSA}\rangle \geq k_B T \ln(2) \sum_j q_j \left(\mathcal{I}^{\mathcal{R}_0\mathcal{S}} - \mathcal{I}^{\mathcal{R}_0\mathcal{S}'}_j\right) \tag{4.5}$$

We assume that the system generating the output using the FSA states as inputs can be modeled as a $\mathcal{L}$-machine from the previous chapter section 3.4, where the initial input states are perfectly correlated to the FSA automata states and the final states are the necessary outputs. The lower bound on dissipation for output generation instantiated in system $\mathcal{M}$ (which is in contact with it's thermal bath at temperature T) is

$$\Delta\langle E^{\mathcal{B}_\mathcal{M}}_{\mathcal{M}}\rangle \geq -k_B T \ln(2) \Delta S^{\mathcal{M}} \tag{4.6}$$

where $\Delta\langle E^{\mathcal{B}_\mathcal{M}}_{\mathcal{M}}\rangle$ is the change in average energy of bath $\mathcal{B}_\mathcal{M}$ during the logical transformation. $-\Delta S^{\mathcal{M}}$ is the loss in von Neumann entropy of the system over the transformation that generates the output.

Thus the total lower bound on dissipation for a Moore machine which generates the output in a separate system is given by the sum of the individual lower bounds. We have

$$\Delta\langle E^{\mathcal{B}}_{Total}\rangle = \Delta\langle E^{\mathcal{B}}_{FSA}\rangle + \Delta\langle E^{\mathcal{B}_\mathcal{M}}_{\mathcal{M}}\rangle \tag{4.7}$$

In the next subsection, this lower bound on dissipation is compared to the lower bound obtained for output generation in a Moore machine in the previous subsection using the example FSA from Fig. 4.6.

4.3.4 Illustrative Example

Consider the 2-bit binary up-counter with reset from the previous section. This FSA has four automata states, two inputs and 2 outputs. As before on each step, the counter increments if the input is $x_1 = 1$ and resets if the input is $x_0 = 0$. The output is 0 for the FSA states 00,01 and 10 and the output is 1 for the state 11. The state diagram for this Moore machine is shown in Fig. 4.6(a) and the corresponding larger FSA in which the outputs are combined as part of the FSA in Fig. 4.6(b), together with the individual input state mappings $\mathcal{L}_1$ and $\mathcal{L}_0$ implemented by this FSA for inputs x_1 and x_0.

The dissipation bound is plotted as a function of the reset probability q_0 in Fig. 4.6(c). The bound is exactly identical to the bound for the counter FSA without the outputs from the previous section. This indicates that there is no additional minimum cost associated with generating outputs in a Moore machine, since the outputs are only dependent on the FSA states. However this result is based on our ability to incorporate the output into the state of the larger FSA. We will leave the case of generating the outputs using a separate output register system for the final defense. The dissipation associated with generating outputs using the automaton state as inputs under a $\mathcal{L}$-machine picture would add an additional cost on top of the minimum dissipation of the FSA state transition.

The dissipation bound for the Moore FSA in Fig. 4.6(a) with a separate output generation process is shown in Fig. 4.3.4. We have plotted the dissipation cost of the FSA in steady state (which is equal to the dissipation cost of the Moore machine in which the output is combined with the FSA state) plus a cost of generating a output

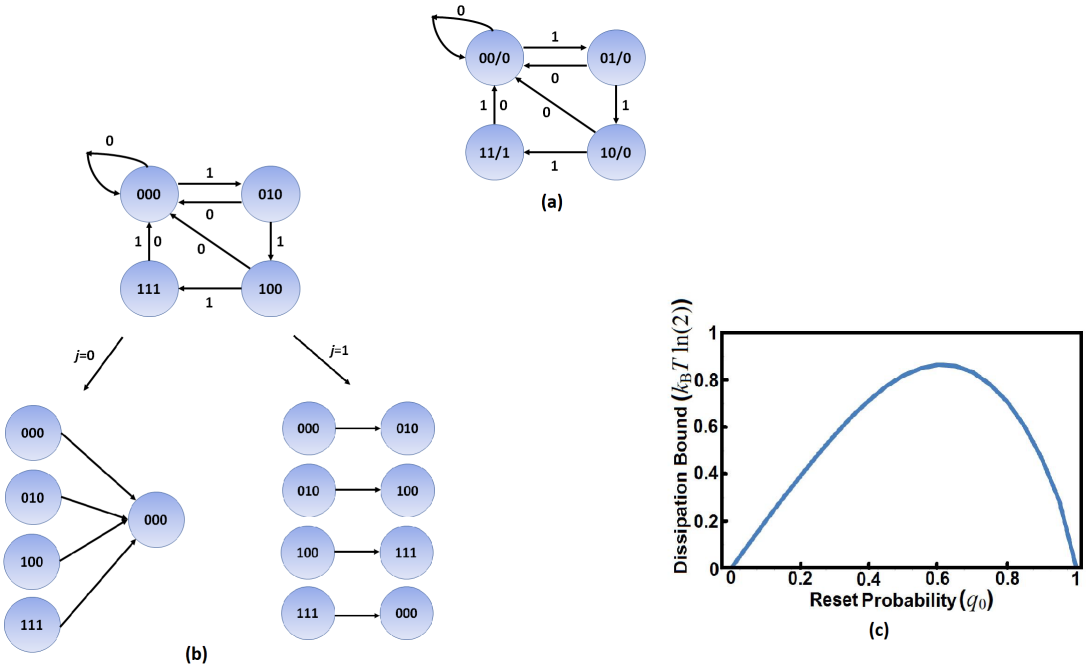

Figure 4.5. (a) State diagram of a simple four-state, two-input up counter Moore FSA with 2 outputs - 0 for states $00, 01, 10$ and 1 for the state 11. (b) The equivalent FSA of the Moore FSA from (a), in which the outputs have been incorporated into the FSA state. (c) Lower bound on the average per-step amount of energy dissipated into the FSA's local environment as a function of the reset probability q_0. This bound identical to the bound of the irreducible FSA without outputs from the previous section.

as a $\mathcal{L}$-machine transformation - all as a function of reset probability (q_0). We can clearly see the excess dissipation spent in generating the output separately.

4.4 Dissipation in Mealy machines

Mealy machines are a type of deterministic FSA in which the next automata state and output are both functions of the current state and the latest input, as shown in Fig. (4.7). This is the major difference with Moore machines, where the output of the FSA depends only upon the current state of the machine. The state transition in a FSA is synchronous i.e associated with a time event like a rising clock edge. Hence the output in a Moore machine changes synchronously with the state of the FSA,

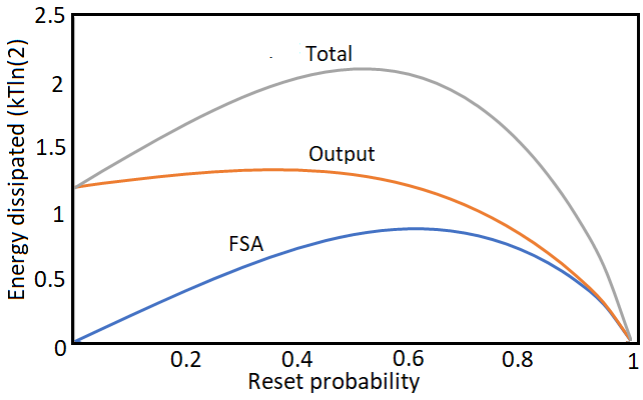

Figure 4.6. Dissipation bound for the Moore FSA with a separate output register for the FSA in Fig. 4.6(a). In the figure, we have the dissipation bound associated with the steady dissipation in the FSA, the bound for the output generation and the bound on the total dissipation which is the sum of the previous two terms.

while in a Mealy machine, it can also change asynchronously with input change. In this section, we will deal with irreducible Mealy machines.

We will define our *Mealy machine cycle* as the period between two successive automata state transitions. In this paper we will calculate the lower bound on dissipation associated with one such cycle. The heat dissipation arising from the irreversible information loss comes from the automata state transition and output generation that occurs at the start of a cycle. Before the completion of the cycle, multiple new inputs can arrive. These new inputs will generate a new output every time, which adds to the heat dissipation. For the purposes of this paper, we will allow for one new input to arrive within the cycle.

4.4.1 Description of Physical Mealy FSA

In this section we will provide the physical description of the abstract Mealy finite state automata $\mathcal{F}_A$. We will formalize the physical realization of this FSA's states and outputs, inputs and operations.

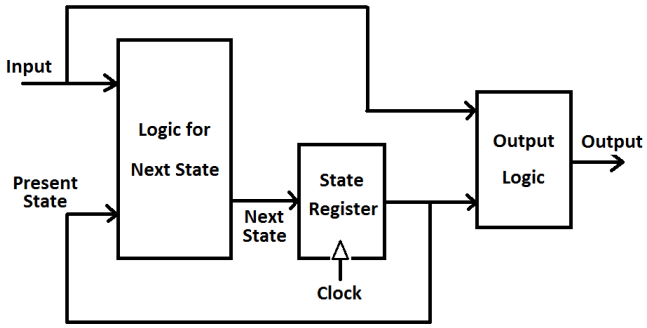

Figure 4.7. Conventional block diagram of a Mealy machine. Both the next state of the automata and the output are functions of the current state and the latest input in the Mealy machine. The state transition is synchronous and depends on the clock signal. However the output change is asynchronous and can occur whenever the input or the state changes.

Automata and Output States: Abstract FSA and output states are physically instantiated in distinguishable orthogonal states $\hat{\sigma}_k^{\mathcal{S}}$ and $\hat{\varrho}_l^{\mathcal{O}}$, of generally quantum mechanical register systems $\mathcal{S}$ and $\mathcal{O}$ respectively. Both systems interact with the local bath $\mathcal{B}$. $\mathcal{B}$ is a finite heat bath in a thermal state $\hat{\rho}^{\mathcal{B}}$ at temperature T at the start of a cycle.

Inputs: Random length-*(n+1)* strings $\vec{X}$ that are physically implemented in the state of a *(n+1)*-partite referent system $\mathcal{R} = \mathcal{R}_0\mathcal{R}_1\mathcal{R}_2$. Subsets $\vec{X}_k = X^{(0)}X^{(1)}...X^{(n-1)}$ of strings that map to machine state $\sigma_k^{\mathcal{S}}$ are grouped into the mixed states $\hat{\rho}_k^{\mathcal{R}_0}$ of the referent $\mathcal{R}_0$, and the new incoming inputs $X^{(n)}$ and $X^{(n+1)}$ are instantiated as a mixture of distinguishable pure states $\hat{x}_j^{\mathcal{R}_1}$ of $\mathcal{R}_1$, and $\hat{x}_j^{\mathcal{R}_2}$ of $\mathcal{R}_2$ respectively.

State Transition and Output Generation: The n-th cycle is realized by the dynamical evolution of the state of $\mathcal{S}$ and $\mathcal{O}$, interacting with their heat baths. Global evolution of $\mathcal{ROSB}$ producing the state transition and output generation is assumed to be governed by the time-dependent Schrodinger equation to ensure consistency with physical law. A different evolution operator is required for the synchronous and asynchronous transformations.

This physical universe is completed with the local bath $\mathcal{B}$ being embedded in a larger environment $\bar{\mathcal{B}}$ [38]. After every evolution, $\mathcal{B}$ is driven from equilibrium due to its interactions with $\mathcal{S}$ and $\mathcal{O}$, and the larger environment $\bar{\mathcal{B}}$ rethermalizes $\mathcal{B}$. This resets the local bath back to the thermal state before another transition, removing all the information about the referents present in the the bath, into the greater environment. The rethermalization process allows us to treat the local heat bath as an ordinary bath with no prior information about the FSA or the output register, at the start of every operation.

The n-th *Mealy machine cycle* is shown in Fig. 4.8. Initially, the current FSA state encoded in the physical state of $\mathcal{S}$ is correlated with $\mathcal{R}_0$, the first *(n-1)* inputs only. The output register state is encoded in the state of $\mathcal{O}$ and depends on the current state of $\mathcal{S}$ and $\mathcal{R}_1$, the latest input which arrived prior to the start of the n-th cycle. The dynamical evolution of $\mathcal{R}_1\mathcal{O}\mathcal{S}\mathcal{B}$ will leave the system $\mathcal{S}$ in the next FSA state, and the output system $\mathcal{O}$ with the new output. This will weaken the correlation between $\mathcal{O}\mathcal{S}$ and $\mathcal{R}_0$, which means irreversible information loss and would necessarily dissipate heat. The new state of $\mathcal{O}$ will be correlated to both the new state of $\mathcal{S}$ and input $\mathcal{R}_1$. During the course of the cycle, the next input $\mathcal{R}_2$ will not induce a change in the state of $\mathcal{S}$, but the state of $\mathcal{O}$ will be conditionally overwritten. We will quantify all the information loss that occurs over the course of this cycle and show that it results in dissipation.

Initial States: At the start of the n-th cycle, the state of $\mathcal{R}\mathcal{O}\mathcal{S}\mathcal{B}$ is described within the density operator formalism as

$$\begin{aligned}\hat{\rho}^{\mathcal{ROSB}} &= \sum_j q_j \left(\hat{x}_j^{\mathcal{R}_1} \otimes \sum_k \pi_k^{(n-1)} \left[\hat{\rho}_k^{\mathcal{R}_0} \otimes \hat{\sigma}_k^{\mathcal{S}} \otimes \hat{\varrho}_{kj}^{\mathcal{O}} \right] \right) \\ &\qquad \otimes \hat{\rho}^{\mathcal{R}_2} \otimes \hat{\rho}_{th}^{\mathcal{B}} \\ &= \sum_j \sum_k q_j \pi_k^{(n-1)} \hat{\rho}_{kj}^{\mathcal{ROSB}}. \end{aligned} \tag{4.8}$$

With $\hat{\rho}_{kj}^{\mathcal{ROSB}} = \hat{\rho}_k^{\mathcal{R}_0} \otimes \hat{\sigma}_k^{\mathcal{S}} \otimes \hat{\varrho}_{kj}^{\mathcal{O}} \otimes \hat{x}_j^{\mathcal{R}_1} \otimes \hat{\rho}^{\mathcal{R}_2} \otimes \hat{\rho}_{th}^{\mathcal{B}}$ and $\hat{\varrho}_{kj}^{\mathcal{O}} = \mathcal{V}_j\{\hat{\sigma}_k^{\mathcal{S}}\}$, the output produced by state $\hat{\sigma}_k^{\mathcal{S}}$ and input $\hat{x}_j^{\mathcal{R}_1}$. [3]

Intermediate States: The n-th synchronous state transition and output generation is a unitary quantum evolution

$\hat{\rho}^{\mathcal{ROSB}'} = \hat{U}_1 \hat{\rho}^{\mathcal{ROSB}} \hat{U}_1^\dagger$ of the state of the system $\mathcal{ROSB}$ involving interactions between $\mathcal{R}_1\mathcal{OS}$ and $\mathcal{B}$ given as $\hat{U}_1 = \hat{I}^{\mathcal{R}_0\mathcal{R}_2} \otimes \hat{U}_1^{\mathcal{R}_1\mathcal{OSB}}$. To implement the abstract FSA $\mathcal{F}_A$ faithfully, $\hat{U}_1$ should have the property that $\hat{\rho}_{kj}^{\mathcal{OS}'} = Tr_{\mathcal{RB}}[\hat{U}_1 \hat{\rho}_{kj}^{\mathcal{ROSB}} \hat{U}_1^\dagger] = \bar{\mathcal{L}}_j\{\hat{\sigma}_k^{\mathcal{S}}\} \otimes \bar{\mathcal{V}}_j \left\{ \bar{\mathcal{L}}_j\{\hat{\sigma}_k^{\mathcal{S}}\} \right\} \in \{\hat{\sigma}^{\mathcal{S}} \otimes \hat{\varrho}^{\mathcal{O}}\}$.

Also $\bar{\mathcal{L}}_j \in \{\bar{\mathcal{L}}\}$ and $\bar{\mathcal{V}}_j \in \{\bar{\mathcal{V}}\}$ are local non-unitary superperators that act on $\mathcal{S}$ and $\mathcal{O}$ respectively to induce state transitions and output generation.

At the end of the unitary evolution, the state of $\mathcal{ROSB}$ is

$$\begin{aligned}\hat{\rho}^{\mathcal{ROSB}'} &= \sum_j q_j \left(\hat{x}_j^{\mathcal{R}_1} \otimes \pi_k^{(n-1)} \left[\hat{\rho}_k^{\mathcal{R}_0} \otimes \hat{\rho}_{kj}^{\mathcal{OSB}'} \right] \right) \otimes \hat{\rho}^{\mathcal{R}_2} \\ &= \sum_j \sum_k q_j \pi_k^{(n-1)} \hat{\rho}_{kj}^{\mathcal{ROSB}'} \end{aligned} \tag{4.9}$$

where

$$\hat{\rho}_{kj}^{\mathcal{ROSB}'} = \hat{\rho}_k^{\mathcal{R}_0} \otimes \bar{\mathcal{L}}_j\{\hat{\sigma}_k^{\mathcal{S}}\} \otimes \bar{\mathcal{V}}_j \left\{ \bar{\mathcal{L}}_j\{\hat{\sigma}_k^{\mathcal{S}}\} \right\} \otimes \hat{x}_j^{\mathcal{R}_1} \otimes \hat{\rho}_{kj}^{\mathcal{B}'} \otimes \hat{\rho}^{\mathcal{R}_2}.$$

The bath $\mathcal{B}$ driven from equilibrium is rethermalized by $\bar{\mathcal{B}}$ before the start of the next unitary evolution.

---

[3] $\hat{\rho}_k^{\mathcal{R}_0}$, $\hat{\sigma}_k^{\mathcal{S}}$, etc. are density operators i.e., positive operators with a unit trace defined on the (complex Hilbert) state spaces of $\mathcal{R}$ and $\mathcal{S}$ and respectively. $\otimes$ denotes the tensor (or Kronecker) product.

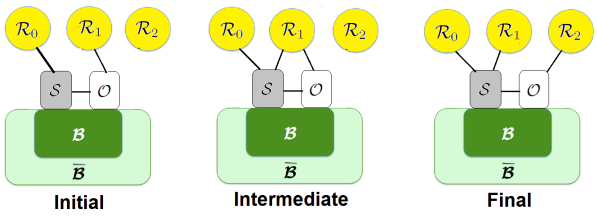

Figure 4.8. Physical description of a *Mealy machine cycle*- It begins with a physical representation of the joint output-state register $\mathcal{OS}$ interacting with the local bath $\mathcal{B}$ (Initial), $\mathcal{S}$ is correlated to $\mathcal{R}_0$ and, $\mathcal{O}$ is correlated to $\mathcal{S}$ and new input, instantiated in referent $\mathcal{R}_1$. The state of $\mathcal{OS}$ is transformed at the start of the cycle where both the state of $\mathcal{S}$ and $\mathcal{O}$ changes by interaction with $\mathcal{R}_1$ and $\mathcal{B}$ (Intermediate). After the larger environment $\bar{\mathcal{B}}$ then rethermalizes $\mathcal{B}$, the system interacts with referent $\mathcal{R}_2$ to produce a new output, with the state of $\mathcal{S}$ unchanged ($\mathcal{O}$ loses correlation with $\mathcal{R}_1$ but $\mathcal{S}$ does not).

Final States: The joint system $\mathcal{ROSB}$ evolves unitarily to generate the new output with no change in the system state $\mathcal{S}$. The final state [4] of the composite is given by

$$\hat{\rho}^{\mathcal{ROSB}''} = \sum_{j'} q_{j'} \left(\hat{x}_{j'}^{\mathcal{R}_2} \otimes \hat{\rho}_{j'}^{\mathcal{R}_0\mathcal{R}_1\mathcal{OSB}''} \right)$$

where

$$\hat{\rho}_{j'}^{\mathcal{R}_0\mathcal{R}_1\mathcal{OSB}''} = \sum_j q_j \left[\hat{x}_j^{\mathcal{R}_1} \otimes \sum_k \pi_k^{(n-1)} \left(\hat{\rho}_k^{\mathcal{R}_0} \otimes \bar{\mathcal{L}}_j\{\hat{\sigma}_k^{\mathcal{S}}\} \otimes \bar{\mathcal{V}}_{j'} \left\{ \bar{\mathcal{L}}_j\{\hat{\sigma}_k^{\mathcal{S}}\} \right\} \right) \right].$$

This asynchronous generation of the new output is given by the unitary $\hat{U}_2 = \sum_{j'} \hat{x}_{j'}^{\mathcal{R}_2} \otimes \hat{I}^{\mathcal{R}_0\mathcal{R}_1} \otimes \hat{U}_{j'}^{\mathcal{OSB}}$ where $\hat{U}_{j'}^{\mathcal{SB}}$ has the property $Tr_{\mathcal{B}}[\hat{U}_{j'}^{\mathcal{SB}} \hat{\rho}^{\mathcal{OSB}'} \hat{U}_{j'}^{\mathcal{SB}\dagger}] = \hat{\rho}_{j'}^{\mathcal{OS}''}$ and

$$\hat{\rho}_{j'}^{\mathcal{OS}''} = \sum_{j,k} q_j \pi_k^{(n-1)} \left[\bar{\mathcal{L}}_j\{\sigma_k^{\mathcal{S}}\} \otimes \bar{\mathcal{V}}_{j'} \left\{ \bar{\mathcal{L}}_j\{\sigma_k^{\mathcal{S}}\} \right\} \right].$$

[4] The intermediate $\hat{\rho}'$ and final $\hat{rho}''$ density operators both characterize the states of the same system $\mathcal{ROSB}$ over the course of the Mealy cycle

The final states are such that $\mathcal{S}$ is correlated to the first n inputs, and $\mathcal{O}$ is to the $(n+1)$-th input and the first n-inputs through $\mathcal{S}$. The bath $\mathcal{B}$ is once again rethermalized by the larger environment $\bar{\mathcal{B}}$, before the start of the next cycle.

4.4.2 Dissipation Bound for Mealy machine over one Cycle

Using the referential approach as before, we will now derive the fundamental dissipation bound for a physical FSA-Mealy machine with an output register over one cycle. In this case, the states of the automata and the output are encoded in the states of $\mathcal{S}$ and $\mathcal{O}$, and the inputs x_j are instantiated in the referent system $\mathcal{R}$. The input will select the transformation that $\mathcal{S}$ and $\mathcal{O}$ will undergo over the course of the cycle.

Theorem-3 For physical FSA $\mathcal{F}_P = \{\mathcal{S}, \mathcal{O}, \mathcal{R}, \{\hat{\sigma}^{\mathcal{S}}\}, \{\hat{\varrho}^{\mathcal{O}}\}, \{\hat{x}^{\mathcal{R}}\}, \{\bar{\mathcal{L}}\}, \{\bar{\mathcal{V}}\}\}$ and input pmf $\{q\}$, the input averaged amount of energy dissipated to a thermal environment $\mathcal{B}$ over one *Mealy machine cycle* is lower bounded in steady state as

$$\Delta\langle E^{\mathcal{B}}\rangle_{cycle} \geq k_B T \ln(2) \left(\sum_j q_j \left[I_j^{\mathcal{R}_0 \mathcal{O}\mathcal{S}} - I_j^{\mathcal{R}_0 \mathcal{O}\mathcal{S}'} \right] + \mathcal{I}^{\mathcal{R}_1 \mathcal{O}\mathcal{S}'} + \sum_j q_j S(\hat{\rho}_j^{\mathcal{O}\mathcal{S}'}) - \sum_{j'} q_{j'} S(\hat{\rho}_{j'}^{\mathcal{O}\mathcal{S}''}) \right) \tag{4.10}$$

where k_B is the Boltzmann constant and T is the environment temperature. $I_j^{\mathcal{R}_0 \mathcal{O}\mathcal{S}} - I_j^{\mathcal{R}_0 \mathcal{O}\mathcal{S}'}$ is, for a state transition induced by input $\hat{x}_j^{\mathcal{R}_1}$, the reduction in quantum mutual information between the joint state of the output and automata register $\mathcal{O}\mathcal{S}$ and the sequence of all past inputs physically instantiated in referent system $\mathcal{R}_0$. $\mathcal{I}^{\mathcal{R}_1 \mathcal{O}\mathcal{S}'}$ is amount of quantum mutual information between the joint state $\mathcal{O}\mathcal{S}$ at the intermediate state and the referent $\mathcal{R}_1$. $S(\hat{\rho}_j^{\mathcal{O}\mathcal{S}'})$ and $S(\hat{\rho}_{j'}^{\mathcal{O}\mathcal{S}''})$ are the self entropies of the states of $\mathcal{O}\mathcal{S}$ associated with the j-th input of $\mathcal{R}_1$ and the j'-th input of $\mathcal{R}_2$, at

the intermdeiate and final stages respectively. The rigorous proof of the theorem-3 is provided in Appendix B.2.

4.4.3 Illustrative Example for Mealy Machines

We illustrate the application of the bound to a simple 2-bit binary up-counter with reset, The counter has four automata states, four outputs and two inputs 0 and 1. On each step, the counter increments if the input is $x_1 = 1$ and resets if the input is $x_1 = 0$. The output at each step is the value of the next state. The state diagram of this FSA is shown in the Fig. (4.9a). We will assume that the automata and output states are physically instantiated as orthogonal pure states.

The lower bound on the average amount of energy dissipated into the bath over every *cycle* reduces to

$$\Delta\langle E^{\mathcal{B}}\rangle_{cycle} \geq k_B T \ln(2) \{q_0 H(\{\pi\}) + H(\{\psi\}) - H(\{\omega\})\}$$

where $H(.)$ is the classical Shannon entropy function for a probability distribution. $\{\pi\}$ is the steady state distribution of the FSA, q_0 is the reset probability, the distributions $\{\omega\} = \{q_0, q_0 \cdot \pi_3, (1-q_0) \cdot \pi_0, (1-q_0) \cdot \pi_1, (1-q_0) \cdot \pi_2\}$ and $\{\psi\} = \{q_0 \cdot (1+\pi_3), (1-q_0) \cdot \pi_0, (1-q_0) \cdot \pi_1, (1-q_0) \cdot \pi_2\}$.

The dissipation bound is plotted as a function of the reset probability q_0 for the counter with and without outputs, in the Fig. (4.9b). For $q_0 = 0$, the bound reaches zero as the FSA reversibly recycles between the counter and output states. The dissipation increases with increasing q_0 as the state evolution becomes more random. The difference between the two cases illustrates the energy dissipation arising from the output generation. This then starts to decrease as the FSA automata and output states are increasingly present in the reset states, and vanishes for $q_0 = 1$ as the states hold no information about it's history to lose.

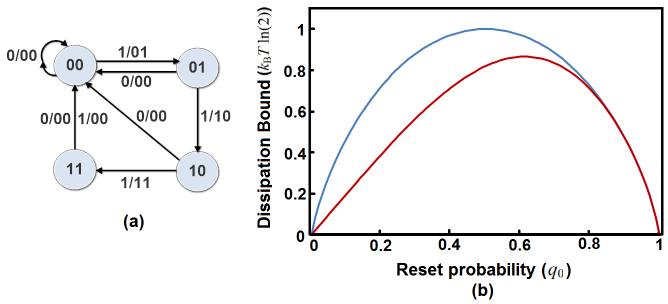

Figure 4.9. (a) State diagram of a simple 2-bit up-counter FSA with four automata states, four output states and two inputs. The counter resets for input 0 and increments for input 1. The output at each step is the new state that the system transitions to. (b) Lower bound on the average per-*cycle* amount of energy dissipated into the local heat bath by the FSA with outputs (blue) and without outputs (red) [138], as a function of the reset probability q_0. The difference between the two bounds illustrates the energy dissipation arising from the output generation.

4.5 Probabilistic Finite State Automata

In this section, a more general finite state automata called probabilistic finite state automata will be described and the fundamental lower bound on energy dissipation for this FSA will be derived. Below we will provide abstract and physical descriptions of probabilistic finite state automata, and use them to obtain the lower bound on dissipation.

4.5.1 Abstract Probabilistic Finite-State Automata

An abstract FSA $\mathcal{F}_A \triangleq \{\{\sigma\}, \{x\}, \{\mathcal{L}\}\}$ as discussed before consists of a finite set $\{\sigma\}$ of states, a set $\{x\}$ of input symbols that induce transitions between states, and a set $\{\mathcal{L}\}$ of transition rules - one for each input - that govern the state transitions. Specifically, for every input x_j there is an "input transition rule" $\mathcal{L}_j$ that can map every "current state" σ_k to multiple "next states" σ'_{kj} at certain probabilities. Outputs are generally defined for FSAs as well, but we do not consider them right now. An FSA is *probabilistic* if input transition rules $\mathcal{L}_j$ can assign more than one next state to every current state with a non-zero probability.

The statistical properties of FSAs driven by random sequences – including the amount of irreversible information loss – depend directly on the input statistics. For input sequences of identically distributed discrete random variables $X^{(n)} = \{q, x\}$ (each with the same symbol set $\{x\}$ and corresponding probability mass function $\{q\}$), the conditional probability that an FSA in state σ_k will transition to state $\sigma_{k'}$ on any given step is

$$p_{k\to k'} = \sum_{j\in\{j\}_{k\to k'}} \pi_{k'|(k,j)} q_j \qquad \Big(\{j\}_{k\to k'} = \{j|\mathcal{L}_j(\sigma_k) = \sigma_{k'}\}\Big).$$

where $\pi_{k'|(k,j)}$ is the probability that the k-th state maps to the k' state for the j-th input, and $\sum_k \pi_{k'|(k,j)} = 1$. We can now generate the statistical transition matrix P with elements $p_{k\to k'}$. As before P satisfies the Markov property, and the "steady-state" occupation probabilities for the FSA states can be obtained from P, either as an eigenvector with an eigenvalue 1, or using the relation $\lim_{n\to\infty} P^n = P_{ss}$.

4.5.2 Physical Probabilistic Finite-State Automata

We will now construct a *physical* description of a probabilistic FSA, defined abstractly as above. We formalize this description after identifying the physical realizations of FSA states, inputs, and transitions.

- **States:** The abstract FSA states σ_k are faithfully represented in distinguishable physical states $\hat{\sigma}_k^{\mathcal{S}}$ of a generally quantum-mechanical register system $\mathcal{S}$, which interacts with its local environment $\mathcal{B}$. Here $\mathcal{B}$ is taken to be a (finite) heat bath nominally in a thermal state $\hat{\rho}_{th}^{\mathcal{B}}$ at temperature T.

- **Inputs:** Random length-n input strings $\vec{X}$ are physically instantiated in the state of a n-partite "referent" system $\mathcal{R} = \mathcal{R}_0\mathcal{R}_1$. The i-th string of $\mathcal{R}_0$ instantiated as $\hat{\rho}_i^{\mathcal{R}_0}$, leads to the FSA state $\hat{\rho}_i^{\mathcal{S}}$ - consisting of distinguishable mixed

states $\hat{\sigma}_k^{\mathcal{S}}$ of $\mathcal{S}$. $X^{(n)}$ is represented by a mixture of pure distinguishable states $\hat{x}_j^{\mathcal{R}_1}$ of $\mathcal{R}_1$.

- **State Transitions:** The n-th state transition is realized by dynamical evolution of the state of $\mathcal{S}$, conditioned on the state of $\mathcal{R}_1$ (i.e. the n-th input) and in interaction with $\mathcal{B}$. Global evolution of the interacting composite $\mathcal{R}_1\mathcal{S}\mathcal{B}$ producing this transition is assumed to be governed by the time-dependent Schrodinger equation to ensure consistency with physical law (implying unitary evolution of the state of $\mathcal{R}_1\mathcal{S}\mathcal{B}$). The n-th input remains encoded in $\mathcal{R}_1$ at the conclusion of the FSA state transition.

The "physical universe" relevant to description of the FSA is completed with the FSA's local environment $\mathcal{B}$, which is embedded in a "greater environment" $\bar{\mathcal{B}}$ and acts to "rethermalize" $\mathcal{B}$ whenever it is driven from equilibrium by interaction with $\mathcal{S}$ during state transitions. $\bar{\mathcal{B}}$ is also taken to include all other subsystems required for global closure of the composite system $\mathcal{R}\mathcal{S}\mathcal{B}\bar{\mathcal{B}}$.

Consider the n-th state transition, which is depicted schematically in Fig. Prior to this transition, the "current" FSA state encoded in the physical state of $\mathcal{S}$ is correlated with $\mathcal{R}_0$ (i.e. the first $n-1$ inputs) but not yet with $\mathcal{R}_1$ (i.e. the n-th input). At the completion of the n-th state transition, correlations will have been created between the state of $\mathcal{S}$ and the state of $\mathcal{R}_1$, and weakens the correlation between $\mathcal{S}$ and $\mathcal{R}_0$. We will quantify this information loss in the next section and show that it necessarily results in dissipation of energy to $\mathcal{B}$, but first provide the formal description of the probabilistic FSA state transitions upon which proof of the dissipation bound is based.

Initial State: Prior to the n-th input, the statistical state of the composite $\mathcal{R}\mathcal{S}\mathcal{B}$ is given by the density operator

$$\hat{\rho}^{\mathcal{R}\mathcal{S}\mathcal{B}} = \hat{\rho}^{\mathcal{R}_0\mathcal{S}} \otimes \hat{\rho}^{\mathcal{R}_1} \otimes \hat{\rho}_{th}^{\mathcal{B}}$$

$$\begin{aligned}\hat{\rho}^{\mathcal{RSB}} &= \left(\sum_i p_i\{\hat{\rho}_i^{\mathcal{R}_0}\otimes\hat{\rho}_i^{\mathcal{S}}\}\right)\otimes\left(\sum_j q_j\hat{x}_j^{\mathcal{R}_1}\right)\otimes\hat{\rho}_{th}^{\mathcal{B}} \\ &= \sum_i\sum_j p_i q_j\hat{\rho}_{ij}^{\mathcal{RSB}}.\end{aligned} \tag{4.11}$$

Here $\hat{\rho}_i^{\mathcal{S}} = \sum_k \pi_{k|i}\hat{\sigma}_k^{\mathcal{S}}$ is a statistical mixture of the states of $\mathcal{S}$ that are correlated to the i-th input string of $\mathcal{R}_0$, that map the initial state of the machine to these FSA states. $\pi_{k|i}$ is the probability that the i-th input maps onto the k-th distinguishable state of the FSA, instantiated as $\hat{\sigma}_k^{\mathcal{S}}$. The FSA state is correlated only to the first $n-1$ inputs, and $\hat{x}_j^{\mathcal{R}_1}$ is the state encoding the n-th FSA input $x_j^{(n)}$ in $\mathcal{R}_1$, and

$$\hat{\rho}_{ij}^{\mathcal{RSB}} = \hat{\rho}_i^{\mathcal{R}_0}\otimes\hat{\rho}_i^{\mathcal{S}}\otimes\hat{x}_j^{\mathcal{R}_1}\otimes\hat{\rho}_{th}^{\mathcal{B}}.$$

State Transition: The n-th state transition is a unitary transformation

$$\hat{\rho}^{\mathcal{RSB}'} = \hat{U}\hat{\rho}^{\mathcal{RSB}}\hat{U}^\dagger$$

of $\mathcal{RSB}$ involving interactions only between $\mathcal{R}_1$, $\mathcal{S}$, and $\mathcal{B}$:

$$\hat{U} = \hat{I}^{\mathcal{R}_0}\otimes\hat{U}^{\mathcal{R}_1\mathcal{SB}}.$$

If the process is to physically implement an FSA defined by abstract states $\sigma_k \in \{\sigma\}$ (physically encoded in register states $\hat{\sigma}_k^{\mathcal{S}} \in \{\hat{\sigma}^{\mathcal{S}}\}$) and by input transition rules $\mathcal{L}_j \in \{\mathcal{L}\}$, then $\hat{U}$ must be such that

$$\hat{\rho}_{ij}^{\mathcal{S}'} = Tr_{\mathcal{RB}}[\hat{U}\hat{\rho}_{ij}^{\mathcal{RSB}}\hat{U}^\dagger].$$

This condition can be written as

$$\hat{\rho}_{ij}^{\mathcal{S}'} = \tilde{\mathcal{L}}_j(\hat{\rho}_i^{\mathcal{S}})$$

to highlight connection to the abstract FSA description, where the $\tilde{\mathcal{L}}_j \in \{\tilde{\mathcal{L}}\}$ are local, nonunitary input transition superoperators that act on $\mathcal{S}$ alone to induce the required state transitions.

Final State: At the conclusion of the n-th state transition, the state of $\mathcal{RSB}$ is:

$$\begin{aligned} \hat{\rho}^{\mathcal{RSB}'} &= \sum_j q_j \left(\hat{x}_j^{\mathcal{R}_1} \otimes \sum_i p_i \left(\hat{\rho}_i^{\mathcal{R}_0} \otimes \hat{\rho}_{ij}^{\mathcal{SB}'} \right) \right) \\ &= \sum_i \sum_j p_i q_j \hat{\rho}_{ij}^{\mathcal{RSB}'} \end{aligned} \tag{4.12}$$

where

$$\hat{\rho}_{ij}^{\mathcal{RSB}'} = \hat{\rho}_i^{\mathcal{R}_0} \otimes \tilde{\mathcal{L}}_j(\hat{\rho}_i^{\mathcal{S}}) \otimes \hat{x}_j^{\mathcal{R}_1} \otimes \hat{\rho}_{ij}^{\mathcal{B}'}.$$

The new register state is correlated with the previous register state (thus the first $n-1$ inputs) *and* the n-th input, as is must be at this stage, and the bath has become correlated with the n-th input and the previous register state.

We conclude this section by noting that each FSA transition is implicitly followed by a spontaneous "rethermalization" of $\mathcal{B}$ by the greater environment $\bar{\mathcal{B}}$. This process "resets" $\mathcal{B}$ to a thermal state before the next FSA transition, washing all information about the history of the FSA from the register's immediate surroundings into the greater environment. Since the "universe" $\mathcal{RSB}\bar{\mathcal{B}}$ is globally closed, this amounts to destruction of correlations between $\mathcal{RS}$ and $\mathcal{B}$ and creation of correlations between $\mathcal{RS}$ and $\bar{\mathcal{B}}$. This rethermalization of the FSA's immediate surroundings by a "greater" environment is a realistic process; it accommodates treatment of a *finite* local environment as an ordinary heat bath - with no memory of past interactions with the

FSA - at the beginning of every state transition. (See [37] for further discussion of this heterogeneous environment model.)

4.5.3 Dissipation Bound for Probabilistic FSAs

In this section, we will state the following theorem for the lower bound on dissipation for the probabilistic FSA described above.

Theorem: For a physical probabilistic FSA $\mathcal{F}_P = \left\{\mathcal{S}, \mathcal{R}, \{\hat{\sigma}^{\mathcal{S}}\}, \{\hat{x}^{\mathcal{R}}\}, \{\tilde{\mathcal{L}}\}\right\}$ and input pmf $\{q\}$, the input-averaged amount of energy dissipated to a thermal environment $\mathcal{B}$ on each state transition is lower bounded in steady state as

$$\Delta\langle E^{\mathcal{B}}\rangle \geq k_B T \ln(2) \left[\sum_j q_j \left(\mathcal{I}^{\mathcal{R}_0\mathcal{S}} - \mathcal{I}_j^{\mathcal{R}_0\mathcal{S}'}\right) + \sum_i p_i (H(\{\pi_{k|i}^{(n-1)}\}) - \sum_{(j)} q_j H(\{\pi_{k'|(i,j)}^{(n)}\}))\right] \tag{4.13}$$

where k_B is the Boltzmann constant, T is the environment temperature, and $\mathcal{I}^{\mathcal{R}_0\mathcal{S}} - \mathcal{I}_j^{\mathcal{R}_0\mathcal{S}'}$ is, for a state transition induced by input $\hat{x}_j^{\mathcal{R}_1}$, the reduction in quantum mutual information between the state of the register system $\mathcal{S}$ and the sequence of all past inputs physically instantiated in referent system $\mathcal{R}_0$. $H(\{\pi_{k|i}^{(n-1)}\})$ is the Shannon entropy of $\{\pi_{k|i}^{(n-1)}\}$, the probability that the i-th input maps to the k-th state of the FSA before the $(n-1)$-th transition. $H(\{\pi_{k|(i,j)}^{(n)}\})$ is Shannon entropy of the distribution $\{\pi_{k|(i,j)}^{(n)}\}$, the probability that the (i,j)-th inputs maps to the k' state after the state transition. The full derivation of this theorem is available in appendix (B.3)

4.6 FSA Computational Efficacy Measures

In this section, we will describe computational efficacy measures for deterministic finite state machines, similar to those described for deterministic L-machines introduce in [37] and discussed in the previous chapter 3, section 3.5. These will allow us to quantify how well the evolution of a physical system accurately instantiates a

particular FSA state transition. The two measures derived here - FSA computational fidelity and representational faithfulness will achieve this and provide a relationship between the efficacy of an instantiation and the lower bound on dissipation.

As before, we can view the deterministic FSA as a set of deterministic $\mathcal{L}$-machines that are conditioned upon the j-th input. Since we know how to characterize the computational efficacy of these $\mathcal{L}$-machines already (from the previous chapter), we can construct the FSA measures as a form of weighted average over the individual $\mathcal{L}$-machines. We will start by providing a different (general) description of the physical deterministic FSA, one in which we do not assume that the physical states representing the automata states are not orthogonal and distinguishable. Prior to the n-th input, let the statistical state of the composite $\mathcal{RSB}$ be given by the density operator

$$\hat{\rho}^{\mathcal{RSB}} = \hat{\rho}^{\mathcal{R}_0\mathcal{S}} \otimes \hat{\rho}^{\mathcal{R}_1} \otimes \hat{\rho}^{\mathcal{B}}_{th}$$

This can be expanded as

$$\begin{aligned} \hat{\rho}^{\mathcal{RSB}} &= \left(\sum_i p_i \{ \hat{\rho}_i^{\mathcal{R}_0} \otimes \hat{\rho}_i^{\mathcal{S}} \} \right) \otimes \left(\sum_j q_j \hat{x}_j^{\mathcal{R}_1} \right) \otimes \hat{\rho}^{\mathcal{B}}_{th} \\ &= \sum_i \sum_j p_i q_j \hat{\rho}_{ij}^{\mathcal{RSB}}. \end{aligned} \tag{4.14}$$

where $\mathcal{S}$, $\mathcal{R}$ and $\mathcal{B}$ are instantiations of the FSA, inputs and thermal bath respectively. We can rewrite the density operator in terms of the k abstract FSA states and the state probability $\pi_k^{(n-1)}$ as

$$\hat{\rho}^{\mathcal{RSB}} = \left(\sum_k \pi_k^{(n-1)} \hat{\rho}_k^{\mathcal{R}_0\mathcal{S}} \right) \otimes \left(\sum_j q_j \hat{x}_j^{\mathcal{R}_1} \right) \otimes \hat{\rho}^{\mathcal{B}}_{th} \tag{4.15}$$

where the density operator of the k-th state is given by $\hat{\rho}_k^{\mathcal{R}_0\mathcal{S}}$ (as opposed to orthogonal $\hat{\sigma}_k^{\mathcal{S}}$'s from the earlier sections)

$$\hat{\rho}_k^{\mathcal{R}_0\mathcal{S}} = \frac{1}{\pi_k^{(n-1)}} \sum_i p_i \pi_{k|i}^{(n-1)} (\hat{\rho}_i^{\mathcal{R}_0} \otimes \hat{\rho}_i^{\mathcal{S}}).$$

where we have $\pi_{k|i}^{(n-1)}$ as the probability that the i-th input string maps to the k-th state of the FSA (assuming we are starting from an initial single state), and we have $\sum_k \pi_{k|i}^{(n-1)} = 1$.

At the conclusion of the n-th state transition, the state of $\mathcal{RSB}$ is:

$$\begin{aligned} \hat{\rho}^{\mathcal{RSB}'} &= \sum_j q_j \left(\hat{x}_j^{\mathcal{R}_1} \otimes \sum_i p_i \left(\hat{\rho}_i^{\mathcal{R}_0} \otimes \hat{\rho}_{ij}^{\mathcal{SB}'} \right) \right) \\ &= \sum_j q_j \left(\hat{x}_j^{\mathcal{R}_1} \otimes \hat{\rho}_j^{\mathcal{R}_0\mathcal{S}'} \right) \end{aligned} \tag{4.16}$$

where $\hat{\rho}_j^{\mathcal{R}_0\mathcal{S}'}$ can be expanded as

$$\begin{aligned} \hat{\rho}_j^{\mathcal{R}_0\mathcal{S}'} &= \sum_i p_i (\hat{\rho}_i^{\mathcal{R}_0} \otimes \hat{\rho}_{ij}^{\mathcal{S}'}) \\ &= \sum_{k'} \pi_{k'|j}^{(n)} \hat{\rho}_{k'|j}^{\mathcal{R}_0\mathcal{S}'} \\ &= \sum_{k'} \pi_{k'|j}^{(n)} \left(\frac{1}{\pi_{k'|j}^{(n)}} \sum_i p_i \pi_{k'|(i,j)}^{(n)} (\hat{\rho}_i^{\mathcal{R}_0} \otimes \hat{\rho}_{ij}^{\mathcal{S}'}) \right) \end{aligned} \tag{4.17}$$

where $\pi_{k'|j}^{(n)}$ is the probability that the j-th inputs maps to the k' state of the FSA at the n-th transition, and $\pi_{k'|(i,j)}^{(n)}$ is the probability that the (i,j)-th inputs of $\mathcal{R}_0\mathcal{R}_1$ maps to the k' of the FSA at the n-th transition. For consistency, we require that $\sum_{k'} \pi_{k'|j}^{(n)} = 1$ and $\sum_{k'} \pi_{k'|(i,j)}^{(n)} = 1$.

As before the dissipation bound for the FSA in steady state is given as

$$\Delta\langle E^{\mathcal{B}}\rangle \geq k_B T \ln(2) \sum_j q_j \left(\mathcal{I}^{\mathcal{R}_0\mathcal{S}} - \mathcal{I}_j^{\mathcal{R}_0\mathcal{S}'} \right)$$

where we now have

$$\mathcal{I}^{\mathcal{R}_0\mathcal{S}} = S(\hat{\rho}^{\mathcal{S}}) - \sum_i p_i S(\hat{\rho}_i^{\mathcal{S}})$$

$$\mathcal{I}_j^{\mathcal{R}_0\mathcal{S}'} = S(\hat{\rho}_j^{\mathcal{S}'}) - \sum_i p_i S(\hat{\rho}_{ij}^{\mathcal{S}'})$$

We can rewrite the terms inside the bracket as $\mathcal{I}^{\mathcal{R}_0\mathcal{S}} - \mathcal{I}_j^{\mathcal{R}_0\mathcal{S}'}$ as

$$\mathcal{I}^{\mathcal{R}_0\mathcal{S}} - \mathcal{I}_j^{\mathcal{R}_0\mathcal{S}'} = \mathcal{I}^{\mathcal{R}_0\mathcal{S}} - H(X^{(n-1)}) + H(X^{(n-1)}) - \mathcal{I}_j^{\mathcal{R}_0\mathcal{S}'}$$

where $H(X^{(n-1)})$ is the Shannon entropy of the variable $X^{(n-1)}$ with probability distribution $\{\pi_k^{(n-1)}\}$, that represents the probability distribution of the k states of the FSA prior the n-th state transition. In the earlier derivation of the lower bound on dissipation for deterministic FSAs, we have $\mathcal{I}^{\mathcal{R}_0\mathcal{S}} = H(X^{(n-1)})$. This is because we implicitly assumed that there are orthogonal physical states $\{\hat{\sigma}_k^{\mathcal{S}}\}$ of $\mathcal{S}$ that are the physical instantiations of the abstract FSA states, and evolution of the system are faithful realization of the FSA state transitions. If the deterministic FSA indeed is not properly physically instantiated (as it is required in order to define efficacy measures), then we should expect for $\mathcal{I}^{\mathcal{R}_0\mathcal{S}} \neq H(X^{(n-1)})$. Clearly we can see that the $\mathcal{I}^{\mathcal{R}_0\mathcal{S}}$ and $H(X^{(n-1)})$ terms are independent of j. Remembering that since FSA can be viewed as $\mathcal{L}$-machines conditioned on the j-th input, the second half of the expression above $H(X^{(n-1)}) - \mathcal{I}_j^{\mathcal{R}_0\mathcal{S}'}$ is very similar to the information loss term used to obtain computational efficacy measures for $\mathcal{L}$-machines in the previous chapter. We can now write the $H(X^{(n-1)})$ for the j-th conditioned $\mathcal{L}$-machine as

$$H(X^{(n-1)}) = H(Y_j^{(n)}) + H(X^{(n-1)}|Y_j^{(n)})$$

where $Y_j^{(n)}$ is a random variable that represents the states of the FSA after the n-th state transition for the j-th input of $\mathcal{R}_1$, and are characterized by the probability distribution $\{\pi_{k'}|j\}$. $H(X^{(n-1)}|Y_j^{(n)})$ is the conditional Shannon entropy of $X^{(n-1)}$ given $Y_j^{(n)}$. We can expand $\mathcal{I}_j^{\mathcal{R}_0\mathcal{S}'}$ in a similar manner

$$\begin{aligned}\mathcal{I}_j^{\mathcal{R}_0\mathcal{S}'} &= S(\hat{\rho}_j^{\mathcal{S}'}) - \sum_i p_i S(\hat{\rho}_{ij}^{\mathcal{S}'}) \\ &= S(\hat{\rho}_j^{\mathcal{S}'}) - \sum_{k'} \pi_{k'|j}^{(n)} S(\hat{\rho}_{k'|j}^{\mathcal{S}'}) + \sum_{k'} \pi_{k'|j}^{(n)} S(\hat{\rho}_{k'|j}^{\mathcal{S}'}) - \sum_i p_i S(\hat{\rho}_{ij}^{\mathcal{S}'})\end{aligned} \tag{4.18}$$

where we defined $\hat{\rho}_{k'|j}^{\mathcal{S}'} = \frac{1}{\pi_{k'|j}^{(n)}} \sum_i p_i \pi_{k'|(i,j)}^{(n)} (\hat{\rho}_i^{\mathcal{R}_0} \otimes \hat{\rho}_{ij}^{\mathcal{S}'})$ and that $\sum_{k'} \pi_{k'|(i,j)}^{(n)} = 1$, we have

$$\begin{aligned}\sum_{k'} \pi_{k'|j}^{(n)} \hat{\rho}_{k'|j}^{\mathcal{S}'} &= \sum_{k'} \pi_{k'|j}^{(n)} \left(\frac{1}{\pi_{k'|j}^{(n)}} \sum_i p_i \pi_{k'|(i,j)}^{(n)} \hat{\rho}_{ij}^{\mathcal{S}'} \right) \\ &= \sum_{k'} \sum_i p_i \pi_{k'|(i,j)}^{(n)} \hat{\rho}_{ij}^{\mathcal{S}'} \\ &= \sum_{k'} \pi_{k'|(i,j)}^{(n)} \sum_i p_i \hat{\rho}_{ij}^{\mathcal{S}'} \\ &= \sum_i p_i \hat{\rho}_{ij}^{\mathcal{S}'}\end{aligned} \tag{4.19}$$

This allows us to write

$$\sum_i p_i S(\hat{\rho}_{ij}^{\mathcal{S}'}) = \sum_{k'} \pi_{k'|j}^{(n)} \left(\frac{1}{\pi_{k'|j}^{(n)}} \sum_i p_i \pi_{k'|(i,j)}^{(n)} S(\hat{\rho}_{ij}^{\mathcal{S}'}) \right)$$

If we define the ensemble $\epsilon_{Y_j^{(n)}}^{\mathcal{S}'} = \{\pi_{k'|j}^{(n)}, \hat{\rho}_{k'|j}^{\mathcal{S}'}\}$ and the ensemble $\epsilon_{k'|j}^{\mathcal{S}'} = \{\frac{p_i \pi_{k'|(i,j)}}{\pi_{k'|j}^{(n)}}, \hat{\rho}_{ij}^{\mathcal{S}'}\}$, we can then write the above expression for $\mathcal{I}_j^{\mathcal{R}_0\mathcal{S}'}$ as

$$\begin{aligned}\mathcal{I}_j^{\mathcal{R}_0\mathcal{S}'} &= S(\hat{\rho}_j^{\mathcal{S}'}) - \sum_{k'} \pi_{k'|j}^{(n)} S(\hat{\rho}_{k'|j}^{\mathcal{S}'}) + \sum_{k'} \pi_{k'|j}^{(n)} S(\hat{\rho}_{k'|j}^{\mathcal{S}'}) - \sum_i p_i S(\hat{\rho}_{ij}^{\mathcal{S}'}) \quad (4.20)\\ &= \chi(\epsilon_{Y_j^{(n)}}^{\mathcal{S}'}) + \sum_{k'} \pi_{k'|j}^{(n)} \chi(\epsilon_{k'|j}^{\mathcal{S}'})\end{aligned}$$

where $\chi(\epsilon_{Y_j^{(n)}}^{\mathcal{S}'})$ and $\chi(\epsilon_{k'|j}^{\mathcal{S}'})$ are the Holevo information associated with the $\epsilon_{Y_j^{(n)}}^{\mathcal{S}'}$ and $\epsilon_{k'|j}^{\mathcal{S}'}$ respectively. Thus we can write part of information loss about $\mathcal{R}_0$ associated with j-th conditioned $\mathcal{L}$ machine of the FSA as

$$H(X^{(n-1)}) - \mathcal{I}_j^{\mathcal{R}_0\mathcal{S}'} = \left[H(Y_j^{(n)}) + H(X^{(n-1)}|Y_j^{(n)})\right] - \left[\chi(\epsilon_{Y_j^{(n)}}^{\mathcal{S}'}) + \sum_{k'} \pi_{k'|j}^{(n)} \chi(\epsilon_{k'|j}^{\mathcal{S}'})\right]$$

Rearranging the above terms as

$$H(X^{(n-1)}) - \mathcal{I}_j^{\mathcal{R}_0\mathcal{S}'} = H(Y_j^{(n)}) - \chi(\epsilon_{Y_j^{(n)}}^{\mathcal{S}'}) + H(X^{(n-1)}|Y_j^{(n)}) - \sum_{k'} \pi_{k'|j}^{(n)} \chi(\epsilon_{k'|j}^{\mathcal{S}'})$$

As defined in the previous chapter, we can define the computational fidelity F_j and representational faithfulness f_j of the j-th $\mathcal{L}$ machine of the FSA as the following

$$F_j = \frac{\chi(\epsilon_{Y_j^{(n)}}^{\mathcal{S}'})}{H(Y_j^{(n)})}$$

$$f_j = 1 - \frac{\sum_{k'} \pi_{k'|j}^{(n)} \chi(\epsilon_{k'|j}^{\mathcal{S}'})}{H(X^{(n-1)}|Y_j^{(n)})}$$

In the above equations,f_j captures whether the logical transformation associated with the j-th $\mathcal{L}$ is faithfully implemented, i.e. whether the physical states of $\mathcal{S}$ that represent the FSA states after the state transition have more information about the present state than that is allowed by the logical state transition mappings. F_j is a measure on the distinguishability of the FSA states irrespective of the faithfulness., and the amount of information of the next FSA states in the j-th $\mathcal{L}$ machine that is encoded in the states of $\mathcal{S}$. Now armed with the fidelity and faithfulness of each of the conditioned $\mathcal{L}$-machines, we can now derive the efficacy measures of the entire deterministic FSA.

4.6.1 FSA Representational Faithfulness

The $\mathcal{L}$-machine representational faithfulness was used to capture whether the computational channel completed the work of implementing a logical transformation $\mathcal{L}$. This would require all device input states belonging to the same logical output of $\mathcal{L}$ must evolve into the same device output state. In FSAs, it would be necessary for all the current FSA states (before the state transition) that map to the same next abstract FSA state (after the state transition), to evolve to the same physical FSA state of $\mathcal{S}$. The FSA faithfulness measure will capture "how well" the physical system realizes this across all the $\mathcal{L}$-machines. This is equivalent to requiring that all of the conditioned $\mathcal{L}$ to be instantiated faithfully. Thus $\forall(j,k) \in \{(j,k)\}_{k'}$, we would have $\mathcal{L}_j(\hat{\sigma}_k^{\mathcal{S}}) = \hat{\rho}_{jk}^{\mathcal{S}'} = \hat{\sigma}_{k'}^{\mathcal{S}'}$. The *FSA representational faithfulness* will be defined as

$$f_{FSA} = 1 - \frac{\sum_j \sum_{k'} \pi_{k'|j}^{(n)} \chi(\epsilon_{k'|j}^{\mathcal{S}'})}{\sum_j H(X^{(n-1)}|Y_j^{(n)})} \tag{4.21}$$

This can be rewritten in terms of f_j as

$$f_{FSA} = \frac{\sum_j q_j f_j H(X^{(n-1)}|Y_j^{(n)})}{\sum_j q_j H(X^{(n-1)}|Y_j^{(n)})} \tag{4.22}$$

Since we have $0 \leq f_j \leq 1$, we will also have $0 \leq f_{FSA} \leq 1$, with $F_{FSA} = 0$ if and only if the $f_j = 0$ for all j. That would mean that each of the individual conditioned $\mathcal{L}$-machines are instantiated completely unfaithfully, and thus the entire FSA is also unfaithfully instantiated. And when $f_j = 1$ for all j, we have the representational faithfulness of the FSA $f_{FSA} = 1$, which indicates that the physical states of $\mathcal{S}$ after the state transition do not contain more information about the previous states of the FSA before the state transition than what is allowed by the abstract state mappings.

4.6.2 FSA Computational Fidelity

The FSA computational fidelity F_{FSA} is measure of the distinguishability of FSA states after the state transition, independent of it's faithfulness. It is related to the amount of information about that states of the FSA after the transition that is encoded in physical state of $\mathcal{S}$. It can be defined as the

$$F_{FSA} = \frac{\sum_j q_j \chi(\epsilon_j^{\mathcal{S}'})}{\sum_j q_j H(Y_j^{(n)})} \tag{4.23}$$

The above equation can be written in terms of the computational fidelity of the conditioned $\mathcal{L}$ machines F_j as

$$F_{FSA} = \frac{\sum_j q_j F_j H(Y_j^{(n)})}{\sum_j q_j H(Y_j^{(n)})} \tag{4.24}$$

From the above equation we can see that similar to the F_j's, $0 \leq F_{FSA} \leq 1$. Also $F_{FSA} = 1$ when all the F_j's are equal to 1. This means that if in all the conditioned $\mathcal{L}$ machines, if the states of the FSA after the state transition are perfectly distinguishable, then the states of FSA as a whole are perfectly distinguishable. Similarly we have $F_{FSA} = 0$, when $F_j = 0$ for all j. Thus if none of the states of the FSA in any of the j $\mathcal{L}$ machines, then the states of the FSA as a whole are also indistinguishable. Along with representational faithfulness, these efficacy measures allows us to quantify how well a FSA state transition is instatiated in the evolution of a physical system $\mathcal{S}$. In the next subsection, we will relate these measures to the information loss that occurs over the state transition.

4.6.3 Information Loss in the FSA in terms of FSA Efficacy Measures

In this section, we will describe the information lost about the past inputs $\mathcal{R}_0$ over a state-transition in terms of the FSA efficacy measures described in the previous

section. This will allow us to directly understand the effect of the efficacy measures on the lower bound on dissipation.

$$\mathcal{I}^{\mathcal{R}_0\mathcal{S}} - \sum_j \mathcal{I}_j^{\mathcal{R}_0\mathcal{S}'} = \mathcal{I}^{\mathcal{R}_0\mathcal{S}} - H(X^{(n-1)}) + H(X^{(n-1)}) - \sum_j \mathcal{I}_j^{\mathcal{R}_0\mathcal{S}'}$$

From the previous subsections, we know that we can write $H(X^{(n-1)}) - \sum_j \mathcal{I}_j^{\mathcal{R}_0\mathcal{S}'}$ as

$$H(X^{(n-1)}) - \sum_j \mathcal{I}_j^{\mathcal{R}_0\mathcal{S}'} = \sum_j q_j \left(H(Y_j^{(n)}) + H(X^{(n-1)}|Y_j^{(n)}) - \chi(\epsilon^{\mathcal{S}'}_{Y_j^{(n)}}) + \sum_{k'} \pi^{(n)}_{k'|j} \chi(\epsilon^{\mathcal{S}'}_{k'|j}) \right) \tag{4.25}$$

Rearranging the terms, we can rewrite the above expression in terms of the FSA efficacy measures as

$$H(X^{(n-1)}) - \sum_j \mathcal{I}_j^{\mathcal{R}_0\mathcal{S}'} = (1 - F_{FSA}) \sum_j q_j H(Y_j^{(n)}) + f_{FSA} \sum_j q_j H(X^{(n-1)}|Y_j^{(n)}) \tag{4.26}$$

Thus the information loss about $\mathcal{R}_0$ conditioned upon $\mathcal{R}_1$ can be written as

$$\begin{aligned} \mathcal{I}^{\mathcal{R}_0\mathcal{S}} - \sum_j \mathcal{I}_j^{\mathcal{R}_0\mathcal{S}'} &= \mathcal{I}^{\mathcal{R}_0\mathcal{S}} - H(X^{(n-1)}) \\ &+ (1 - F_{FSA}) \sum_j q_j H(Y_j^{(n)}) + f_{FSA} \sum_j q_j H(X^{(n-1)}|Y_j^{(n)}) \end{aligned} \tag{4.27}$$

As was the case with information loss in the $\mathcal{L}$ machine, the first term in the above equation represents the undesirable information loss associated with the indistinguishability of the FSA states, and the second term corresponds to the necessary information loss that is needed to faithfully implement the state transition. When the FSA state transitions are implemented perfectly, $\mathcal{I}^{\mathcal{R}_0\mathcal{S}} - H(X^{(n-1)})$ and we would have $F_{FSA} = 1$ and $f_{FSA} = 1$. Thus the conditioned information loss is given by

$\mathcal{I}^{\mathcal{R}_0\mathcal{S}} - \sum_j \mathcal{I}_j^{\mathcal{R}_0\mathcal{S}'} = \sum_j q_j H(X^{(n-1)}|Y_j^{(n)})$. If the state transition is implemented unfaithfully $f_{FSA} = 0$, but with perfectly distinguishable states $F_{FSA} = 1$, we have $\mathcal{I}^{\mathcal{R}_0\mathcal{S}} - \sum_j \mathcal{I}_j^{\mathcal{R}_0\mathcal{S}'} = 0$ if $\mathcal{I}^{\mathcal{R}_0\mathcal{S}} = H(X^{(n-1)})$.

4.6.4 Lower Bounds on Energy Dissipation in Terms of Efficacy Measures

In Eq. (4.27), we have related the conditioned information loss in a state transition with the efficacy which indicated how well that FSA state transition was achieved. Since this information loss is directly related with the heat dissipation to the environment for a FSA in steady state, substituting Eq. (4.27) in Eq. (B.1), we get

$$\begin{aligned} \Delta\langle E^{\mathcal{B}}\rangle &\geq k_B T \ln(2)[\mathcal{I}^{\mathcal{R}_0\mathcal{S}} - H(X^{(n-1)}) \\ &+ (1 - F_{FSA}) \sum_j q_j H(Y_j^{(n)}) + f_{FSA} \sum_j q_j H(X^{(n-1)}|Y_j^{(n)})] \end{aligned} \tag{4.28}$$

The above equation provides a very important relation between the lower bound on the physical cost the user must pay to achieve the FSA state transition in terms of the efficacy measures fidelity and faithfulness that quantify how well the state transition has been achieved.

4.7 Dissipation in FSA with Correlated Inputs

Throughout this chapter, we have described both deterministic and probabilistic FSA that are driven by IID inputs. However in a lot of cases, the inputs to the FSA are often temporally correlated, especially in learning operations which are characterized by inputs with significant spatial and temporal correlations. In this section, we will provide the lower bound on dissipation in finite state automata for temporally correlated inputs. The physical probabilistic/deterministic FSA will be described as before, except for the fact that the inputs $\mathcal{R}$ will not be independent in time anymore as shown in Fig. 4.10.

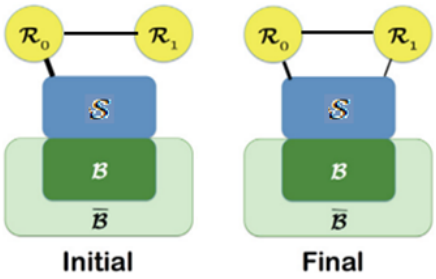

Figure 4.10. Physical description of an FSA with correlated driving inputs undergoing a state transition. The system $\mathcal{S}$, which registers the FSA state, is initially correlated with previous inputs physically encoded in $\mathcal{R}_0$ and is indirectly correlated to $\mathcal{R}_1$ through $\mathcal{R}_0$. The quantum mutual information between $\mathcal{R}_1$ and $\mathcal{S}$ before the state transition $I^{\mathcal{R}_1\mathcal{S}}$ can be seen as a prediction component.

Theorem: For physical FSA $\mathcal{F}_P = \left\{\mathcal{S}, \mathcal{R}, \{\hat{\rho}^{\mathcal{S}}\}, \{\hat{x}^{\mathcal{R}}\}, \{\tilde{\mathcal{L}}\}\right\}$ and input pmf $\{q\}$, the input-averaged amount of energy dissipated to a thermal environment $\mathcal{B}$ on each state transition is lower bounded as

$$\Delta\langle E^{\mathcal{B}}\rangle \geq k_B T \ln(2) \left(\sum_j q_j [S(\hat{\rho}_j^{\mathcal{S}}) - S(\hat{\rho}_j^{\mathcal{S}'})]\right) \tag{4.29}$$

where k_B is the Boltzmann constant and T is the environment temperature. For a state transition induced by input $\hat{x}_j^{\mathcal{R}_1}$, $\hat{\rho}_j^{\mathcal{S}}$ and $\hat{\rho}_j^{\mathcal{S}'}$ are the density operators associated with the j-th input before and after the state transition. This can be rewritten in information theoretic terms as

$$\Delta\langle E^{\mathcal{B}}\rangle \geq k_B T \ln(2) \left[-\Delta S^{\mathcal{S}} + \Delta\mathcal{I}^{\mathcal{R}_1\mathcal{S}}\right] \tag{4.30}$$

where $-\Delta S^{\mathcal{S}}$ is the reduction in von Neumann entropy of the system $\mathcal{S}$ over the transition, and $\Delta\mathcal{I}^{\mathcal{R}_1\mathcal{S}} = \mathcal{I}^{\mathcal{R}_1\mathcal{S}'} - \mathcal{I}^{\mathcal{R}_1\mathcal{S}}$ is the change in quantum mutual information between the system $\mathcal{S}$ and the latest input $\mathcal{R}_1$. The quantum mutual information between $\mathcal{S}$ and $\mathcal{R}_1$ before the state transition $\mathcal{I}^{\mathcal{R}_1\mathcal{S}}$, can be seen as a measure of prediction of the next input $\mathcal{R}_1$ by the system $\mathcal{S}$. In the next few sections, we will

describe a simple machine trying to learn the external input pixel values and calculate the lower bound on dissipation for both temporally correlated and independent inputs. The detailed proof of this theorem is available in appendix (B.4).

4.7.1 Illustrative Example: A Simple Learning Machine

We now apply the lower bound on dissipation for FSA with correlated inputs to a device with a functionality and input environment inspired by learning applications. This example is adapted from the author's work in [44]. Consider a system comprising an array of simplified artificial synapses. The system is a functionally enhanced memory tasked with learning or creating a model of a slowly changing environment from partial observations. While learning is essential, most experiences do not cause a given synapse to change state, and we will exploit this low probability of actual learning to lower the minimum energy of operation. The environment comprises of an array of $n \times n$ (here n=3) data items or pixels that take the values 1 and +1. We will evaluate two different scenarios for the environment, one where all the pixels are spatially independent and the other where the pixels in a row are perfectly correlated. Observations are of one pixel (or row) at a time, with probability p that a specific pixel (respectively, row) is observed in each step in cases of spatially independent (respectively, correlated) pixels. The system has an internal $n \times n$ array of functionally enhanced storage cells and shift registers that drives both the row and column of the internal array with the observed pixel value of 1 or +1. When the selected cell receives $(1, 1)$ or $(+1, +1)$, it remembers the stimulus value. Each pixel in the environment changes with time at a rate corresponding to a probability q of a change per observation. The system will be modeled in steady state, so an initial condition is not needed. The system could drive multiple rows and columns at once and include both 1 and +1 data values in the same observation, but this will not be considered here. An implementation of the example system is illustrated in Fig. 2A,

which is an $n \times n$ array of the synapse machines in a framework that transmits data past the array as shown.

The system monitors a stream of $2n$ parallel data inputs from the environment (one for each row and column), which is assumed to be ongoing and which is not destroyed or erased by the system. For the case of single pixel observations, the stream provides a single nonzero, 0/1, stimulus on each set of $2n$ data inputs as shown in Fig. 4.11(a) to write into the corresponding core. (In the case of the spatially correlated environment, the stream contains multiple 0 inputs to update an entire row of cores with the same value.) As the data flows downward through the $2n$ shift registers, the values on the bottom row are translated into current in the blue and red wires. The wires become rows and columns of an array tilted at 45 degrees where the row-column intersections each flow through the center of a core. Each core flips to align with its magnetic field, but only if the field is above a threshold and a core will not flip if it is already in the correct state. The system would be engineered to flip magnetization at 1.5 units of current flowing through each core. Thus, a core exposed to +1 on the row wire and +1 on the column wire will have total current +2 and would flip magnetization to the green state provided it was not in the right state already. Vice versa for -1 and a red state. Magnetic cores dissipate energy when they change state, but nearly zero energy otherwise. Unless the two currents are in the same direction, the total current will be below the threshold and there will no state change and no energy dissipation associated with core state changes. Fig. 4.11(a) illustrates the system processing the data, specifically at step n. Steps 1-3 cause the system to learn pixels, setting the three non-white cores shown in figure; the white cores are irrelevant to the discussion and could be either red or green. The system then experiences a long sequence of steps containing repeating known pixels. In the last row of Table I, the learning machine observes a change in the external data set. The $\{bottom, left\}$ pixel changes from 1 to +1 and is recorded as the leftmost

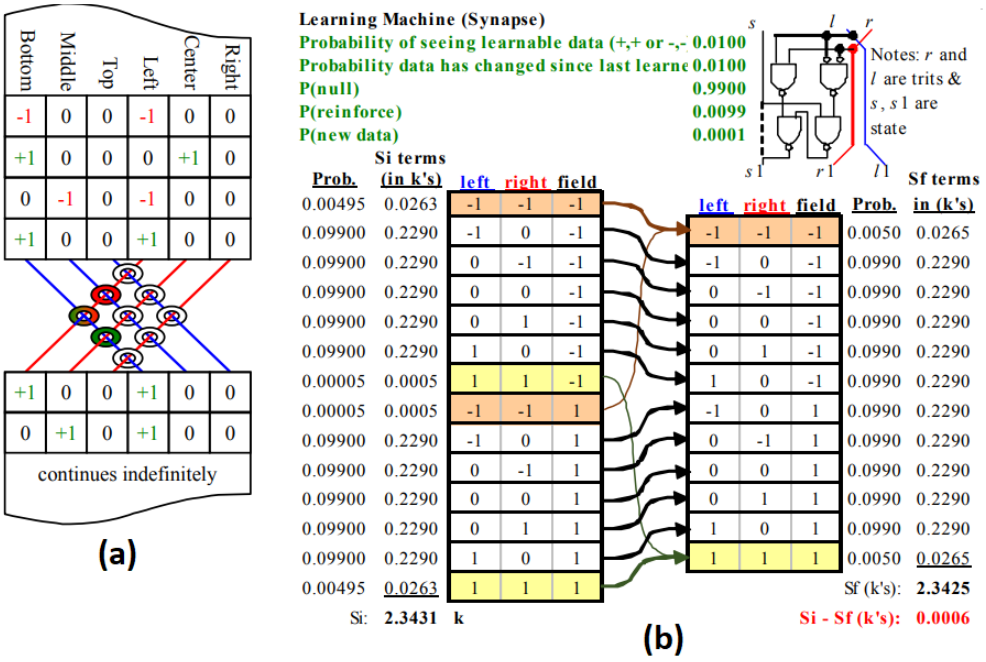

Figure 4.11. (a) The learning machine described in the section using the 3×3 cores receiving $2 \times 3 = 6$ inputs from the environment [44]. (b) The dissipation analysis using the bound for FSA with temporally correlated inputs described in this section. Also included in the top-right corner is the gate equivalent circuit of every learning core.

core in Fig. 4.11(a) flips. We now consider lower bounds on the energy dissipation for this learning machine.

4.7.2 Dissipation Analysis for Learning Machine

In this section, we obtain lower dissipation bounds for the learning machine described above. We will start with a limiting dissipation analysis of a single core, and then calculate the same for the entire learning machine and elucidate the differences in the dissipation for the two different pixel environment and the input stream cases mentioned in the previous section. Each magnetic core behaves as a finite-state automaton, as does the entire learning machine. Thus we use the dissipation bounds obtained earlier in the chapter for FSA driven by IID information sources [138], as well

as by inputs with temporal correlations that would be common for learning scenarios in slow-changing environments.

The FSA description of each core is as follows: The FSA state corresponds to the current magnetization state of the core. FSA inputs l and r correspond to the current states in the blue and red wire respectively. The next state of the core s' depends upon its current state s and the input values on the wires. We use the random variables S, S', L and R for a statistical description of the current and next state of the core, and for the two inputs, respectively. Assuming that the magnetization states of the core are perfectly distinguishable, the minimum energy dissipated into the environment (thermal bath $\mathcal{B}$ at temperature T) as the core (in steady state) undergoes a transition from s to s is

$$\langle \Delta E^{\mathcal{B}} \rangle \geq k_B T \ln(2) \left[H(S|LR) H(S'|LR) \right]$$

where $H(S|LR)$ and $H(S'|LR)$ are the conditional Shannon entropies of the core state distribution given the inputs, before and after the state transition respectively. The inputs $(l, r) = (+1, +1)$ and $(l, r) = (1, 1)$ write $+1$ and 1 into the core states respectively, regardless of the previous state. This merging of the core states for certain l and r inputs is the source of the irreversibility and energy dissipation into the environment.

We have calculated the limiting dissipation for the learning machine with $p = 0.01$ and $q = 0.01$, where p is the probability of seeing learnable data, i.e. the probability of seeing the inputs $(l, r) = (+1, +1)$ or $(l, r) = (1, 1)$. And q is the probability that given the presence of learnable data, the data value changes in the environment since the last time that data was observed. The input probabilities are functions of p, q, and the steady state core state distribution is $P(S = +1) = P(S = 1) = 0.5$. The lower bound on energy dissipation calculated for a single core of the learning machine both from the FSA description and the modified Landauer-like analysis (Fig. 4.11(b)) is $\langle \Delta E^{\mathcal{B}} \rangle \geq 0.0006 k_B T$ per operation. The $1,000\times$ difference between

the limiting dissipation for the magnetic core and the "$k_BT \ln 2$ per operation" rule of thumb stems largely from the input probabilities selected for this learning example, which correspond to learning with a slowly-changing environment. We now extend our analysis to the entire learning machine for the two scenarios introduced in the previous section. The magnetic cores are assumed not to interact with one other. In the first case, the pixels in the 3×3 environment are spatially independent and the cores updated one at a time randomly. The limiting dissipation bound for the entire learning machine will be equal to the sum of the dissipation bounds for the nine individual cores. For $p = 0.01$ and $q = 0.01$, we have the lower bound on the energy dissipated into the environment for the nine-core learning machine to be $\langle \Delta E^{\mathcal{B}} \rangle \geq 9 \times 0.0006kT = 0.0054kT$. In the second case, updating an entire row with correlated inputs, will produce correlations between the cores of each row. As a result, the limiting dissipation of the entire learning machine will be < 9 times that of a single core. Using the same values for p and q as before, we have the lower bound on the energy dissipation of the learning machine of $\langle \Delta E^{\mathcal{B}} \rangle \geq 0.00168k_BT$. Thus, the limiting dissipation values for variations of the learning machine can vary significantly, depending upon the characteristics of the input environment and the updating scheme employed, even for a fixed limiting dissipation values for the individual cores.

We propose the following **Principle of aggregation** conjecture:

The minimum energy dissipation of a function will always be less than or equal to the minimum for a realization as a disaggregated group of lower level (often non-optimal) logical primitives like NAND and NOr.

The proof of this conjecture is beyond the scope of this dissertation, but represents an important problem that needs to be rigorously solved. However we will illustrate this principle using the following example. Consider the magnetic core from the learning machine. Each of the nine cores is functionally equivalent to the logic circuit in Fig. 4.11(b), comprised of NAND primitives (two of which use three-valued inputs). A

dissipation analysis of this circuit using the same input distribution as the magnetic core implementation, and assuming that the gate operations are not conditioned upon l and r inputs, gives a dissipation bound of $\langle \Delta E^{\mathcal{B}} \rangle \geq 2.8939 k_B T$ (obtained by calculating the lower bound on dissipation for each gate in the circuit given the probability distributions of its inputs). This is much greater than $0.0006 k_B T$, the large difference attributable to a highly non-optimal disaggregation of the logic function using gate-level primitives. This dramatically illustrates both the aggregation principle and the need for careful analysis and interpretation of the fundamental lower bound on computation. Furthermore analysis of this type will pave the way moving forward in identifying optimal primitives for implementation of different functions.

4.8 Summary

In this chapter, we introduced a physical description of finite state state automata and derived a fundamental lower bound on the average energy dissipated per state transition in a finite-state automaton was obtained for deterministic FSAs without an output, Mealy machine with outputs and irreducible probabilistic FSA driven by random, classical input strings. The bound, which follows from dynamical laws and entropic inequalities alone, depends on the input-averaged amount of information the FSA loses about its own history on each step as well as the temperature of the FSAs local environment. In the case of the deterministic FSA, the quantity corresponding to input-averaged information loss was proposed as a measure of the computational irreversibility of an FSA driven by an input source with specified statistics. We then calculated the bounds for a simple 2-bit counter with and without an output.

Following this, we introduced computational efficacy measures for the FSA, similar to the ones discussed in the previous chapter for $\mathcal{L}$-machines - the FSA computational fidelity and representational faithfulness and established the relationship between these measures and the lower bound on dissipation. We concluded the chapter by

deriving the lower bound on dissipation for a FSA that is driven by a source that produces temporally correlated inputs. The correlation between the FSA system $\mathcal{S}$ and the incoming input forms a prediction component, which lowers the dissipation bound. The theoretical tools obtained in this chapter can be used to explore limits on the inherent dissipative characteristics of new and unfamiliar approaches to digital computation, independent of considerations related to technological implementation. In the next chapter, we will use these tools to study the fundamental lower bounds on dissipation in neural networks, and establish the ultimate dissipation limits of learning.

CHAPTER 5

DISSIPATION IN NEURAL NETWORKS

With the future of computing heavily geared towards applications that involve handling and learning from large amounts of data, nanoscale implementation of machine learning algorithms in neuromorphic hardware will greatly increase the efficiency with which large amounts of data can be handled and learned from [53]. Possible implementations include phase change material (PCM) [54], spin torque [55], memristor-based [56] and optical [57] neuromorphic systems. As exploration of these and other emerging computing paradigms intensifies, evaluation of their energy efficiency limits will become increasingly important. We must know where these limits lie for complex systems realized in existing and emerging nanocomputing paradigms - including neuromorphic paradigms - if we are to comparatively assess their ultimate potential for energy efficient computation.

In the previous chapter, the fundamental lower bounds on dissipation for FSA were derived and calculated for simple example cases. In this chapter, we will extend the that work and present results on fundamental, technology independent dissipation limits associated with training and testing feedforward neural networks will be presented and evaluated for a simple perceptron on a classification task. Dissipation costs associated with the use of Hopfield and Boltzmann networks as content addressable memories are also studied. Such analyses represent first steps in the determination of fundamental efficiency limits for complex neuromorphic systems. The chapter is organized as follows - we will first introduce feedforward networks, and their training and testing procedures, followed by Hopfield and Boltzmann recurrent

neural networks. We will then proceed to describe training and testing in feedforward and recurrent network as state transitions, and calculate the lower bound on energy dissipation associated with each of these networks during both the training and testing phases. We then study the effect of the learning rate, input probability distributions, update policy and pseudo-temperature on the dissipation bound with illustrative examples. The chapter will end with a discussion on the identification of learning rules from thermodynamic cost functions.

5.1 Neural Networks and Threshold Logic

The idea of artificial neural networks were derived from their biological counterparts. They are networks of functions in which the nodes are made up of the simplest kind of computing units that are a generalization of the common logic gates used in conventional computing. These units usually operate by comparing their total input with a threshold and are known as threshold logic. The directed edges of the network between the nodes are weighted. The goal of this network is to behave like a "mapping machine" and model the n-input, m-output function $\mathcal{F} : \mathcal{R}^n \rightarrow \mathcal{R}^m$.

The function computed at every nodal unit in the network is a simple function of the n-incoming inputs. Since the inputs have to reduced to a single numerical value in threshold computing units, they are divided into two functional parts. The first part is the integration function f that reduces the n arguments to a single value, and the output or activation function g produces the output of this node taking that single value as its argument. A common integration function used is the weighted addition function (where the inputs are weighted according to the strength of the directed edges in the network. The whole operation can also be viewed as a matrix multiplication). The McCulloh-Pitts neuron is one of the simplest forms of threshold units which uses the weighted addition function for f, and the activation function is

the non-linear Heaviside step function with threshold θ. Thus the output of these units are always a 0 or a 1.

A simple and popular classification of these neural networks is based on the edges between these nodes and the flow of information through the network. There are feedforward networks in which there no cycles and information flows from the output to one node, through the directed edge to the input of the next node. The computation is well defined and there is no need to synchronize the computing units. These networks are extremely powerful since it can be shown that any function $\mathcal{F} : \{0,1\}^n \rightarrow \{0,1\}$ can be implemented with a network of McCulloh Pitts neurons of two layers [58], [59]. The other type of neural networks are recurrent neural networks, in which the connections do form a cycle. The output of a node is fed back to the input and is a form of recursive computation. In addition to the interconnection between the nodes, the temporal step at the current state of the network have to be also taken into account for computing. In the following sections, we will delve deeper into both types of neural networks and derive a lower bound on the dissipation associated with learning and computing desired functions.

5.2 Feedforward Neural Networks: Perceptron

As defined previously, a feedforward network is a type of artificial neural network wherein connections between the neuronal nodes do not form a cycle and the information moves in only one direction - forward, from the input nodes, through the hidden nodes (if any) and to the output nodes. The simplest kind of feedforward neural network are single layer perceptron networks in which the inputs are fed directly to the outputs via a series of weights. They can be constructed with the McCulloh-Pitts neurons explained in the previous section. The sum of the products of the weights (along the edges between the nodes) and the inputs is added together in each node, and if the value is above the threshold μ, the neuron unit fires and produces the

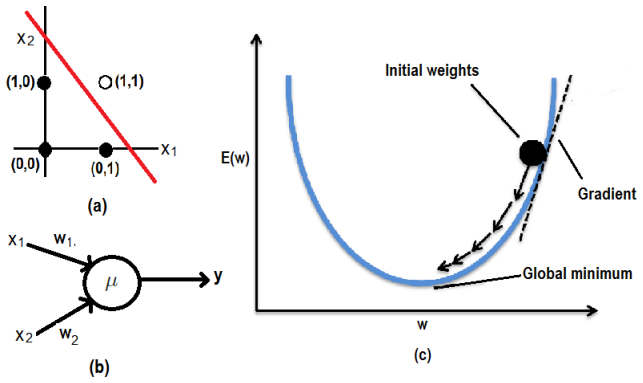

Figure 5.1. (a) AND gate as classification task. (b) Single node simple perceptron with inputs (x_1, x_2), weights (w_1, w_2) and output y required to perform the AND classification task. (c) Gradient descent on the squared error function E to obtain the global minimum.

output value y of '1', otherwise it takes the deactivated value '0' [58]; as indicated below

$$y = \begin{cases} 1 & : w \cdot x - \mu \geq 0 \\ 0 & : w \cdot x - \mu < 0 \end{cases}$$

where $w \cdot x = \sum_i w_i x_i$ is the dot product between the weights $\{w\}$ and inputs $\{x\}$, and μ is the bias.

Perceptron neural networks are linear classifiers and can be used to perform simple classification tasks if the labeled data vectors (collection of inputs $\{x\}$ and the correct classification outputs d) are linearly separable, like in an AND or OR gate. If the vectors are not linearly separable, we cannot find the right weights for which all vectors are classified properly. The most famous example of the single-layer perceptron's inability to solve problems with linearly non-separable vectors is the Boolean exclusive-or problem. This however can be solved by using a multilayer perceptron, trained with backpropagation algorithms.

Training and Testing Phase in Perceptron Networks

In supervised machine learning techniques, there are two major phases called training and testing. The labeled dataset is divided into two parts, one for each phase. In the training phase, the training data is used to train the weights in the neural network to perform the required classification task. The weights $\{w\}$ of the network are trained by a simple learning algorithm, implementing a form of gradient descent on an error function E (using a step size determined by the learning rate parameter η). There are a wide range of error functions like sqaured error, log likelihood, cross entropy and distance metrics that are commonly used. We will use the squared error between the calculated output y and sample output data d to create an adjustment to the i-th weight at time t, to produce the new weight as shown below [58]. The training phase is completed when the error function is below an acceptable threshold, and we move on to the testing phase.

$$E = \tfrac{1}{2}(d-y)^2$$

$$w_i(t+1) = w_i(t) + \eta \tfrac{dE}{dw_i} \rightarrow w_i(t+1) = w_i(t) + \eta(d-y)x_i$$

The trained perceptron is then applied on the unseen testing data to determine how well the network performs. It is important to reteirate that the weights are not updated during this phase. If the test error is below an acceptable value, then the network is ready to be used for the classification task. However if the test error is above this threshold, we have to retrain the network with changed parameters and datasets to ensure success on the next attempt. We will now build towards a fundamental lower bound on dissipation in the next section for this type of neural network.

5.3 Lower Bound on Dissipation in FeedForward Perceptron

Comprehensive analyses of neuromorphic implementations of perceptron networks should include the energy costs of training the network, as well as the costs of classification by the trained networks in the testing phase. To obtain fundamental lower bound on dissipation costs associated with training and testing phases, we will describe both the feedforward networks and the discretized weights (the weights are discretized by being ultimately realized as a state in a physical system) as a deterministic finite-state automata (FSA), and use the FSA formulation from the previous chapter to obtain the lower bound on dissipation associated with training and testing these networks.

5.3.1 Training Phase

Initially in the training phase, the weights can be randomly initialized (preferably to small values), and the labeled data containing inputs and their corresponding outputs, is used to obtain weights that minimize the error function E. This entails using the inputs and the current weights to generate the calculated output, and then use gradient descent on the error function to change the weights. Thus the cost of training the network should include the cost of generating the outputs, as well as the cost of using the outputs to train the network. In a neuromorphic system, the neural network nodes are physically instantiated in the states of the system $\mathcal{S}$ and the generation of the output using the weights and the training data can be described as state transitions of a FSA. In order to be able to differentiate between different node and weight values, we assume that the physical states instantiating the neural network are perfectly distinguishable. The lower bound on dissipation in the the t-th training step to generate the output is given by

$$\Delta\langle E_t^{\mathcal{B}}\rangle \geq k_B T \ln(2) \left[H(S^{(t)}) - H(S^{(t+1)}|R_t)\right] \tag{5.1}$$

where $H(S^{(t)})$ is the Shannon entropy of the state distribution associated with the neural nodes (where the neural node states correspond to the joint state of the network given that each node can take one of two binary states), before the t-th training step, $H(S^{(t+1)}|R_t)$ is the conditional entropy of the neural node state distribution after the output generation (given R_t, the training data used to train the weights at time t). k_B is the Boltzmann constant and T is the temperature of the thermal bath $\mathcal{B}$.

If the discretized weights of the neural network are also physically instantiated in the physical states of a system $\mathcal{W}$, like in a neuropmorphic system in which the weights are realized as the memristance value of the memristor synapses [63], then the training of the weights can be described as state transitions of another deterministic FSA (Fig. 5.2), with the transitions being dependent on the training learning rate parameter η and training sample distribution. We have the bias fixed at $\mu = 1.5$ here, but it can be defined as an additional weight with a constant input value of 1, and the optimal value can be obtained using the same gradient descent procedure as before. There is a clear difference in the FSA, with a change in the learning parameter (Fig. 5.2). For a large value of $\eta = 2$, we see that the gradient descent procedure cannot find the optimal set of weights to minimize the error function. For the lower value of $\eta = 0.5$, the optimal weights are $(1, 1)$ and the perceptron is trained to perform the AND classification task.

We assume that the physical states instantiating the weights are perfectly distinguishable, and the lower bound on dissipation for training the weights in the t-th training step is

$$\Delta\langle E_t^{\mathcal{B}}\rangle \geq k_B T \ln(2) \left[H(W^{(t)}) - H(W^{(t+1)}|R_t)\right] \tag{5.2}$$

where $H(W^{(t)})$ is the Shannon entropy of the weight distribution before the t-th training step, $H(W^{(t+1)}|R_t)$ is the conditional entropy of the weight distribution after the weight update (given R_t, the training data used to train the weights at time t). In

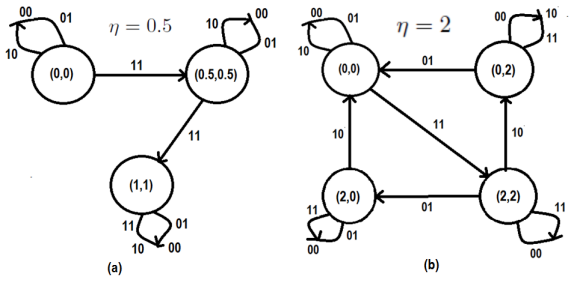

Figure 5.2. (a) State transitions for learning rate $\eta = 0.5$. $(w_1, w_2) = (1, 1)$ with a bias of $\mu = 1.5$ is the trained value of weights that performs the classification properly. (b) State transitions for learning rate $\eta = 2$. There is no single value of the weights with this η that achieves proper classification.

this paper, we will focus more on this cost of training the weights, once the output has been generated and less on the cost of generating the output itself. The lower bound on the total cost of training the weights in the training phase $\Delta\langle E^{\mathcal{B}}_{total}\rangle$, is obtained as the sum of lower bounds over N different time steps $\Delta\langle E^{\mathcal{B}}_{total}\rangle = \sum_{t=1}^{N} \Delta\langle E^{\mathcal{B}}_{t}\rangle$. In section 5.6, we will describe the variation of $\Delta\langle E^{\mathcal{B}}_{total}\rangle$ for training the weights over multiple time steps with varying values of η and different training data distributions. We next move to the testing phase.

5.3.2 Testing Phase

As stated earlier, in the testing phase the trained weights are not changed anymore and the only dissipation cost is associated with generating the outputs on the test dataset. Once again, the generation of the output values by the physical instantiations of the neural network nodes can be described as FSA state transition [64], and the lower bound on dissipation in the m-th testing test is given as

$$\Delta\langle E^{\mathcal{B}}_{m}\rangle \geq k_B T \ln(2) \left[H(S^{(m)}) - H(S^{(m+1)}|R_m)\right] \tag{5.3}$$

where $\mathcal{R}_m$ now refers to the input from the test data used in the m-th testing step. In the AND classification task, once the correct weights have been learned, the lower

bound on dissipation for an uniform distribution on the test data inputs is equal to $0.562kT$ per time step. This is lower than the value of $0.824kT$ per time step, calculated using Landauer's principle for a physical realization of an AND gate with an uniform distribution of inputs.

5.3.3 Lower Bound on Dissipation for Varying Learning Rates

To ensure perceptron convergence to optimal weights, the learning rate parameter is reduced according to a learning rate schedule. These include linear and exponential down scaling of the learning rate, independent of the input information. There are also adaptive learning rate techniques like Newton's Hessian method, ADAGRAD and ADADELTA [65] which use first and second order information from the inputs to continuously adapt the learning rate. The choice of the initial learning rate and the subsequent schedule is of extreme importance while training a network for a task. If the learning rate is made too small, a very large number of steps are required to converge to the optimal weights. On the other hand, if the rate is too large then weights might not converge at all as seen in the case of $\eta = 2$ from the previous section. It is thus important to study the effect of changing the learning parameter during training on the dissipation lower bound. Since the FSA state transitions depend upon the value of the learning rate parameter η, changing it's value changes the FSA description of the weights and the associated dissipation. The change in the FSA description of the discretized weights when η is changed from 0.5 to 1 during training is plotted in Fig. 4. We will work under the assumption that the learning parameter is not physically instantiated and changed once externally during training, independent of the inputs. It is also required that the physical system $\mathcal{W}$, in which the discretized weights are instantiated in, have enough number of distinguishable states to accommodate all the different values of weights that are generated as the learning rate is changed. Comparison of the lower bounds in dissipation for the case

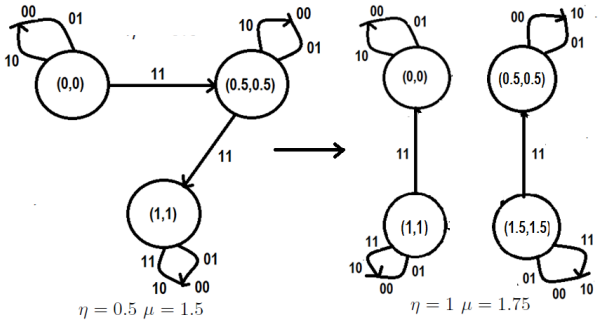

Figure 5.3. FSA description of the neural network weights during training with $\eta = 0.5$ and $\mu = 1.5$, and of the weights after the η and μ are changed during the training phase to 1 and 1.75 respectively.

of fixed η, and the case where η is changed is presented in the results section (Fig. 7).

While we have described the learning parameter being changed externally, it is important to note that in neuromorphic computing systems, the learning parameter might be physically instantiated in a register system and can be changed according to an adaptive schedule. The cost in this case would also include the dissipation associated with this physical instantiation, and will be discussed in detail in future works.

5.4 Recurrent Neural Networks - Hopfield & Boltzmann Networks

Recurrent neural networks are those neural networks where the connections between neuronal units form a directed cycle. They exhibit dynamic temporal behaviour and can use their internal memory to process an arbitrary sequence of inputs. We will introduce here two very popular recurrent networks - Hopfield and Boltzmann networks, in this section and explore the dissipation cost associated with their use as content addressable memories.

5.4.1 Hopfield Networks

Hopfield network is a form of an artificial neural network that was popularized by John Hopfield in 1982. It serves as a robust content addressable memory with binary threshold nodes. The activation value of the neuronal units are thus +1 for firing, and −1 for not firing instead of a 1 and 0. The next state of a nodal unit i - x_i is given as

$$x_{i+1} = \begin{cases} 1 & : w_i \cdot x - \theta_i \geq 0 \\ -1 & : w_i \cdot x - \theta_i < 0 \end{cases}$$

where $w_i \cdot x = \sum_{j=1}^{N} w_{ij} x_j$ and θ_i is the bias of unit i [58]. The Hopfield network consists of N completely coupled units, i.e each unit is connected to every other unit except itself. The network is symmetric i.e $w_{ij} = w_{ji}$ and $w_{ii} = 0$.

The energy function E of the state $x = \{x_1, x_2, ..., x_N\}$ can be defined as

$$E(x) = -\tfrac{1}{2} \sum_{j=1}^{N} \sum_{i=1}^{N} w_{ij} x_i x_j + \sum_{i=1}^{N} \theta_i x_i$$

For learning n patterns, we can calculate the necessary weights of the network that will minimize the energy function given above by solving for the equation $\frac{dE}{dw_{ij}} = 0$. This gives us

$$w_{ij} = \tfrac{1}{n} \sum_{k=1}^{n} \epsilon_i^k \epsilon_j^k$$

where ϵ_i^k is the i-th bit of the k-th pattern. The above learning rule is called the Hebbian learning rule, named after Donald Hebb who proposed a synaptic mechanism for learning in the brain where *"increase in synaptic efficacy arises from the pre-synaptic cell's repeated and persistent stimulation of the post-synaptic cell."* [58] [59]. This can be summarized as "Cells that fire together, wire together." For the weights set to these values, the energy of the network is minimized at any one of the k patterns.

The Hopfield model is isomorphic to an Ising model for magnetism at zero temperature. The Ising model can be used to describe those systems made of particles capable of adopting one of two states. The atoms in ferromagnetic material can be modeled as particles of either spin 1/2 (up) or spin 1/2 (down). The spin points in the direction of the magnetic field. All the atoms interact with each other, and this interaction causes some of the atoms to flip their spin until equilibrium is reached and the total magnetization of the material (which is the sum of the individual spins) reaches a constant level. Under these conditions, we can show that the energy function for the Ising model is isomorphic or has the same form as the energy function of Hopfield networks. The potential energy of certain state $(x_1, x_2, ..., x_n)$ of an Ising system is of the form

$$E = -\tfrac{1}{2}\sum_{1,j}^{n} w_{ij}x_i x_j + \sum_{i}^{n} -h^* x_i$$

where w_{ij} represents the magnitude of the magnetic coupling between the atoms labeled i and j, and h^* is the external field. The two systems are equivalent dynamically, but only in the case of zero temperature, when the system behaves in a deterministic manner at each state update.

In the Hopfield model, each individual unit preserves its state until they are asynchronously selected for an update. Under these update dynamics, the network is guaranteed to converge to the minimum of the energy function, and the state of the network will not change after that. However Hopfield networks suffer from the problem of multiple spurious local minima in which the system might find itself stuck in and unable to retrieve the required stored pattern. The number of patterns that can be stored faithfully is dependent on the number of neurons (N) and their connections. It was shown that the capacity of these networks was about 0.138N (approximately 138 vectors can be recalled from storage for every 1000 nodes) for the Hebbian learning rule [58]. Therefore, mistakes will occur if one tries to store a large number of vectors exceeding this capacity. Perfect recalls and higher capacity of $> 0.14N$,

can be achieved in the network by using the Storkey learning method [69]. In the following section, we will consider Boltzmann machines, which uses a time-varying pseudo-temperature parameter and stochastic state updates as in the full Ising model to overcome some of the limitations of the Hopfield networks.

5.4.2 Boltzmann Networks

In order to avoid local minima and arrive at the global one, noise is introduced into the dynamics of the network. As the system converges to states of lower energy, transitions to higher energy state are occasionally allowed to help skip out of local minima, in a statistical sense. Neural network models with such stochastic dynamics are called *Boltzmann machines* [58]. The Boltzmann machine is thus an Hopfield network consisting of N units $x_1, x_2, ..., x_N$, in which each unit is updated asynchronously using the update rule

$$x_i = \begin{cases} 1 & \text{with probability } p_i \\ -1 & \text{with probability } 1 - p_i \end{cases}$$

$$p_i = 1/\left[1 + exp\left(\frac{-(w_i \cdot x - \theta_i)}{T_p}\right)\right]$$

where T_p is a positive constant and a measure of the noise introduced, often referred to as *pseudo-temperature* (not to be confused with the actual physical temperature). The energy function of the Boltzmann network is the same as the energy function of the Hopfield network, and with the use of non-zero temperatures and stochastic updates, it is closer to the full Ising model. For extremely small values of $T_p \approx 0$, the Boltzmann network will behave like an Hopfield network. In order to achieve the required final states in a Boltzmann network from the initial states, we use *simulated annealing*. This approach takes it's inspiration from the real-life phenomenon of annealing processes, used to form crystals. The updates are started with a high pseudo-temperature T_p, and reduced according to an annealing schedule. As the

temperature $T_p \rightarrow 0$, the system state distribution is heavily concentrated in the global minimum. The choice of a good annealing schedule, according to which T_p varies over time is very important. We will explore the exponential and logarithmic annealing schedules in this paper.

$$\text{Exponential: } T_p^n = T_p^0 \alpha^k \text{ with } 0.8 < \alpha < 0.9$$

$$\text{Logarithmic: } T_p^n = \frac{T_p^0}{1+\alpha \log(1+k)} \text{ with } \alpha > 1$$

5.5 Lower Bound on Dissipation in Hopfield and Boltzmann networks

The analysis of Hopfield and Boltzmann neural networks in this paper will only focus on the dissipation incurred in the use of these networks as a content addressable memory (CAM). We will not focus on the costs associated with the learning the necessary weights in this case. There are closed form expressions for the weights using the Hebbian rule, that can be calculated directly (without the need for training) for these networks depending upon the input pattern that needs to be stored and appropriately instantiated [70]. The network of N units is realized in the physical system $\mathcal{S}$, and is a deterministic FSA (in the case of Hopfield networks) and probabilistic FSA (in the case of Boltzmann networks) of 2^N states. In both Hopfield and Boltzmann networks, the weights are initialized at the required values depending on the target pattern, and they remain fixed throughout the entire update process. The initial distribution of the patterns are instantiated in the states of $\mathcal{S}$. The signal asynchronously updating one of the N units at each time instant t will be instantiated as the referent $\mathcal{R}_t$. The update dynamics are implemented using unitary Hamiltonians that evolve the current state of the network to the required next state according to the weight-dependent transition rules. The system $\mathcal{S}$ is in contact with a thermal bath at temperature T. If we assume that the physical states of the network are orthogonal pure states i.e.

distinguishable, the t-th network update can be modeled as a FSA state transition and the lower bound on dissipation is given by

$$\Delta\langle E_t^{\mathcal{B}}\rangle \geq k_B T \ln(2) \left[H(S^{(t)}) - H(S^{(t+1)}|R_t)\right]$$

where $H(S^{(t)})$ is the Shannon entropy of the state distribution of $\mathcal{S}$ at time t, and $H(S^{(t+1)}|R_t)$ is the entropy of the updated state of $\mathcal{S}$ conditioned on the update signal $\mathcal{R}_t$. The lower bound on the total dissipation $\Delta\langle E_{total}^{\mathcal{B}}\rangle$ for the entire process as the network evolves from an initial distribution to the final state corresponding to a minima is

$$\Delta\langle E_{total}^{\mathcal{B}}\rangle = \Delta\langle E_0^{\mathcal{B}}\rangle + \Delta\langle E_1^{\mathcal{B}}\rangle + ... + \Delta\langle E_t^{\mathcal{B}}\rangle$$

The 3 - node, 8 state neural network shown in Fig. 5 will be studied in this paper. The network has two stable low energy states, with the lowest being the necessary global minima. While the Hopfield network will get stuck in one of these two states, the problem is avoided by using simulated annealing in the Boltzmann network. The effect of node update policy and pseudo-temperature for this example network will be presented in the results section.

5.6 Illustrative Example & Results

The formulation from the previous sections are used to calculate the lower bound on total dissipation associated with feedfoward and recurrent neural networks. In Fig. 5.6, for the feedforward network from Fig. 5.2, we present the lower bound on the dissipation per time step in training for both generating the output, the subsequent weight change and the total cost of both for the first 70 training time steps when the learning rate is set to $\eta = 1$. We can see that as we progress through the training phase, the cost of the weight training increases, reaches a maximum and decreases as the right weights are learned. The cost of generating the outputs in each time

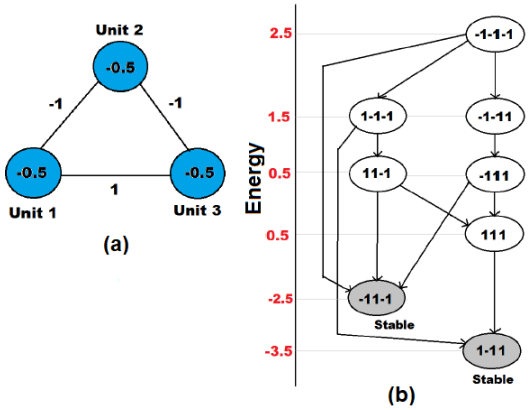

Figure 5.4. (a) Recurrent neural network with 3 nodes and the corresponding weights between the nodes. The thresholds for each of the nodes are also indicated within the node. (b) FSA state transition diagram of the neural network in (a), with 8 states and the two low energy stable states (gray).

step continuously increases and levels of at $0.811kT$ per time step once the correct weights are learned. The total cost over 70 time steps for generating the output is $36.31kT$, training the weight $1.4kT$ and the overall cost of training is $37.71kT$. For the remaining results in this paper, we will focus more on the cost of the weight changes, as this would constitute learning in such networks.

Lower bounds on the dissipation per time step for training the weights are shown for the first 60 time steps in Fig. 5.6(a). Results are shown for an uniform distribution of the training data, and values $\eta = 0.5$, 1, and 2 of the learning parameter, assumed fixed. Respective lower bounds on the total weight training costs $\Delta\langle E^{\mathcal{B}}_{total}\rangle$ are $11.2kT$, $4.2kT$ and $47.98kT$. The dissipation-per-training-step increases, reaches a maximum, and then decreases as the neural networks learn the correct weights from the training data. There is no fundamental lower bound on training cost once the right weights have been learned. For $\eta = 2$, the dissipation never decreases as this choice of learning parameter cannot find the optimal weights. $\eta = 1$ is a better choice of the learning parameter, since it allows for learning of the right weights and at a lower total dissipation than $\eta = 0.5$.

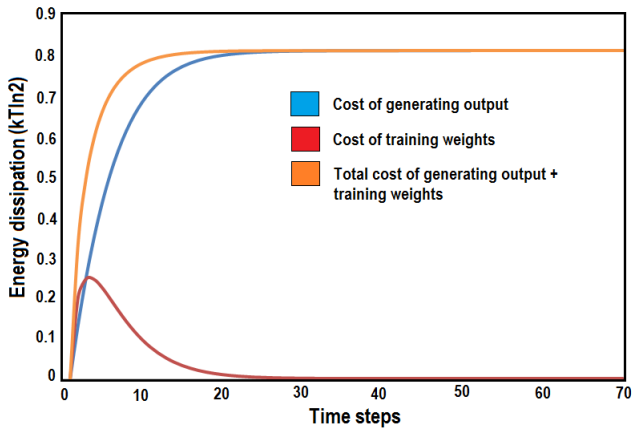

Figure 5.5. Dissipation lower bound on output generation, weight training and total cost over 70 time steps for simple perceptron learning the AND classification task at learning rate of $\eta = 1$.

The difference in the dissipation bound with change in training data distribution is explored in Fig. 5.6(b). There is a significant increase in $\Delta\langle E^{\mathcal{B}}_{total}\rangle$ as the probability of the 11 input is changed from 0.25 to 0.05. Since the input 11 is the only input that maps to the output 1, any decrease in its probability significantly reduces the chances of the right weights being learned, hence increasing the energy cost of training. These results indicate that the learning parameter and the training distribution can thus be changed in an optimal manner to balance learning of the correct weights, with the minimum energy dissipation associated with doing so.

In Fig. 5.6, we see the variation in the lower bound in distribution for different initial weight distributions over 30 time steps. We can see that the dissipation lower bound reduces as the initial starting distributions are more skewed towards the optimal weights. For the initial distribution of $p(+1,+1) = 0, 0.5, 0.33$ and 0.75 respectively, we have the corresponding total lower bounds on learning the optimal weights to be $5.0264k_BT\ln(2)$, $4.62k_BT\ln(2)$, $3.7238k_BT\ln(2)$ and $2.1433k_BT\ln(2)$ respectively.

In Fig. 5.6, we see the variation in the dissipation lower bound per time step between a fixed learning rate of $\eta = 0.5$ (dashed) and the case where the learning

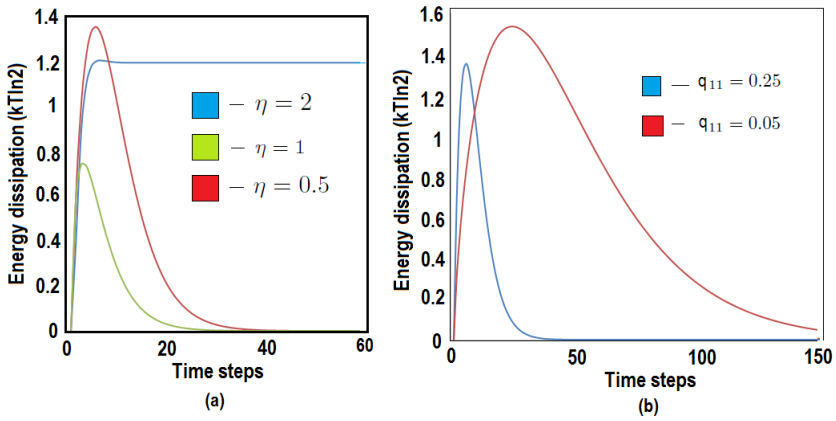

Figure 5.6. (a) Dissipation lower bound of simple perceptron learning the AND classification for 60 time steps, for different values of the learning parameter η during training. (b) Dissipation lower bound of same perceptron for 150 steps, for $\eta = 0.5$ and different training data distributions.

rate is changed externally from $\eta = 0.5$ to $\eta = 1$ (bold) during training. The total dissipation bound $\Delta\langle E^{\mathcal{B}}_{total}\rangle$ over 100 time steps for the fixed η case is $11.24kT$ and $7.28kT$ for the latter. Thus while the learning rate of $\eta = 0.5$ did allow for convergence towards the optimal weights, the value was too small and required more time steps to converge resulting in a larger total dissipation. When η is changed from 0.5 to 1, the new parameter has a greater step size and allows for a quicker convergence to optimal weights at a lower cost. If we were to decrease the learning rate from $\eta = 2$ to $\eta = 1$, the weights converge to an optimal value of $w = (1, 1)$ and the lower bound on total dissipation for changing the weights decreases significantly from the fixed case value of $60.83kT$ to $24.66kT$. Thus we see the effects of both increasing and decreasing the learning rate parameter on the associated total dissipation. Decreasing η is accompanied by reduction in the total dissipation bound, if the new value allows for the weights to converge to an optimal value that minimizes the cost function. Very small values of η will achieve convergence, but will result in an higher dissipation and must be properly tuned.

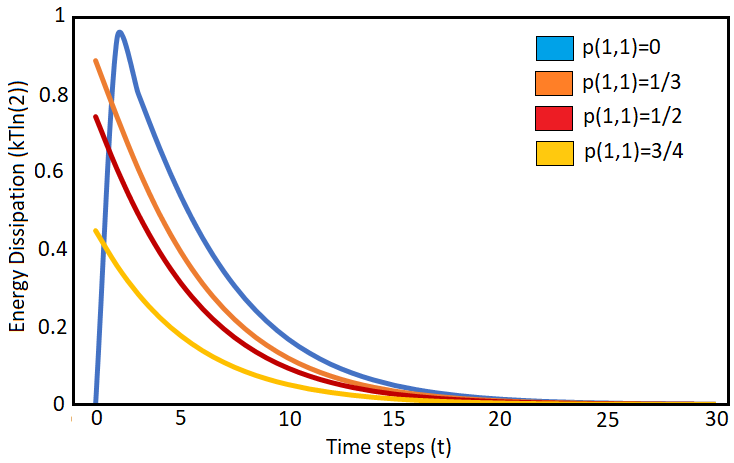

Figure 5.7. Dissipation bounds for different initial starting weight distributions over 30 time steps.

The results for the lower bound on dissipation in using the Hopfield network from Fig. 5.5 as a CAM, is shown in Fig. 5.6. This particular network contains two low energy stable states and thus the final steady state distribution should contain both the states with a non-zero probability. However the local minima can be avoided by using simulated annealing, which will be examined next. The network is initialized randomly and updated asynchronously, and the variation in the dissipation with respect to the update policies are studied. In the random update policy, all 3 nodes have an equal probability of being updated at any time step. In skewed policy - 1, all 3 nodes have a non-zero probability of being updated at any time step with one node being preferred over others, and in skewed policy - 2, only 2 nodes have a non-zero probability of being updated. The total dissipation over 100 times steps associated with each of those policies are $1.994kT$, $1.9926kT$ and $1.1158kT$ respectively. The dissipation decreases as the update policy changes from random to being skewed. Thus the choice of an optimal update policy can be made taking into account both the dissipation costs involved and the rate of convergence to the minima.

In Fig. 5.6, the lower bound on dissipation associated with different annealing schedules over multiple time-steps in a Boltzmann network are calculated. The net-

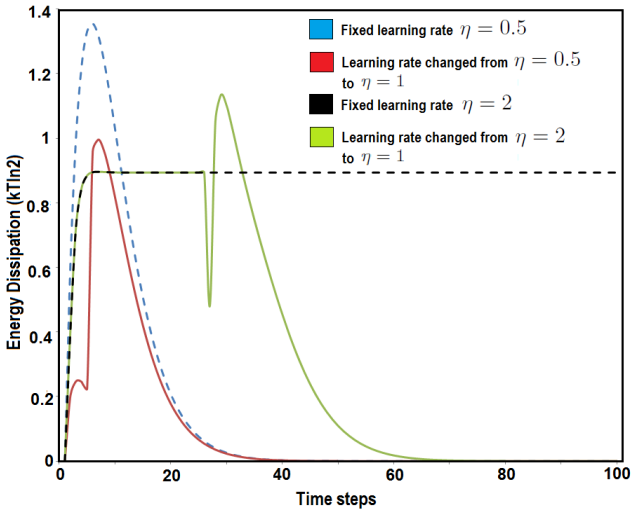

Figure 5.8. Dissipation lower bound for 100 time steps, for a fixed learning rate of $\eta = 0.5$ (dashed blue) and $\eta = 2$ (dashed black), and the case of learning rate changing from $\eta = 0.5$ to $\eta = 1$ (solid red) and learning rate changing from $\eta = 0.5$ to $\eta = 1$ (solid green).

work is initialized in a random state, and the starting pseudo-temperature is $T_p^0 = 5$. The two annealing schedules used to reduce the pseudo-temperature, that have been studied here are the exponential and logarithmic schedules. From the figure, it is very clear that the lower bound on dissipation in each of the time-steps are different for the two schedules. This is to be expected since the dissipation costs are dependent on the state-transition probabilities, and those are functions of pseudo-temperature and the annealing schedules. The sum of the dissipation bounds over the different time-steps will give us the total dissipation $\Delta\langle E^{\mathcal{B}}_{total}\rangle$ of the simulated annealing process using the respective schedule. $\Delta\langle E^{\mathcal{B}}_{total}\rangle \geq 5.775kT$ and $5.438kT$ over 30 time steps for the exponential and logarithmic schedules respectively. As in the case of the node update policy, a dynamic annealing schedule to optimally reduce energy dissipation can be evaluated.

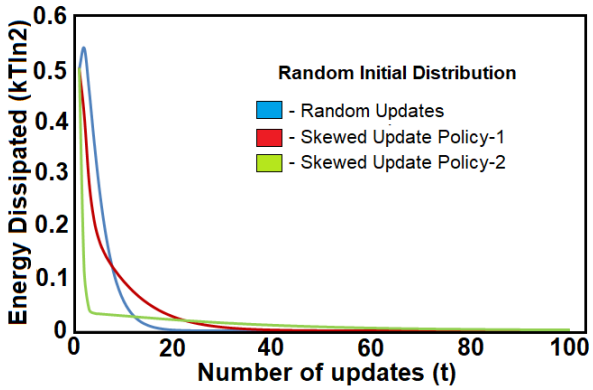

Figure 5.9. Variation in the lower bound on dissipation in an Hopfield network, with different asynchronous update policies over many 100 time steps. The random update policy is in blue, skewed update policy - 1 in red and skewed update policy - 2 in green.

5.7 Towards Thermodynamic Objective Functions

Machine learning techniques rely on performing some form of gradient descent to minimize a pre-defined cost function. The ones used in Hopfield networks were modeled after the total energy function in physical Ising spin systems [58] and emphasize the rich historical connection between machine learning and physics. While we used quadratic squared error as our cost function in this paper, a wide range of options are now available depending upon the task at hand. If the discretized weights in the neural network are modeled as a FSA, then the learning rule and the learning rate η schedule can be obtained from how the training data input is encoded in the weight states and the state transition mappings of the weights. We are interested in deriving learning rules from a physically grounded approach, and looking for the input encodings and transition mappings of the weights that will minimize the total dissipation cost of training the weights $\Delta\langle E^{\mathcal{B}}_{total}\rangle$ (we will ignore the cost of generating the outputs for now). In order to ensure that we do not get only trivial solutions, we will impose a memory constraint on the weights. From the principle of optimality, we have that the dynamic state encodings and transition mappings between the weight states has to be optimal for every time step, in order to be optimal for the entire

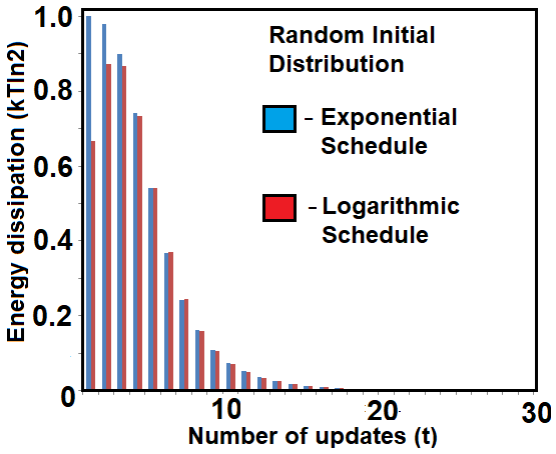

Figure 5.10. Lower bound on dissipation for simulated annealing in a Boltzmann network with 3 neural nodes over 30 time steps. The dissipation in each time step is compared between the exponential (blue) and logarithmic (red) annealing schedules.

training phase. Thus the problem of finding the optimal state encoding to minimize the total dissipation requires finding the optimal solution to minimize the dissipation at the t-th step.

If the discretized weights of the neural network are physically instantiated in the distinguishable states of a system $\mathcal{W}$, then from Theorem-3 from section 4.7, the lower bound on dissipation in each time step of training the weights with temporally correlated inputs can be written as

$$\Delta\langle E_t^{\mathcal{B}}\rangle \geq k_B T \ln(2) \left[-\Delta H^{\mathcal{W}} + \mathcal{I}^{\mathcal{R}_1\mathcal{W}'} - \mathcal{I}^{\mathcal{R}_1\mathcal{W}}\right]$$

where $\mathcal{I}^{\mathcal{R}_1\mathcal{W}}$ and $\mathcal{I}^{\mathcal{R}_1\mathcal{W}'}$ is the Shannon mutual information between the latest training data input $\mathcal{R}_1$ and the weights $\mathcal{W}$ before and after the t-th training step. $-\Delta H^{\mathcal{W}}$ is the difference in Shannon entropy of the state distribution of the weights, before and after the training step. It is evident from the above expression that the fundamental lower bound on dissipation for training the neural network weights according to the learning rules is zero, once the weights that minimize the chosen cost function have been learned. In order to obtain the optimal state encoding $p(w_{(t-1)}^{\mathcal{W}}|i^{\mathcal{R}_0})$ of the past inputs $\mathcal{R}_0$ in the weights $\mathcal{W}$, to minimize $\Delta\langle E_t^{\mathcal{B}}\rangle$, we construct the following

Lagrangian $\mathcal{L} = \Delta\langle E_t^{\mathcal{B}}\rangle + \beta(\mathcal{I}^{\mathcal{R}_0\mathcal{W}} - I_t)$, where $\mathcal{I}^{\mathcal{R}_0\mathcal{W}} = I_t$ is the memory constraint at time t, and β is the dissipation-memory tradeoff parameter. The solution can be obtained by solving the following constrained optimization problem.

$$\text{Minimize}_{p(w_{(t-1)}^{\mathcal{W}}|i^{\mathcal{R}_0})}\{\Delta\langle E_t^{\mathcal{B}}\rangle + \beta(\mathcal{I}^{\mathcal{R}_0\mathcal{W}} - I_t)\}$$

Under simple assumptions, we can show that solving this problem is equivalent to maximizing $\mathcal{I}^{\mathcal{R}_1\mathcal{W}} - \beta(\mathcal{I}^{\mathcal{R}_0\mathcal{W}} - I_t)$, with respect to the state encoding $p(w_{(t-1)}^{\mathcal{W}}|i^{\mathcal{R}_0})$. This is the Information Bottleneck algorithm discussed in section 2.3.1 of this dissertation and has been widely used in clustering problems [140], predictive inference and deep learning [67]. The ideas presented above clearly elucidate the deep connection between the physical cost of energy dissipation and learning algorithms. A detailed discussion of using thermodynamics as the central concept of learning, is extremely necessary and the focus of this dissertation in the next two chapters, as well as [68].

5.8 Conclusion

In this chapter, the fundamental energy costs of training in different types of neural networks were explored in a framework that describes physical implementations of such networks (with discretized weights) as FSA. Two types of networks were studied - feedforward perceptrons, and recurrent Hopfield and Boltzmann networks. The fundamental lower bounds on energy dissipation were calculated for a simple perceptron, learning the AND classification task. This was followed by an analysis of the dissipation costs associated with the use of Hopfield and Boltzmann networks as content addressable memory. This physically grounded approach has provided fundamental bounds on the dissipative costs necessarily incurred, that are independent of implementation details. While focused on simple networks in this chapter, the FSA description of neural networks are more general and can be applied to a wider class of systems like multilayer neural networks. These bounds are an important first

step towards determining the ultimate performance limits of neuromorphic systems and identifying sources of inefficiency. Identification of neural network learning algorithms that minimize the dissipative cost of training were also discussed. This final concept will influence the work in the next couple of chapters where we analyze the fundamental connections between physical intelligence and thermodynamics.

CHAPTER 6

A THERMODYNAMIC TREATMENT OF INTELLIGENT SYSTEMS

6.1 Introduction

In this age of big data, the computing industry has shifted its focus towards tasks that involve handling and learning from data. The tremendous progress made in the field of machine learning to perform numerous learning tasks has been due to two major driving factors. The first is the emergence of extremely sophisticated learning algorithms for supervised learning and reinforcement learning techniques. However a large number of these algorithms achieve learning by performing gradient descent on a task dependent energy or loss function, and are problem specific and narrow in applicability. They also require significantly large labeled datasets to train them. The community has now sets its target in improving the understanding of unsupervised learning, and development of cost functions that are applicable over a large range of tasks. The other is the availability of powerful specialized hardware like graphic (GPU) and tensor (TPU) processing units have made realization of these resource intensive algorithms feasible. As we approach the physical limits to scaling and dissipation, we have started to look away from these conventional computing devices and architectures as solutions to our new computing tasks. Understanding how complex biological systems are able to process and learn from data will improve our ability to build intelligent systems. In the previous chapter, I presented results on the lower bounds on dissipation in neural networks and introduced the concept of using energy dissipation as the sole objective function to be optimized for learning

algorithm in neural networks. In this chapter, I will look to extend that idea further into a more general picture and obtain the thermodynamic constraints under which learning emerges in a physical system.

In the recent past, there has been increased research into developing fundamental relationships between thermodynamics, information theory and neurobiological systems [71],[72]. Since intelligent processes are physical as well, extending such work to establish the thermodynamic conditions under which a physical system exhibit learning capabilities would be extremely beneficial in the design and fabrication of intelligent systems, and usher in the new paradigm of *thermodynamic computing* [74]. In Section 6.2 and 6.3, clear definitions of the different terms that will be used in this paper will be provided. Recent progress made in fluctuation theorems to describe driven non-equilibrium systems will be presented in section 6.4 and their implications discussed. In section 6.5 - 6.7, the fundamental lower bound on dissipation for a physical system implementing a finite state automata over a state transition is used to analyze two important concepts associated with intelligence - adaptive learning and predictive inference. In section 6.8, the results from the previous sections are extended to study active agents with the ability to act on their environment. The chapter concludes in section 6.10 with a brief discussion of the results and what they entail for the future.

In this chapter and through the rest of this dissertation, we will define *intelligence* as comprising of two important and distinct phenomenon - one that involves the accumulation/learning of information from the environment and other which is the use of this accumulated information to predict future inputs. We will refer to these two components as adaptive learning and predictive inference respectively and discuss them in detail in later sections. As this chapter seeks to bridge the gap between thermodynamics, information theory and the learning capabilities of complex systems, it is important to clearly introduce some of the concepts that will be used in this

chapter. These will include definitions for complex systems, measures used to quantify complexity, self-assembly and self-organization processes and thermodynamics.

6.2 Complex Systems & Complexity

Complex systems are those systems in which "large networks of highly interacting non-linear components with no central control, simple rules of operation give rise to complex collective behavior, sophisticated information processing, and adaptation via learning or evolution" [139]. This makes them and their relationship to their environment difficult to model. They cal also be defined as a system "that exhibits notrivial emergent and self-organizing behaviours. Good examples of such systems include biological organisms, cellular automata, cities, financial markets, world wide web and artificial neural networks. Significant research has been carried out on characterizing the complexity of such systems and understanding their behavior.

While there are some distinct properties of complex systems, it is not yet very clear on how to best quantify this complexity in a complex system. A number of measures have been suggested, but there is not a single universally accepted metric of complexity. A good measure of complexity must be low in the cases of high amounts of order and randomness, and higher for a suitable combination of the two. A non-exhaustive list has been provided in [**?**], some of which are - size, entropy, algorithmic information content, logical depth, thermodynamic depth, fractal dimension, computational capacity and statistical complexity. In this chapter, we will us statistical and information-theoretic measures to measure the complexity of those systems of interest [**?**]. The statistical complexity of the system measures the minimum amount of information about the past behavior of the system that is required to predict the future behavior of the system. For a system $\mathcal{A}$ interacting with external signals $\mathcal{R}$, the mutual information $\mathcal{I}^{\mathcal{RA}}$ can be used to calculate this amount of information. A

more complex system will require greater correlation (memory) of the signals in the past the system interacted with to predict the future.

6.3 Self-assembly & Self-organization

Self-organization and self-assembly are terms that have become extremely popular across different fields in the recent years. While both describe processes that can give rise to collective order from dynamic small-scale interactions, they often have different meanings ascribed to them and worse, even used interchangeably. Since biological systems are self-organized, it is necessary to briefly define the process and distinguish it from self-assembly. From [139], self-organization (often referred to as dynamic self-assembly) is defined as "*a dissipative non-equilibrium order at macroscopic levels, because of collective, non-linear interactions between multiple microscopic components. This order is induced by interplay between intrinsic and extrinsic factors, and decays upon removal of the energy source. In this context, microscopic and macroscopic are relative*" (Fig. 6.1(b)).

Self-assembly is a "*non-dissipative process that produces structural order on a macroscopic level, because of collective interactions between multiple (usually microscopic) components that do not change their character upon integration into the self-assembled structure. This process is spontaneous because the energy of unassembled components is higher than the self-assembled structure, which is in static equilibrium, persisting without the need for energy input*" (Fig. 6.1(a)). According to the second law of thermodynamics, for any spontaneous process that occurs in a closed system, the increase in entropy translates to a decrease in free-energy $\Delta F = \Delta E - T\Delta S < 0$. Self-assembly processes are characterized as free-energy minimization, with the minimum value attained at equilibrium (where $\Delta F = 0$). On the other hand, self-organized systems are generally not in equilibrium and are characterized by exchange of both matter and energy with the environment. These processes are characterized

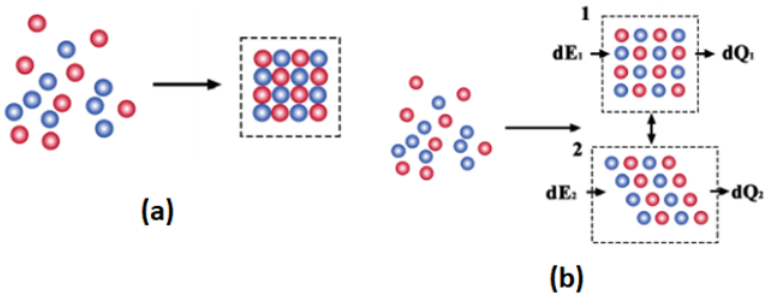

Figure 6.1. (a) Self-assembly process characterized by no external driving fields and the spontaneous evolution to the equilibrium state of minimum free energy. (b) Non-equilibrium self-organization process in which the external fields produces different structures. The process is dissipative and the system loses its order when the external energy source is removed.

by non-equilibrium thermodynamics which will be discussed in detail in the next section.

6.4 Non-equilibrium Thermodynamics & Fluctuation Theorems

Non-equilibrium thermodynamics is the branch of thermodynamics that deal with systems that are not in thermodynamic equilibrium, but can be described using variables that are used to describe equilibrium systems. Almost all systems, including self-organized systems found in nature are not in equilibrium, for they change over time and subject to flux of matter and energy to and from other systems. Hence the thermodynamic study of non-equilibrium systems requires more general concepts than those that are dealt with by equilibrium thermodynamics.

The Crooks fluctuation theorem represent an important breakthrough in the field. The theorem establishes the relationship between the relative likelihoods of different dynamical paths or trajectories that the microstates of a non-equilibrium system could traverse, to the entropy production associated with those trajectories [135]. Us-

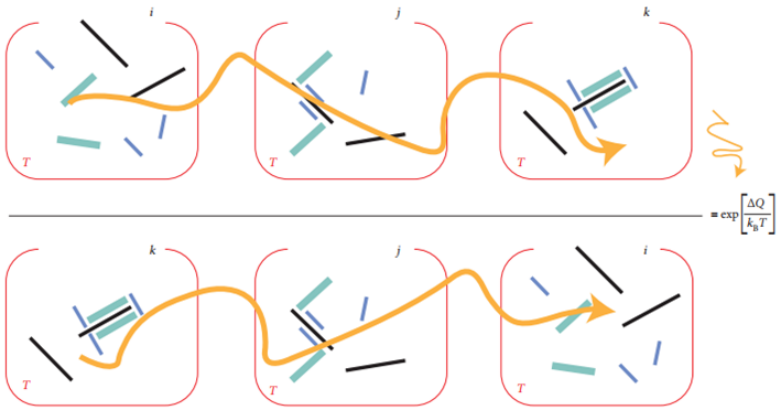

Figure 6.2. The Crooks Fluctuation theorem provides a quantitative relationship between the likelihoods of the forward and reverse trajectory of microstates when driven by an external field with the heat dissipated ΔQ into the thermal bath as the system traverses the trajectory.

ing time-reversal symmetry and conservation of energy, Crooks derived the following relationship

$$\frac{\pi(\gamma)}{\pi(\gamma^*)} = exp\left[\frac{\Delta Q(\gamma)}{k_B T}\right]$$

where the left hand side is the ratio of the relative likelihoods of a certain trajectory γ of microstates (a sequence of microstates over time) to its time reversed trajectory γ^*, and $\Delta Q(\gamma)$ is the heat dissipated into the thermal reservoir (at temperature T) as the system traverses trajectory γ and shown in Fig. 6.2. The relationship above indicates that a certain forward trajectory is more likely than the time reversed one by an exponential factor of the heat $\Delta Q(\gamma)$. The relationship is extremely powerful as it holds even in the presence of external fields driving the system. The Jarzynski equality can be seen as a special case of the Crook's fluctuation theorem, and given as

$$e^{-\Delta F/k_B T} = e^{-\bar{W}/k_B T}$$

where ΔF is the free energy difference between two states, and $\bar{W}$ is the average work done on the system. Applying Jensen's inequality ($\phi(E[X]) \leq E[\phi(X)]$ for a

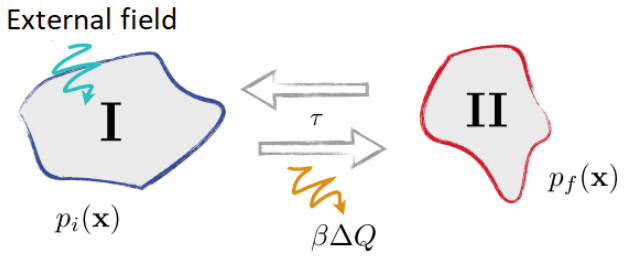

Figure 6.3. The macrostate fluctuation theorem quantifies the relationship between the likelihood of driving a system in macrostate I (with microstate distribution $p_i(x)$) to macrostate II (with microstate distribution $p_f(x)$) in time τ with the internal entropy change in the system and the heat dissipated ΔQ into the bath [80].

random variable X and convex function ϕ) to the above equality, we get the second law of thermodynamics $\Delta F \leq \bar{W}$. The Jarzynski and Crooks theorems have both been verified using biomolecular and simulation experiments.

While Crook's Fluctuation theorem dealt with microtrajectores in systems, England developed a generalization of this relationship to understand the likelihood ratios associated with transition between macrostates (shown in Fig. 6.3) by integrating over all microstates under a macrostate and over all relevant trajectories. These were used to study self-organized systems in [79] and [80], where the authors discuss using the relationship to develop a dissipation-driven theory of adaptation. The result central to these papers is the following equation, which relates the statistics of arbitrary macro-observables to the dissipation.

$$\frac{\pi(II^* \to I^*)}{\pi(I \to II)} = \langle e^{\ln\left[\frac{p_f(j|II)}{p_{in}(i|I)}\right]} \langle e^{-\beta \Delta Q_{i\to j}} \rangle \rangle_{I \to II} \tag{6.1}$$

I and II correspond to a macro-observable of the system driven by an Hamiltonian at two different times. Each of these macro-observables correspond to a set of microstates $\{i\}_I$ and $\{j\}_{II}$ respectively. $p_0(i|I)$ and $p_f(j|II)$ are probability distributions of the microstates given the different macroscopic variable values of I and II respectively at the two different time instances. $\langle \ln \left[\frac{p_f(j|II)}{p_{in}(i|I)}\right] \rangle_{I \to II}$ is the difference in internal

entropy of the system and we will denote it as ΔS. $\langle\langle e^{-\beta \Delta Q_{i \to j}} \rangle\rangle_{I \to II}$ corresponds to the exponential of the dissipation associated with transitions from microstate i in I, to microstate j in II averaged over all trajectories from i to j (the average calculated by the inner expectation bracket) and over all microstates $i \in I$ to $j \in II$ (calculated by the outer expectation bracket). $\beta = 1/k_B T$, where k_B is the Boltzmann constant and T is the temperature of the thermal bath that the system is in contact with. $\pi(II^* \to I^*)$ is the probability of the reverse process, if the system is driven by the time-reversed Hamiltonian of the forward process.

England focused on the non-equilibrium constraints, and studied the conditions under which the system is more likely to be in one macrostate over another, and used it to propose a dissipation driven theory of adaptation [79]. While there is some good intuitive notion behind this hypothesis and simulation experiments are now being performed to verify it, we will use information theoretic measures to cast adaptation as *adaptive learning*, and characterize the connection to the dissipation in a later section. In the next section, we will introduce a physical description of passive agents as FSA, and use these fluctuation theorems to obtain thermodynamic constraints for emergence of predictive inference in self-organized systems.

6.5 Passive Agents as Finite State Automata

Agency is the capacity of the system to act on it's environment. Passive agents are systems that simply interact with external signals from the environment, but cannot affect which future inputs the system interacts with. Active agents are those agents that interact with the environment and determine the future inputs that the systems interacts with and a lot more interesting to study. Before we move onto active agents, it is important to understand how learning behavior might emerge in passive agents that simply interact with external signals. In order to do so, we model the passive agent as physical finite state automata.

6.5.1 Physical FSA Description of Passive Agents

Passive agents like traditional FSAs described in previous sections are characterized by their internal states, the inputs that they receive, and the mapping that defines transitions between these internal states. The *physical* description of a passive agent is identical to the physical descriptions of Markov FSA that has been described in the chapter 4. We once again cast that description of the states, inputs and transitions below

- **Internal States:** The internal FSA states of the agent are faithfully represented in the distinguishable physical states of a quantum-mechanical system $\mathcal{S}$. The Markov property implies that the next state of the FSA depends only upon the current state of $\mathcal{S}$, and the next input. The system $\mathcal{S}$ interacts with it's environment $\mathcal{B}$, a (finite) heat bath nominally in a thermal state $\hat{\rho}_{th}^{\mathcal{B}}$ at temperature T.

- **Inputs:** Input strings $\vec{X}$ that the agents interacts with are physically instantiated in the state of a "referent" system $\mathcal{R} = \mathcal{R}_0\mathcal{R}_1$. Subsets $\vec{X}_k = X^{(1)}X^{(2)}...X^{(t-1)}$ of strings leading to the current FSA state are represented by $\mathcal{R}_0$, and $X^{(t)}$ is represented by $\mathcal{R}_1$ as before. In general, we assume that $\mathcal{R}_0$ and $\mathcal{R}_1$ are correlated.

- **State Transitions:** The t-th state transition is realized by dynamical evolution of the state of $\mathcal{S}$, conditioned on the state of $\mathcal{R}_1$ and in interaction with $\mathcal{B}$. Global evolution of the interacting composite $\mathcal{R}_1\mathcal{S}\mathcal{B}$ producing this transition is assumed to be governed by the time-dependent Schrodinger equation to ensure consistency with physical law. The t-th input remains encoded in $\mathcal{R}_1$ at the conclusion of the FSA state transition.

The "physical universe" relevant to description of the FSA is completed with the FSA's thermal bath $\mathcal{B}$, embedded in a "greater environment" $\bar{\mathcal{B}}$ which acts to

"rethermalize" $\mathcal{B}$ whenever it is driven from equilibrium by interaction with $\mathcal{S}$ during the state transitions.

6.5.2 Lower Bound on Dissipation for Passive Agents

Since passive agents are simply modeled as a Markov FSA from the previous chapters, we have the lower bound on dissipation for the agent as it interacts with external inputs to be the lower bound on dissipation over a state transition for the FSA instantiated in a general quantum system $\mathcal{S}$ [138] as

$$\Delta\langle E^{\mathcal{B}}\rangle \geq k_B T \ln(2) \left[-\Delta S^{\mathcal{S}} + \Delta\mathcal{I}^{\mathcal{R}_1\mathcal{S}}\right]$$

where $\mathcal{B}$ is a thermal bath at temperature T, and $\Delta\langle E^{\mathcal{B}}\rangle$ is the change in the expected energy of the bath and captures the dissipation associated with the state transition. $-\Delta S^{\mathcal{S}}$ is the change in von Neumann entropy of the system, and $\Delta\mathcal{I}^{\mathcal{R}_1\mathcal{S}}$ is the change in quantum mutual information between the system $\mathcal{S}$ and the input signal $\mathcal{R}_1$ driving the transition. $\mathcal{I}^{\mathcal{R}_1\mathcal{S}}$ is the correlation between the incoming input $\mathcal{R}_1$ and the agent $\mathcal{S}$ before the transition, and can be seen as a measure of *prediction of future inputs by the agent*. A much detailed description of the physical instantiation of a FSA, and derivation of the bound is provided in the previous chapters.

Since we intend to leave the detailed analysis of emergence of learning behavior in quantum systems for the future, we will continue forward with classical systems. A similar bound on dissipation for a Markov FSA has been derived by the authors in [82] from a completely classical perspective. The lower bound derived in [82] is equivalent to the equation below

$$\Delta\langle E^{\mathcal{B}}\rangle \geq k_B T \ln(2) \left[-\Delta H^{\mathcal{S}} + \Delta\mathcal{I}^{\mathcal{R}_1\mathcal{S}}\right] \tag{6.2}$$

where $-\Delta H^{\mathcal{S}}$ is the change in classical Shannon entropy, and $\Delta I^{\mathcal{R}_1\mathcal{S}}$ is the change in classical Shannon mutual information between $\mathcal{R}_1$ and $\mathcal{S}$. The dependency of the

dissipation bound on the mutual information between the $\mathcal{R}_1$ and $\mathcal{S}$, especially the prediction component, provides significant insight into the conditions under which learning is achieved in self-organized systems. These conditions will be discussed in the next section.

6.6 Dissipation Driven Adaptation & Learning

In this section, we will briefly discuss Jeremy England's dissipation driven adaptation hypothesis, cast adaptation as adaptive learning and extend upon the results from [79]. Engand studied the non-equilibrium conditions under which the system is more likely to be in one macrostate over another, and used it to propose the following [79] - "*while any given change in shape for the system is mostly random, the most durable and irreversible of these shifts in configuration occur when the system happens to be momentarily better at absorbing and dissipating work. With the passage of time, the 'memory' of these less erasable changes accumulates preferentially, and the system increasingly adopts shapes that resemble those in its history where dissipation occurred. Looking backward at the likely history of a product of this non-equilibrium process, the structure will appear to us like it has self-organized into a state that is well adapted to the environmental conditions. This is the phenomenon of dissipative adaptation.*'

The above hypothesis can be stated mathematically as follows. Taking negative logarithm on both sides of the fluctuation theorem Eq.(6.1), we have

$$-\ln\left[\frac{\pi(II^* \to I^*)}{\pi(I \to II)}\right] = -\ln\langle e^{\ln\left[\frac{p_f(j|II)}{p_0(i|I)}\right]}\langle e^{-\beta\Delta Q_{i\to j}}\rangle\rangle_{I\to II}$$

Using cumulant generating function - $\log E[e^{tX}] = \sum_{n=0}^{+\infty}\tau_n\frac{t^n}{n!} = \mu t + \sigma^2\frac{t^2}{2} + ...$ where τ_n is n-th moment, and μ and σ^2 correspond to the first (mean) and second (variance) moment respectively.

$$
\begin{aligned}
&- \ln\langle\langle e^{\ln\left[\frac{p_f(j|II)}{p_0(i|I)}\right]} e^{-\beta\Delta Q_{i\to j}}\rangle\rangle_{I\to II} \\
&\quad = \beta\langle\Delta Q - \frac{1}{\beta}\ln\left[\frac{p_f(j|II)}{p_0(i|I)}\right]\rangle_{I\to II} - \frac{\beta^2}{2}\sigma^2_{I\to II} + ... \\
&\quad = \kappa_{I\to II} - \gamma_{I\to II}
\end{aligned}
$$

where $-\frac{1}{\beta}\langle\ln\left[\frac{p_f(j|II)}{p_0(i|I)}\right]\rangle_{I\to II} = k_B T\ln(2)\Delta S_{I\to II}$, is the change in entropy of the microstates of the system, as system evolves from macrostate I to II. We define $\kappa_{I\to II} = \beta\langle\Delta Q\rangle_{I\to II} + k_B T\ln(2)\Delta S_{I\to II}$ as the sum of the mean of dissipation into the bath and the change in internal microstate entropy for transitions under the external drive, averaged over all trajectories and microstates. $\gamma_{I\to II} = \kappa_{I\to II} + \ln\langle e^{\ln\left[\frac{p_f(j|II)}{p_0(i|I)}\right]}\langle e^{-\beta\Delta Q_{i\to j}}\rangle\rangle_{I\to II}$ represents the fluctuations about the mean $\kappa_{I\to II}$. Now we can see that from the above equation, that a system is more likely to be in macrostate II than I, if $\kappa_{I\to II} >> \gamma_{I\to II}(\approx 0)$ - this is referred to as the condition of *reliable high dissipation* when $\Delta\langle Q\rangle >> 0$ and fluctuations about this average dissipation is low.

The condition of reliable high dissipation has been studied experimentally in [17] (though the quantity in focus was the rate of change of entropy and not dissipation), as well through simulation in toy chemical spaces [136] where the property of adaptation was characterized as resonance with the driving field. In this dissertation, evolution will be viewed as combination of processes of homeostasis (where the macrostate I of system does not change in time) and adaptation which entail irreversible macrostate changes in a system (from $I \to II$) in which the final state appears to more 'fitter' than the starting state. The latter will be viewed as the process of *learning the environment* over different time-scales and also referred to as adaptive learning. The concept of viewing evolution and adaptation as learning is not novel and extensively discussed in [85], [86], [87], [88], [89]. In [85], Valiant seeks to explain Darwinian

evolution as a special learning mechanism that he refers to as Probably Approximately Correct or PAC learning. In [88], Harper discusses the dynamics of the population following the replicator equation [90] as an inference process. He showed that as such a population approaches an 'evolutionary optimum' (corresponding to maximum fitness), the amount of Shannon information it has 'left to learn' about the optimal population is non-increasing. We would like to simply provide a thermodynamic context to this learning process.

In the previous section, the lower bound on dissipation in passive agents under the FSA model was presented. We can show that when the Shannon entropy change remains fixed and focusing only on the information part of the lower bound, we have the average energy dissipation into the bath to be directly proportional to $\Delta I^{\mathcal{R}_1\mathcal{S}}$. $\Delta I^{\mathcal{R}_1\mathcal{S}}$ characterizes the change in correlation between the physical system $\mathcal{S}$ and the driving signal $\mathcal{R}_1$, with an increase in correlation indicating the 'learning of the driving signal'. The reliable high dissipation condition $\kappa_{I\to II} >> 0$ can be satisfied by this increase in driving signal (environmental) correlation since $\kappa_{I\to II} \propto \Delta I^{\mathcal{R}_1\mathcal{S}}$. If we assume that the increased correlation with the environment is a sign of increasing fitness over time, then such processes can we be viewed as adaptive learning in the system of the environment driving it.

While dissipation driven adaptation forms an important aspect of explaining the correlation exhibited by physical systems to the input signals driving the system, it does not explain the ability of physical agents to utilize information to exhibit a key aspect of intelligence - prediction. As stated earlier in this section, we will view evolution of a system as a combination of adaptive learning and homeostatic processes. In the next section, we will derive the thermodynamic conditions that will capture the relationship between dissipation, homeostasis and prediction in both passive and active agents.

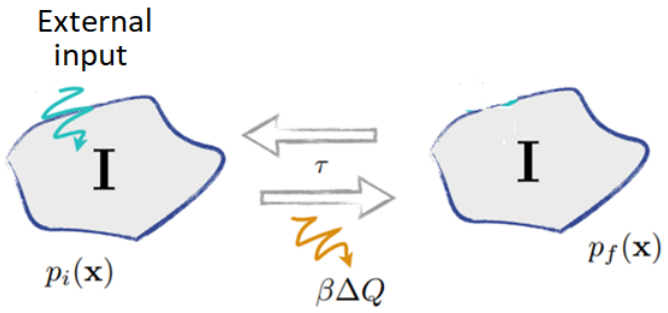

Figure 6.4. Homeostasis of a physical system which maintains the macro-observable at I while being driven by an external field over a time period τ and dissipates ΔQ into the bath. The system is characterized by an initial microstate distribution of $p_i(x)$ and a final distribution of $p_f(x)$.

6.7 Dissipation, Homeostasis and Prediction in Passive Agents

In the previous section, we presented the fluctuation theorem for macrostate-to-macrostate transitions from [79] and it's relationship to adaptive learning. It was also mentioned that the evolution of systems can be divided into adaptive learning and maintaining homeostasis. We will now use the same fluctuation theorems to further explore the homeostatic conditions under which prediction capabilities would emerge in self-organized systems. We define *homeostasis* as the property of a system in which a variable is actively regulated to remain very nearly constant. It's a defining feature of living systems, fundamental to the field of cybernetics [137] and it's role in neuronal plasticity has gained prominence [137]. Let us consider the case of a classical system in a non-equilibrium state driven by an external Hamiltonian in which the value of a macroscopic variable(s) corresponding to the macrostate I remains fixed over a finite time period. This corresponds to the system being homeostatic with respect to that macrostate I.

In the Eq.(6.1), for homeostasis we set $I = II$ (as shown in the Fig. 6.4) to get $\frac{\pi(I\to II)}{\pi(II\to I)} = 1$ and

$$\langle e^{\ln\left[\frac{p_f(j|I)}{p_0(i|I)}\right]} \langle e^{-\beta\Delta Q_{i\to j}}\rangle_{i\to j}\rangle_{I\to I} = 1$$

Taking negative logarithm on both sides and using the cumulant generating function, we get

$$-\ln\langle e^{\ln\left[\frac{p_f(j|I)}{p_0(i|I)}\right]}\langle e^{-\beta\Delta Q_{i\to j}}\rangle_{i\to j}\rangle_{I\to I}=0$$

$$\begin{aligned}
&-\ln\langle\langle e^{\ln\left[\frac{p_f(j|I)}{p_0(i|I)}\right]}e^{-\beta\Delta Q_{i\to j}}\rangle\rangle_{I\to I}\\
&\quad=\beta\langle\Delta Q-\frac{1}{\beta}\ln\left[\frac{p_f(j|I)}{p_0(i|I)}\right]\rangle_{I\to I}-\frac{\beta^2}{2}\sigma^2_{I\to I}+...\\
&\quad=\kappa-\gamma\\
&\quad=0
\end{aligned}$$

where $-\frac{1}{\beta}\langle\ln\left[\frac{p_f(j|I)}{p_0(i|I)}\right]\rangle_{I\to I}=k_BT\ln(2)\Delta H$, is the change in entropy of the microstates of the system, as macrostate I is maintained. We define $\kappa=\langle\Delta Q\rangle_{I\to I}+k_BT\ln(2)\Delta H$ as the sum of the mean of dissipation into the bath and the change in internal entropy for transitions under the external drive, averaged over all trajectories and microstates. $\gamma=\kappa+\ln\langle e^{\ln\left[\frac{p_f(j|I)}{p_0(i|I)}\right]}\langle e^{-\beta\Delta Q_{i\to j}}\rangle\rangle_{I\to I}$ represents the fluctuations about the mean κ. If κ is low when the macroscopic variable corresponding to I remains fixed, we would require the fluctuations γ to be low as well since $\kappa=\gamma$. This also implies that the value of $\beta\langle\Delta Q\rangle_{I\to I}+k_BT\ln(2)\Delta H$ over different micro-trajectories are concentrated around the mean. We will call this special case of the fluctuation theorem under homeostasis when $\kappa=\gamma$ to be *reliable low dissipation* (similar to how $\kappa>>\gamma$ was called reliable high dissipation).

The exact value of the mean dissipation into the bath $\langle\Delta Q\rangle$, as the system is driven by a Hamiltonian is very system specific, and can vary significantly. Since a system in a non-equilibrium state driven by an external field can be modeled as a FSA, we can substitute the lower bound on dissipation for a FSA from Eq.(6.3) for the actual mean dissipation, to obtain a lower bound on κ and gain insight into

the correlation between the system and the external driving signals, when the lower bound on κ is minimized.

$$\begin{aligned}\langle \Delta Q \rangle_{I\to I} + k_B T \ln(2) \Delta H^{\mathcal{S}}_{I\to I} &\geq k_B T \ln(2) \left[-\Delta H^{\mathcal{S}}_{Shannon} + \Delta \mathcal{I}^{\mathcal{R}_1 \mathcal{S}} \right] \\ &+ k_B T \ln(2) \Delta H^{\mathcal{S}}_{I\to I}\end{aligned}$$

It is important to note that the internal Shannon entropy terms ($H^{\mathcal{S}}_{Shannon}$) in the lower bound for κ, and the macrostate entropy $H^{\mathcal{S}}_{I\to I}$ from the fluctuation theorem expression need out cancel out as discussed in Chapter-2. Since it is possible to express the macrostate entropy in terms of the Shannon entropy, the two terms can be combined together. Using this previous discussion we can write (assuming both terms are in the same units) $\Delta H^{\mathcal{S}}_{I\to I} - \Delta H^{\mathcal{S}}_{Shannon} = -\alpha \Delta H^{\mathcal{S}}_{Shannon}$. Thus we have

$$\kappa = \langle \Delta Q \rangle_{I\to I} + k_B T \ln(2) \Delta H^{\mathcal{S}}_{I\to I} \geq k_B T \ln(2) \left[-\alpha \Delta H^{\mathcal{S}}_{Shannon} + \Delta \mathcal{I}^{\mathcal{R}_1 \mathcal{S}} \right]$$

A criticism of the results obtained here would be that we are using the lower bound on dissipation, rather than the actual dissipation. It is important to note that since we were looking at the case where κ and the fluctuations about the mean γ are low, the lower bound expression is not a bad approximation of the actual dissipation. In fact, many biological systems operate near the limits of energy efficiency [83], and the bound might only be a few orders below the actual energy dissipation. Furthermore, the change in information and non-information bearing entropy terms in the lower bound provide significant insight into the components associated with the actual dissipation. The $\Delta \mathcal{I}^{\mathcal{R}_1 \mathcal{S}}$ term in the lower bound is dependent on the correlation between the system $\mathcal{S}$ and the external signals $\mathcal{R} = \mathcal{R}_0 \mathcal{R}_1$ that drives this transition. $-\Delta H^{\mathcal{S}}_{Shannon}$ corresponds to the change in the distribution of states that are correlated to the driving input.

It is of interest to us to understand how the driving inputs are mapped onto the system (in a FSA model) when the system satisfies the reliable low dissipation condition. At some time t, let $\mathcal{R}_0$ correspond to the driving signals that have produced the current state of the system $\mathcal{S}$, and $\mathcal{R}_1$ be the incoming signal that will drive the next state transition. $\mathcal{R}_0$ and $\mathcal{R}_1$ are assumed to be correlated, which is often the case. Let there be a finite number of states $\{k_t^{\mathcal{S}}\}$ of $\mathcal{S}$ such that the past inputs map onto these states using the mapping $p(k_t^{\mathcal{S}}|i^{\mathcal{R}_0})$ and we assume that amount of information between $\mathcal{R}_0$ and $\mathcal{S}$ - $\mathcal{I}^{\mathcal{R}_0\mathcal{S}}$ at time t is equal to I_t. $\mathcal{I}^{\mathcal{R}_0\mathcal{S}}$ can be viewed as a measure of the system's memory and viewed as the result of dissipation driven adaptive learning. We are now interested in how $\mathcal{R}_0$ is encoded in the current state of the system $\mathcal{S}$, so that the lower bound on κ, for the transition driven by $\mathcal{R}_1$ is minimized. This is equivalent to calculating the mapping $p(k_t^{\mathcal{S}}|i^{\mathcal{R}_0})$ of the i-th input of $\mathcal{R}_0$ to the k-th state of $\mathcal{S}$ at time t, that will minimize $\kappa \geq k_B T \ln(2) \left[-\alpha \Delta H_{Shannon}^{\mathcal{S}} + \Delta \mathcal{I}^{\mathcal{R}_1\mathcal{S}}\right]$. The problem reduces to solving a constrained optimization problem with the Lagrangian $\mathcal{L}$

$$\mathcal{L} = \kappa + \lambda(\mathcal{I}^{\mathcal{R}_0\mathcal{S}} - I_t)$$

Replacing κ with the expression for the lower bound, we get

$$\min_{p(k_t^{\mathcal{S}}|i^{\mathcal{R}_0})}(k_B T \ln(2) \left[-\alpha \Delta H_{Shannon}^{\mathcal{S}} \Delta \mathcal{I}^{\mathcal{R}_1\mathcal{S}}\right] + \lambda(\mathcal{I}^{\mathcal{R}_0\mathcal{S}} - I_t))$$

where λ is the trade-off parameter between κ and $\mathcal{I}^{\mathcal{R}_0\mathcal{S}}$. Since we are interested in the mapping $p(k_t^{\mathcal{S}}|i^{\mathcal{R}_0})$ that minimizes the Lagrangian for any state transition mapping, the solution is independent of $\mathcal{I}^{\mathcal{R}_1\mathcal{S}'}$ and $H_{Shannon}^{\mathcal{S}'}$, and the problem reduces to

$$\max_{p(k_t^{\mathcal{S}}|i^{\mathcal{R}_0})}(\mathcal{I}^{\mathcal{R}_1\mathcal{S}} - \alpha H_{Shannon}^{\mathcal{S}} - \lambda(\mathcal{I}^{\mathcal{R}_0\mathcal{S}} - I_t))$$

The constraint optimization problem above is the information bottleneck algorithm, an information theoretic technique to achieve optimal data representation and predictive inference [92]. The solution for $p(k_t^{\mathcal{S}}|i^{\mathcal{R}_0})$ is of the form

$$p(k_t^{\mathcal{S}}|i^{\mathcal{R}_0}) \propto e^{\frac{-1}{\lambda}\left[D_{KL}[p(j^{\mathcal{R}_1}|i^{\mathcal{R}_0})|p(j^{\mathcal{R}_1}|k_t^{\mathcal{S}})]-\ln p(k_t^{\mathcal{S}})\right]}$$

where $D_{KL}[a|b]$ is the Kullback-Liebler (KL) divergence between distributions $\{a\}$ and $\{b\}$.$D_{KL}[a|b] \geq 0$ for any two distributions with equality achieved when $\{a\} = \{b\}$. The above solution indicates that in an homeostatic system with a finite amount of memory, when κ is minimized with respect to the mapping $p(k_t^{\mathcal{S}}|i^{\mathcal{R}_0})$ of the input $i^{\mathcal{R}_0}$ to state $k_t^{\mathcal{S}}$ we have that

(a) $p(k_t^{\mathcal{S}}|i^{\mathcal{R}_0})$ is higher if $D_{KL}[p(j^{\mathcal{R}_1}|i^{\mathcal{R}_0})|p(j^{\mathcal{R}_1}|k_t^{\mathcal{S}})]$ is lower. This entails that the actual conditional distribution of the input $\{p(j^{\mathcal{R}_1}|i^{\mathcal{R}_0})\}$ be similar to the predicted distribution of the next inputs by the system state $\{p(j^{\mathcal{R}_1}|k_t^{\mathcal{S}})\}$. Thus the probability of i-th input of $\mathcal{R}_0$ to the k-th state of $\mathcal{S}$ is greater if that state allows for better prediction of the next driving input $j^{\mathcal{R}_1}$.

(b) $p(k_t^{\mathcal{S}}|i^{\mathcal{R}_0})$ is higher if $\ln p(k_t^{\mathcal{S}})$ is lower. This is achieved when $p(k_t^{\mathcal{S}}) \to 1$ indicating a preference of sparser distributions of $\{p(k^t)\}$ over broader distributions.

Thus we see that in systems for which the lower bound on κ is minimized i.e. the reliable low dissipation condition, the mapping of the past signals in the states of the system is skewed towards better prediction the next driving input and sparse representations.

The transition state probability $p(l_{t+1}^{\mathcal{S}}|k_t^{\mathcal{S}}, j^{\mathcal{R}_1})$ is the probability that the k-th state of $\mathcal{S}$ at time t maps to the l-th state at time $(t+1)$ when driven by the j-th input of $\mathcal{R}_1$, and characterizes the system's temporal dynamics with respect to the driving signal. The state transition probabilities $p(l_{t+1}^{\mathcal{S}}|k_t^{\mathcal{S}}, j^{\mathcal{R}_1})$ when the system maintains the reliable low dissipation over a finite time period can be obtained from the following differential equation

$$\frac{dp(l_{t+1}^{\mathcal{S}}|i^{\mathcal{R}_0}, j^{\mathcal{R}_1})}{dp(k_t^{\mathcal{S}}|i^{\mathcal{R}_0})} = p(l_{t+1}^{\mathcal{S}}|p(k_t^{\mathcal{S}}, j^{\mathcal{R}_1}) + p(k_t^{\mathcal{S}}|i^{\mathcal{R}_0})\frac{dp(l_{t+1}^{\mathcal{S}}|p(k_t^{\mathcal{S}}, j^{\mathcal{R}_1})}{dp(k_t^{\mathcal{S}}|i^{\mathcal{R}_0})}$$

This equation is obtained under the assumption that the $p(l^{\mathcal{S}}_{t+1}|p(k^{\mathcal{S}}_t, j^{\mathcal{R}_1})$ is time-varying and dependent on past mappings $p(k^{\mathcal{S}}_t|i^{\mathcal{R}_0})$ i.e. $\frac{dp(l^{\mathcal{S}}_{t+1}|p(k^{\mathcal{S}}_t, j^{\mathcal{R}_1})}{dp(k^{\mathcal{S}}_t|i^{\mathcal{R}_0})} \neq 0$. The solution to the above equation is extremely specific to the choice of distributions and is often not tractable. However if we assume that $\frac{dp(l^{\mathcal{S}}_{t+1}|i^{\mathcal{R}_0}, j^{\mathcal{R}_1})}{dp(k^{\mathcal{S}}_t|i^{\mathcal{R}_0})} = Q[p(k^{\mathcal{S}}_t|i^{\mathcal{R}_0})]$ where Q is some function of $p(k^{\mathcal{S}}_t|i^{\mathcal{R}_0})$, then we can say that

$$\begin{aligned} p(l^{\mathcal{S}}_{t+1}|k^{\mathcal{S}}_t, j^{\mathcal{R}_1}) \quad &\propto \quad e^{\frac{-1}{\lambda}\{D_{KL}[p(m^{\mathcal{R}_2}|i^{\mathcal{R}_0} j^{\mathcal{R}_1})|p(m^{\mathcal{R}_2}|l^{\mathcal{S}}_{t+1})]\}} \\ &\quad \times e^{\frac{1}{\lambda'}\{D_{KL}[p(j^{\mathcal{R}_1}|i^{\mathcal{R}_0})|p(j^{\mathcal{R}_1}|k^{\mathcal{S}}_t)]\}} \end{aligned}$$

The first term on the right hand side corresponds to the 'generative' component maximizing prediction as before. In addition, $p(l^{\mathcal{S}}_{t+1}|k^{\mathcal{S}}_t, j^{\mathcal{R}_1})$ is higher, if the second term, $D_{KL}[p(j^{\mathcal{R}_1}|i^{\mathcal{R}_0})|p(j^{\mathcal{R}_1}|k^{\mathcal{S}}_t)]$ is higher. This term corresponds to a 'recognition' component minimizing past errors. We thus see that when a macroscopic variable is homeostatically maintained in an energy efficient low dissipation manner in a driven non-equilibrium systems, the dynamics of the homeostatic system can be characterized by the state-transition probabilities that exhibit top-down generative and bottom-up recognition components. This is exactly what is seen systems like the human brain which exhibit prediction-centric intelligence [142], [97].

Similar results have also been proposed for information engines in [141]. The solutions here are related to the unsupervised learning techniques from the Helmholtz machine [143] and variational autoencoders [144], that use generative and recognition components to learn optimal encodings of the underlying structure in unlabeled data. Unlike these algorithms, where learning is the a priori goal, the results presented here hint that the preference for predictive dynamics of the driving signals in self-organized systems is the result of specific thermodynamic conditions. Rather than looking to make the physical implementations of learning algorithms more energy efficient, we should recognize that *physical systems that satisfy the reliable low dissipation ther-*

modynamic condition exhibit dynamics with a preference for predictive inference and sparse representations. We will next illustrate the reliable low dissipation condition and the preferred encodings of the input in the states of the system with a simple example. Following that we will briefly extend these results to active agents and then proceed to define a new computing paradigm based on these thermodynamic conditions.

6.7.1 Illustrative Example

The reliable low dissipation condition is powerful since it brings together 3 very important ideas in dissipation, homeostasis and prediction in a non-equilibrium scenario. Given that these are ultimately thermodynamical conditions, they need to be experimentally tested and verified by measuring observables such as heat dissipation, entropy (or suitable substitutes) and temporal dynamics in self-organized systems. There is additional complexity of identifying the suitable spatial and temporal scale to study these conditions in both biological and artificial systems. Given that, we will illustrate the implications of the theoretical results from the above section with a simple example of a system with two states.

Consider the system $\mathcal{S}$ with two states $k = 0, 1$. Let us assume these are the states of the system when $\mathcal{S}$ satisfies the reliable low dissipation condition. The system is being driven by a string of inputs made up of 0's and 1's. We will discuss a couple of cases with different conditional distributions for the inputs and see how that influences the mapping onto the states of the system. Now from the reliable low dissipation condition we know that the encoding of the i-th input on the k-th state of the system is dependent on the closeness of the prediction of the next input to the actual conditional distribution of the input.

$$p(k_t^{\mathcal{S}}|i^{\mathcal{R}_0}) \propto e^{\frac{-1}{\lambda}\left[D_{KL}[p(j^{\mathcal{R}_1}|i^{\mathcal{R}_0})|p(j^{\mathcal{R}_1}|k_t^{\mathcal{S}})]-\ln p(k_t^{\mathcal{S}})\right]}$$

We will start by focusing on the Kullback-Liebler divergence component of the equation keeping in mind that the other exponent is a sparsity criterion driving the system towards narrower state distributions. The D_{KL} can be expanded as

$$D_{KL}[p(j^{\mathcal{R}_1}|i^{\mathcal{R}_0})|p(j^{\mathcal{R}_1}|k_t^{\mathcal{S}})] = \sum_{j=0,1} p(j^{\mathcal{R}_1}|i^{\mathcal{R}_0}) \log_2 \left[\frac{p(j^{\mathcal{R}_1}|i^{\mathcal{R}_0})}{p(j^{\mathcal{R}_1}|k_t^{\mathcal{S}})}\right]$$

Writing out the terms clearly we have

$$D_{KL}[p(j^{\mathcal{R}_1}|i^{\mathcal{R}_0})|p(j^{\mathcal{R}_1}|k_t^{\mathcal{S}})] = p(0^{\mathcal{R}_1}|i^{\mathcal{R}_0}) \log_2 \left[\frac{p(0^{\mathcal{R}_1}|i^{\mathcal{R}_0})}{p(1^{\mathcal{R}_1}|k_t^{\mathcal{S}})}\right] + p(1^{\mathcal{R}_1}|i^{\mathcal{R}_0}) \log_2 \left[\frac{p(1^{\mathcal{R}_1}|i^{\mathcal{R}_0})}{p(1^{\mathcal{R}_1}|k_t^{\mathcal{S}})}\right]$$

For $p(0_t^{\mathcal{S}}|0^{\mathcal{R}_0})$, we have

$$D_{KL}[p(j^{\mathcal{R}_1}|0^{\mathcal{R}_0})|p(j^{\mathcal{R}_1}|0_t^{\mathcal{S}})] = p(0^{\mathcal{R}_1}|0^{\mathcal{R}_0}) \log_2 \left[\frac{p(0^{\mathcal{R}_1}|0^{\mathcal{R}_0})}{p(1^{\mathcal{R}_1}|0_t^{\mathcal{S}})}\right] + p(1^{\mathcal{R}_1}|0^{\mathcal{R}_0}) \log_2 \left[\frac{p(1^{\mathcal{R}_1}|0^{\mathcal{R}_0})}{p(1^{\mathcal{R}_1}|0_t^{\mathcal{S}})}\right]$$

For the input conditional distribution given by $p(0^{\mathcal{R}_1}|0^{\mathcal{R}_0}) = 1$, $p(1^{\mathcal{R}_1}|0^{\mathcal{R}_0}) = 0$, $p(0^{\mathcal{R}_1}|1^{\mathcal{R}_0}) = 0$ and $p(1^{\mathcal{R}_1}|1^{\mathcal{R}_0}) = 1$ we get

$$D_{KL}[p(j^{\mathcal{R}_1}|0^{\mathcal{R}_0})|p(j^{\mathcal{R}_1}|0_t^{\mathcal{S}})] = p(0^{\mathcal{R}_1}|0^{\mathcal{R}_0}) \log_2 \left[\frac{p(0^{\mathcal{R}_1}|0^{\mathcal{R}_0})}{p(1^{\mathcal{R}_1}|0_t^{\mathcal{S}})}\right]$$

$$D_{KL}[p(j^{\mathcal{R}_1}|0^{\mathcal{R}_0})|p(j^{\mathcal{R}_1}|0_t^{\mathcal{S}})] = -\log_2 p(0^{\mathcal{R}_1}|0_t^{\mathcal{S}})$$

Thus we have $p(0_t^{\mathcal{S}}|0^{\mathcal{R}_0}) \propto e^{\frac{1}{\lambda}[\log_2 p(1^{\mathcal{R}_1}|0_t^{\mathcal{S}})]}$. This implies that the probability that $0^{\mathcal{R}_0}$ maps to the state $0_t^{\mathcal{S}}$ depends upon the conditional probability that $0^{\mathcal{R}_1}$ is predicted by the state $0_t^{\mathcal{S}}$. For simplicity, let us assume that the prediction by every state is the state itself. This gives us that $p(0^{\mathcal{R}_1}|0_t^{\mathcal{S}}) = p(1^{\mathcal{R}_1}|1_t^{\mathcal{S}}) = 1$. Under this assumption for the prediction by the states, we have $D_{KL} = [p(j^{\mathcal{R}_1}|0^{\mathcal{R}_0})|p(j^{\mathcal{R}_1}|0_t^{\mathcal{S}})] = -\log_2 p(0^{\mathcal{R}_1}|0_t^{\mathcal{S}}) = 0$ is minimized. Other factors being equal, we have that for the given input conditional distribution and state prediction $p(0_t^{\mathcal{S}}|0^{\mathcal{R}_0}) \to 1$. Likewise we can show that $p(1_t^{\mathcal{S}}|1^{\mathcal{R}_0}) \to 1$ and conversely $p(1_t^{\mathcal{S}}|0^{\mathcal{R}_0}) \to 0$. Of course these mappings would be very different if the mapping of the state into the outputs were different as well. But we are beginning to see the input-FSA state mappings that are preferred under specific thermodynamic conditions. Like any other scientific hypothesis, further experimental evidence is specific self-organized networks will be necessary to confirm these results.

6.8 Active Agents as Finite State Automata

Active agents are those agents that can interact and affect the inputs they receive through their actions. These are very important since we would require our intelligent systems to be able to interact, learn and act intelligently in a dynamic environment. In this paper, we are interested in active agents with a fixed set of actions. In this section, I will provide a physical finite state automata (FSA) description of such an agent, and the lower bound on dissipation over state transitions.

6.8.1 Physical FSA Description of Active Agents

Active agents are characterized by their internal states, the inputs that they receive, a mapping that defines transitions between these internal states, and state-dependent actions that influence future inputs. We construct a very general *physical* description of an active agent as a Markov FSA by identifying the physical realizations of the agent states, inputs, and state transitions.

- **Internal States:** The internal FSA states are faithfully represented in the distinguishable physical states of a joint quantum-mechanical system $\mathcal{SA}$. The Markov property implies that the next state of the FSA depends only upon the current state of $\mathcal{SA}$, and the next input. We will also assume (without loss in generality) that these internal states can be divided into - *action states* of system $\mathcal{A}$ that affect the next input to the joint system, and *sensory states* of system $\mathcal{S}$ that do not affect the incoming inputs. State transitions in both $\mathcal{S}$ and $\mathcal{A}$ are however dependent on the input.). The mapping from the states of a physical system $\mathcal{A}$ to the action policy will not be explored in detail, beyond the assumption that they are fixed for the given system. The joint system interacts with it's environment $\mathcal{B}$, a (finite) heat bath nominally in a thermal state $\hat{\rho}_{th}^{\mathcal{B}}$ at temperature T.

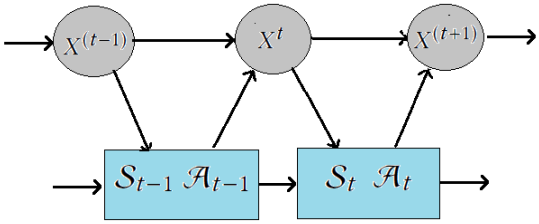

Figure 6.5. Finite state automata description of an active agent $\mathcal{SA}$ with sensory and action states.

- **Inputs:** Input strings $\vec{X}$ are physically instantiated in the state of a "referent" system $\mathcal{R} = \mathcal{R}_0\mathcal{R}_1$, which can be regarded as a physical "input tape" that holds the output of a classical information source. Subsets $\vec{X}_k = X^{(1)}X^{(2)}...X^{(t-1)}$ of strings leading to the current FSA state are represented by $\mathcal{R}_0$, and $X^{(t)}$ is represented by $\mathcal{R}_1$. In general, we assume that $\mathcal{R}_0$ and $\mathcal{R}_1$ are correlated, and that the input distribution of $\mathcal{R}_1$ is dependent on the current state of the subsystem $\mathcal{A}$ and $\mathcal{R}_0$.

- **State Transitions:** The t-th state transition is realized by dynamical evolution of the state of $\mathcal{SA}$, conditioned on the state of $\mathcal{R}_1$ and in interaction with $\mathcal{B}$. Global evolution of the interacting composite $\mathcal{R}_1\mathcal{SAB}$ producing this transition is assumed to be governed by the time-dependent Schrodinger equation to ensure consistency with physical law. The t-th input remains encoded in $\mathcal{R}_1$ at the conclusion of the FSA state transitions.

To complete the "physical universe" relevant to description of the FSA, we will once again have that the FSA's local environment $\mathcal{B}$ is embedded in a "greater environment" $\bar{\mathcal{B}}$ which acts to "rethermalize" $\mathcal{B}$ whenever it is driven from equilibrium by interaction with $\mathcal{S}$ during state transitions.

6.8.2 Lower Bound on Dissipation for Active Agents

In this section, I will present the lower bound on dissipation over a state transition for the *interactive* FSA described above, instantiated in a general quantum system $\mathcal{SA}$, based on the derivation in [138].

$$\Delta\langle E^{\mathcal{B}}\rangle \geq k_B T \ln(2) \left[-\Delta S^{\mathcal{SA}} + \Delta\mathcal{I}^{\mathcal{R}_1\mathcal{SA}}\right]$$

where $\mathcal{B}$ is a thermal bath at temperature T, and $\Delta\langle E^{\mathcal{B}}\rangle$ is the change in the expected energy of the bath and captures the dissipation associated with the state transition. $-\Delta S^{\mathcal{SA}}$ is the change in von Neumann entropy of the system, and $\Delta\mathcal{I}^{\mathcal{R}_1\mathcal{SA}}$ is the change in quantum mutual information between the system $\mathcal{SA}$ and the input signal $\mathcal{R}_1$ driving the transition. $\mathcal{I}^{\mathcal{R}_1\mathcal{SA}}$ is the correlation between $\mathcal{R}_1$ and $\mathcal{SA}$, before the transition and can be seen as a measure of *prediction* of the incoming input. A much detailed description of the physical instantiation of a FSA, and derivation of the bound is provided in [138].

We will continue forward with classical systems and leave the detailed analysis of the emergence of learning in quantum systems for the future. A similar bound on dissipation for a Markov FSA has been derived by the authors in [82] from a completely classical perspective. The lower bound derived in [82] is equivalent to the equation below

$$\Delta\langle E^{\mathcal{B}}\rangle \geq k_B T \ln(2) \left[-\Delta H^{\mathcal{SA}} + \Delta\mathcal{I}^{\mathcal{R}_1\mathcal{SA}}\right] \tag{6.3}$$

where $-\Delta H^{\mathcal{SA}}$ is the change in Shannon entropy, and $\Delta I^{\mathcal{R}_1\mathcal{SA}}$ is the change in classical Shannon mutual information between $\mathcal{R}_1$ and $\mathcal{SA}$. In the next section, we will discuss the relationship between the reliable low dissipation conditions and prediction dynamics in active agents.

6.9 Dissipation, Homeostasis and Prediction in Active Agents

We will now use the fluctuation theorems to study the physical conditions under which prediction capabilities would emerge in self-organized active agents. Consider the case of a classical system in a non-equilibrium homeostatic state with macrostate I fixed over a finite time period. Setting $I = II$ in Eq.(6.1) again we get $\pi(I \to II) = \pi(II \to I)$ and

$$\langle e^{\ln\left[\frac{p_f(j|I)}{p_0(i|I)}\right]} \langle e^{-\beta \Delta Q_{i\to j}} \rangle_{i\to j} \rangle_{I\to I} = 1$$

Taking negative logarithm on both sides and using the cumulant generating function, this equation can be reduced to

$$-\ln\langle e^{\ln\left[\frac{p_f(j|I)}{p_0(i|I)}\right]} \langle e^{-\beta \Delta Q_{i\to j}} \rangle_{i\to j} \rangle_{I\to I} = 0$$

$$\begin{aligned} &-\ln\langle\langle e^{\ln\left[\frac{p_f(j|I)}{p_0(i|I)}\right]} e^{-\beta \Delta Q_{i\to j}} \rangle\rangle_{I\to I} \\ &\quad = \kappa_I - \gamma_I \\ &\quad = 0 \end{aligned}$$

where κ_I and γ_I represent the mean and fluctuations about the mean as before, when the system maintains it's macrostate I over a finite period of time. If κ_I is low when the macroscopic variable I remains fixed, the fluctuations γ_I needs to be low as well to satisfy $\kappa_I = \gamma_I$. This implies that the value of $\beta\langle\Delta Q\rangle_{I\to I} + k_B T \ln(2) \Delta H_{I\to I}$ over different micro-trajectories are concentrated around the mean, and once again corresponds to the condition of *reliable low dissipation* for active agents.

As before, I will model the agent system in a non-equilibrium state as an interactive FSA, and substitute the FSA lower bound on dissipation from Eq.(6.3) to obtain a lower bound on κ. We use this substitution to gain insight into the correlation

between the system and the external driving signals, when the lower bound on κ is minimized under homeostasis.

$$\begin{aligned}\kappa &= \langle \Delta Q \rangle_{I \to I} + k_B T \ln(2) \Delta H_{I \to I} \\ &\geq k_B T \ln(2) \left[-\Delta H^{\mathcal{SA}}_{Shannon} + \Delta \mathcal{I}^{\mathcal{R}_1 \mathcal{SA}} + \Delta H_{I \to I} \right]\end{aligned}$$

The macrostate entropy $H_{I \to I}$ is related to the Shannon entropy $H^{\mathcal{SA}}_{Shannon}$ depending upon the choice of macro-observable I as well as the information bearing variable. We will characterize their relationship as $\Delta H_{I \to I} = -\alpha \Delta H^{\mathcal{SA}}_{Shannon}$.

We will make similar assumptions for the FSA models active agents of as we did for passive agents. At some time t, let $\mathcal{R}_0$ correspond to the driving signals that have produced the current state of the system $\mathcal{SA}$, and $\mathcal{R}_1$ be the incoming signal that will drive the next state transition. $\mathcal{R}_0$ and $\mathcal{R}_1$ are assumed to be correlated, which is often the case. Let there be a finite number of states $\{k^{\mathcal{S}}_t, l^{\mathcal{A}}_t\}$ of $\mathcal{SA}$ such that the past inputs map onto these states using the mapping $p(k^{\mathcal{S}}_t, l^{\mathcal{A}}_t | i^{\mathcal{R}_0})$ and we assume that amount of information between $\mathcal{R}_0$ and $\mathcal{SA}$ - $\mathcal{I}^{\mathcal{R}_0 \mathcal{SA}}$ at time t is equal to I_t. $\mathcal{I}^{\mathcal{R}_0 \mathcal{SA}}$ can be viewed as a measure of the system's memory and viewed as the result of dissipation driven adaptive learning. We are now interested in how $\mathcal{R}_0$ is encoded in the current state of the system $\mathcal{SA}$, so that the lower bound on κ, for the transition driven by $\mathcal{R}_1$ is minimized. This is equivalent to calculating the mapping $p(k^{\mathcal{SA}}_t | i^{\mathcal{R}_0})$ of the i-th input of $\mathcal{R}_0$ to the (k, l)-th state of $\mathcal{SA}$ at time t, that will minimize $\kappa \geq k_B T \ln(2) \left[-\alpha \Delta H^{\mathcal{S}}_{Shannon} + \Delta \mathcal{I}^{\mathcal{R}_1 \mathcal{S}} \right]$.

The problem can again be cast as a constrained optimization problem with the Lagrangian

$$\mathcal{L} = \kappa + \lambda (\mathcal{I}^{\mathcal{R}_0 \mathcal{SA}} - I_t)$$

Replacing κ with the expression for the lower bound, we get

$$\min_{p(k_t^{\mathcal{S}}, l_t^{\mathcal{A}}|i^{\mathcal{R}_0})} \left[-\Delta H_{Shannon}^{\mathcal{SA}} + \Delta \mathcal{I}^{\mathcal{R}_1\mathcal{SA}} + \Delta H_{I\to I} + \lambda(\mathcal{I}^{\mathcal{R}_0\mathcal{SA}} - I_t)\right]$$

where λ is the trade-off parameter between κ and $\mathcal{I}^{\mathcal{R}_0\mathcal{SA}}$. Since we are interested in the mapping $p(k_t^{\mathcal{S}}, l_t^{\mathcal{A}}|i^{\mathcal{R}_0})$ that minimizes the Lagrangian for any state transition mapping, the solution is independent of $\mathcal{I}^{\mathcal{R}_1\mathcal{S}'}$ and $H^{\mathcal{S}'\mathcal{A}'}$. Thus the problem reduces to

$$\min_{p(k_t^{\mathcal{S}}, l_t^{\mathcal{A}}|i^{\mathcal{R}_0})}(\alpha H^{\mathcal{SA}} - \mathcal{I}^{\mathcal{R}_1\mathcal{SA}} + \lambda(\mathcal{I}^{\mathcal{R}_0\mathcal{SA}} - I_t))$$

This is equivalent to

$$\max_{p(k_t^{\mathcal{S}}, l_t^{\mathcal{A}}|i^{\mathcal{R}_0})}(\mathcal{I}^{\mathcal{R}_1\mathcal{SA}} - H^{\mathcal{SA}} - \lambda(\mathcal{I}^{\mathcal{R}_0\mathcal{SA}} - I_t))$$

The constraint optimization problem above is the information bottleneck algorithm, an information theoretic technique to achieve interactive learning [92]. The solution for $p(k_t^{\mathcal{S}}, l_t^{\mathcal{A}}|i^{\mathcal{R}_0})$ is of the form

$$\begin{aligned} p(k_t^{\mathcal{S}}, l_t^{\mathcal{A}}|i^{\mathcal{R}_0}) \quad &\propto \quad e^{\frac{-1}{\lambda}\left[D_{KL}[p(j^{\mathcal{R}_1}|i^{\mathcal{R}_0})|p(j^{\mathcal{R}_1}|k_t^{\mathcal{S}}, l_t^{\mathcal{A}})] - \ln p(k_t^{\mathcal{S}}, l_t^{\mathcal{A}})\right]} \\ &\quad \times e^{\frac{1}{\lambda}\left[D_{KL}[p(j^{\mathcal{R}_1}|i^{\mathcal{R}_0}, l_t^{\mathcal{A}})|q(j^{\mathcal{R}_1})]\right]} \end{aligned} \tag{6.4}$$

where $D_{KL}[a|b]$ is the Kullback-Liebler (KL) divergence between distributions $\{a\}$ and $\{b\}$. The optimal encoding of $i^{\mathcal{R}_0}$ to state $k_t^{\mathcal{S}} l_t^{\mathcal{A}}$, $p(k_t^{\mathcal{S}}, l_t^{\mathcal{A}}|i^{\mathcal{R}_0})$ is dependent on 3 factors

- *Exploitation* - $D_{KL}[p(j^{\mathcal{R}_1}|i^{\mathcal{R}_0})|p(j^{\mathcal{R}_1}|k_t^{\mathcal{S}}, l_t^{\mathcal{A}})]$ is low. This implies that those input $\mathcal{R}_0$ - state $\mathcal{SA}$ mappings that improve prediction of the next driving input $j^{\mathcal{R}_1}$ (true conditional distribution $\{p(j^{\mathcal{R}_1}|i^{\mathcal{R}_0})\}$ and predicted distribution $\{p(j^{\mathcal{R}_1}|k_t^{\mathcal{S}}, l_t^{\mathcal{A}})\}$ are similar) are preferred. We identify this as the exploitation

component, that looks to maximize the prediction of future inputs from past ones in both the sensory and action states. This is very interesting to analyze particularly for the states that influence action. The states of A that maximize the prediction of the next input produced by a fixed action policy, are the states and by extension the actions that are more probable. This might seem a little counter-intuitive at first but align well with the predictions of free-energy principle in neuroscience [146]. At this juncture, it is a reasonable to argue that the action that maximizes input prediction in certain cases, might be one that produces a passive agent with no action. This problem is addressed by the exploration component in the equation that is discussed next.

- *Exploration* - $D_{KL}[p(j^{\mathcal{R}_1}|i^{\mathcal{R}_0}, l_t^{\mathcal{A}})|q(j^{\mathcal{R}_1})]$ is higher. This entails those action states that allow for the conditional future distribution $p(j^{\mathcal{R}_1}|i^{\mathcal{R}_0}, l_t^{\mathcal{A}})$ to deviate from the average distribution of future inputs $p(j^{\mathcal{R}_1})$ are preferred. This would correspond to an *exploration* component in state encodings. The conditional distribution of inputs produced by action states of $\mathcal{A}$ - $\{p(j^{\mathcal{R}_1}|i^{\mathcal{R}_0}, l_t^{\mathcal{A}}\}$ that differ from the average future input $\{|q(j^{\mathcal{R}_1})\}$ are more preferred. This strikes a tradeoff with the exploitation component and ensure a system does not remain passive. Also note that the exploration component arose as part of the optimal solution that minimizes κ, and did not have to be artificially introduced as it often the case in techniques like reinforcement learning [101].

- *Sparsity* - $\ln p(k_t^{\mathcal{S}}, l_t^{\mathcal{A}})$ is higher. This happens when $p(k_t^{\mathcal{S}}, l_t^{\mathcal{A}}) \rightarrow 1$ for a particular state of the system $\mathcal{SA}$ which indicates that sparse distributions of the system states are preferred over dense distributions. The importance of sparse representations in intelligent systems has been well analyzed in [98], and plays a crucial role in the design of neuromorphic system [99].

We see that in active agents that satisfy the condition of reliable low dissipation, the encoding of the past signals in the states of the agent system is a balanced trade-off between exploitation, exploration and sparsity factors. It is important to note that while these components themselves have been observed in other learning approaches [102], it has been derived from the fluctuation theorems without assuming prediction as the intended goal beforehand.

The transition state probability $p(u^{\mathcal{S}}_{t+1}v^{\mathcal{A}}_{t+1}|k^{\mathcal{S}}_t l^{\mathcal{A}}_t, j^{\mathcal{R}_1})$ is the probability that the $(k^{\mathcal{S}}, l^{\mathcal{A}})$-th state of $\mathcal{SA}$ at time t, maps to the $(u^{\mathcal{S}}, v^{\mathcal{A}})$-th state at time $(t+1)$ when driven by the j-th input of $\mathcal{R}_1$, and characterizes the system dynamics with respect to the driving signal. These are dependent upon all three related factors of exploitation, exploration and sparsity. The dependency of the state transition probabilities on the exploitation components, when the system maintains the thermodynamic conditions described above are explored below.

$$p(u^{\mathcal{S}}_{t+1}v^{\mathcal{A}}_{t+1}|k^{\mathcal{S}}_t l^{\mathcal{A}}_t, j^{\mathcal{R}_1}) \propto \; e^{\frac{-1}{\lambda}\{D_{KL}[p(m^{\mathcal{R}_2}|i^{\mathcal{R}_0}j^{\mathcal{R}_1}, v^{\mathcal{A}}_{t+1})|p(m^{\mathcal{R}_2}|u^{\mathcal{S}}_{t+1}v^{\mathcal{A}}_{t+1})]\}} \times e^{\frac{1}{\lambda'}\{D_{KL}[p(j^{\mathcal{R}_1}|i^{\mathcal{R}_0}, l^{\mathcal{A}}_t)|p(j^{\mathcal{R}_1}|k^{\mathcal{S}}_t l^{\mathcal{A}}_t)]\}}$$

The first term on the right hand side corresponds to the 'generative' component of states in $\mathcal{SA}$ maximizing prediction as before. In addition to these, $p(u^{\mathcal{S}}_{t+1}v^{\mathcal{A}}_{t+1}|k^{\mathcal{S}}_t l^{\mathcal{A}}_t, j^{\mathcal{R}_1})$ increases if the second D_{KL} term is higher, corresponding to a 'recognition' component minimizing past errors. The dynamics of the homeostatic system in which a macroscopic variable is maintained in an energy efficient manner, are characterized by state-transition probabilities that exhibit both top-down generative and bottom-up recognition components. Similar components have been observed in the cerebral cortex of the human brain, and also used in unsupervised learning techniques like the Helmholtz machine [143] and variational autoencoders [144]. An analysis of the state transition probability dependence with respect to *exploration* components, indicates

that there is a trade off between exploration components of the two time steps, as shown below.

$$p(u^{\mathcal{S}}_{t+1}v^{\mathcal{A}}_{t+1}|k^{\mathcal{S}}_t l^{\mathcal{A}}_t, j^{\mathcal{R}_1}) \propto e^{\frac{1}{\lambda}\{D_{KL}[p(m^{\mathcal{R}_2}|i^{\mathcal{R}_0}j^{\mathcal{R}_1}, v^{\mathcal{A}}_{t+1})|p(m^{\mathcal{R}_2}])\}} \times e^{\frac{-1}{\lambda'}\{D_{KL}[p(j^{\mathcal{R}_1}|i^{\mathcal{R}_0}, l^{\mathcal{A}}_t)|p(j^{\mathcal{R}_1})]\}}$$

The results presented in this section show that *in active agents that satisfy the reliable low dissipation condition for homeostasis, the temporal dynamics of the system exhibit a tradeoff between exploitation, exploration and sparsity.* By not assuming the prediction dynamics apriori and/or artificially introducing the exploration component, the results here indicate significant first steps in identifying a task independent objective functions for active agents to determine their action policies. These results are closely tied to ideas of intrinsic motivation in agents [103], [104]. In order to make the jump from our current successes with narrow intelligence to an artificial general intelligence, we need agents to be capable of *lifelong learning.* Some defining characteristics of lifelong learning include - continuous learning and adaptation to the changing environments, and the capability to generalize and apply the accumulated knowledge to new situations. In the next section, we will review the problem of catastrophic forgetting and how these thermodynamic conditions might provide a strong base to build off for such lifelong learning agents.

6.10 Reliable Low Dissipation and Catastrophic Forgetting

One of the biggest disadvantages of current machine learning algorithms is the problem of catastrophic forgetting - the tendency of an artificial neural network to completely and abruptly forget previously learned information upon learning new information. The problem had been identified decades ago [145], and is a manifestation

of the stability-plasticity dilemma in connectionist networks, which encode information in a distributed manner. It corresponds to the trade-off between the system's ability to hold onto past information when presented with new information (stability), and it's ability to generalize and infer from it's inputs (plasticity). In order to make the jump from our current techniques for narrow intelligence to an artificial general intelligence, it is vitally important to address this problem. Success would allow for the realization of systems that exhibit continuous incremental learning and adaptation to the changing environments, and the capability to generalize and apply the accumulated knowledge to new situations. Ongoing research into new algorithms and architectures have produced progress in transfer and lifelong learning to overcome the stability-plasticity problem, but have achieved limited success [126].

In this section, I will focus on the relationship between catastrophic forgetting and the thermodynamic conditions discussed above. In a Markov finite state automata description of an agent, the input encoding scheme to overcome the stability-plasticity problem in the system would maximize the system's ability to predict future inputs (plasticity) for a finite amount of memory (stability). This can be framed as a constrained optimization problem of the form

$$\max_{Encoding} \text{ (Plasticity) - } \beta \text{ (Stability)}$$

Using mutual information based measures for plasticity ($\mathcal{I}^{\mathcal{R}_1\mathcal{S}}$) and stability ($\mathcal{I}^{\mathcal{R}_0\mathcal{S}}$) in a Markov state machine descriptions, the above optimization problem for encoding $p(k^{\mathcal{S}}|i^{\mathcal{R}_0})$ can be stated as the following

$$\max_{p(k^{\mathcal{S}}|i^{\mathcal{R}_0})} \mathcal{I}^{\mathcal{R}_1\mathcal{S}} - \beta\mathcal{I}^{\mathcal{R}_0\mathcal{S}}$$

We know this to be the past-future information bottleneck problem, and has been suggested as a suitable objective function for lifelong learning [107]. It is shown to be capable of explicitly capturing the underlying predictive structure of a process [108].

In the previous section, we have seen that systems realizing the physical condition of reliable low dissipation $\kappa = \gamma$, is equivalent to implementing the above information bottleneck. While the memory in such systems will be achieved through reliable high dissipation, the ability to generalize and predict new inputs matches up with the optimal solution. Thus physical systems that exhibit reliable low dissipation have the ability to transfer knowledge, overcome the stability-plasticity problem, and offer a compelling candidate system for lifelong learning.

6.11 Discussion & Conclusion

In this chapter, we introduced the non-equilibrium conditions for predictive inference in physical systems using the macrostate fluctuation theorems. The reliable low dissipation condition represents a powerful relationship between homeostasis, energy efficiency and prediction in passive agents. According to the condition, the energy efficient dynamics of a system maintaining homeostasis is the predictive behavior the systems exhibits of future inputs. Furthermore we were able to extract the generative and error corrective components of the predictive dynamics exhibited under the low dissipation condition. In addition to the reliable low dissipation condition, we also added greater context to the reliable high dissipation condition proposed under the dissipation driven adaptation. The reliable low dissipation condition was then further extended to active agents to explain exploitation-exploration trade-offs in the optimal action policy of these agents. Finally we proposed the reliable low dissipation condition as a possible solution to the stability-plasticity dilemma and achieve lifelong learning.

There is a need for a fundamental shift in our strategies to achieve learning and intelligence in computing systems. A possible radical new strategy to move forward is that of thermodynamic computing, an engineering framework that seeks to combine information theory and thermodynamics and place it at the very heart of comput-

ing moving forward. In order to achieve physical systems capable of learning in a dynamic environment, we need to move away from physical implementation of task dependent algorithms, and focus more on processes to realize thermodynamic conditions like those presented in this chapter under which physical systems will exhibit optimal predictive encoding of the external signals. We need to identify more conditions and translate these results into effective guidelines and design principles for self-organization processes. In the next chapter, we will delve in the philosophical underpinnings of underlying our current computational approaches to building intelligent systems and then situate our new thermodynamic approach in this landscape which help us understand these ideas in the larger contet. I will further develop the ideas presented in this chapter and discuss the engineering methodologies required to build thermodynamic computers moving forward.

CHAPTER 7

THERMODYNAMIC INTELLIGENCE FRAMEWORK

7.1 Introduction

The *Rebooting Computing Initiative* was started as IEEE Future Directions Group in 2012 with the stated goal of rethinking the computer "from soup to nuts, including all aspects from device to user interface," working from a holistic viewpoint taking into account evolutionary and revolutionary approaches [115]. This need for and unique opportunity to reboot computing has been driven by two important factors that we mentioned in the previous chapter - the shift in the focus of the computing industry away from traditional logical and mathematical operations towards data-centric applications which require the handling and learning from large amounts of data. The second is the imminent and inevitable end of Moore's law brought about by physical limits to CMOS scaling (death by scaling) and energy efficiency (death by heat). The last time we were here was over six decades ago, when we first started designing computing systems. It was the perfect storm of the right task of performing large mathematical operations meeting the necessary system architecture and CMOS device technology that led to the first computer technology revolution. Needless to say, we have made tremendous progress in those six decades, to find ourselves right where we started - a new task of producing intelligent systems, looking for new architectures and devices to effectively achieve them in.

In the previous chapter, we introduced the non-equilibrium reliable high and low dissipation conditions for adaptive learning and predictive inference in sections 6.6 and 6.7 respectively. These conditions provide an alternative physical description of

intelligence and will look to challenge the central premise of our current computational approaches to building intelligent systems - the brain is a computer and intelligence can be described and achieved through an algorithm. This approach is based on the computational theory of the brain, and will be discussed in detail in the sections to follow. Extremely sophisticated machine learning algorithms have been developed and implemented on traditional computer hardware to very impressive effect. In spite of the massive improvement in performance afforded by machine learning techniques in the recent past, these state of art algorithms are heavily outperformed by the human brain over a wide range of tasks, and with respect to energy efficiency of implementation. The growing consensus in the field is that if human general intelligence and beyond is the goal, we have to look beyond modifying existing techniques and more towards living systems. The next major technological leap will require a fundamental shift away from our current notions on intelligence, and figure out how it can be optimally realized in artificial systems.

The goal of this chapter is to explore the fundamental ideas at the foundations of our current approaches to artificial intelligence and place the thermodynamic constraints from the previous chapter in larger context. Exploration of what computing is and what it entails will help us understand the nature of intelligence that is realized with our current computational approaches. It will also enable us to distinguish this type of intelligence from our own. Identifying the difference will allow us to address it, and explore new theoretical frameworks and approaches to engineering intelligent systems. Just as the first computer technology revolution was built on the theoretical foundations of the Turing machine, it is necessary to ask if we need to look beyond this existing framework, and seek new ideas in order to build efficient intelligent systems and embark on another technology revolution. I must note that this chapter will have a strong philosophical flavor to it, which is unavoidable given the nature of the foundational issues that the chapter looks to tackle. While one might be forgiven for

not pondering about the philosophical implications of all our engineering endeavors (Is a bridge, still a bridge if no one uses it?), that cannot be the case when it comes to artificial intelligence. Such technology has tremendous potential to change the world, and it is necessary that we understand these complex systems we are designing. Furthermore, I hope to show that understanding these fundamental (often philosophical) differences will help us in designing intelligent systems in a more optimal manner.

The chapter will be organized as follows. We will start with a quick review of the computational theory of the brain, and explore the fundamental differences between observer dependent and independent computing. We will then build on the idea of observer dependent computing to talk about the current state machine learning algorithms. Also included are discussions on the differences between these algorithms and the human brain, as well as introduction of a simulation-emulation scale for their implementations. Using the results on the fundamental non-equilibrium conditions for learning from the previous chapter, we will introduce the framework of thermodynamic intelligence and explain the ability of these systems to exhibit observer-independent intelligence. Finally we will explore the realization of thermodynamic intelligence in physical systems under an engineering framework called thermodynamic computing. The chapter concludes with a summary of the results and a brief discussion of important issues to consider moving forward.

7.2 Computational Theory of the Brain

The quest for artificial intelligence lies at the very heart of the field of computing. Alan Turing, one of the earliest pioneers of computer science was very interested in the existence of programs that would make machines indistinguishable from human in their intelligence, and proposed the famous Turing test in his seminal paper "Computing machinery and intelligence" [116]. Turing was also interested in the idea of training connectionist networks for learning and proposed the unorganized Type-A

and Type-B machines [117]. The collection of ideas which understands and describes our intelligence as a computation, being run on the hardware of the human brain is called the computational theory of the brain/mind [118], and forms the central tenet behind all of our current approaches to AI. Under this picture, the computations achieved by the brain to produce our intelligence is the only thing that matters, and the nature of the hardware that produces them is not of consequence. If these computations were properly identified and implemented in any hardware system like our modern day digital computers, the systems achieves the same intelligence as the human brain. In the quest for an artificial general intelligence, it is this last statement that I think is important to analyze and possibly challenge. In order to do that, we have to go back to the very fundamentals of what computing is, and what it entails.

7.2.1 Revisiting the Fundamentals of Computing

Information and computing - the bedrocks of our field, are probably two of the most widely used and often misunderstood terms in science, engineering and philosophy. For most of us, computing in a practical useful sense corresponds to what is achieved by our modern day digital 'computers' - artificial silicon based systems that have been designed to execute programs. Many decades ago we would refer to people as computers, given their ability to compute. This would be one of the reasons for which the title of Turing's seminal paper was 'computing machinery', rather than computers. Turing machines were invented as a way to model our own ability to follow a set of instructions and execute mathematical or logical operations, it heavily influenced the invention of the first computing machines. However over time, with advancements in engineering, we were able to come up with ways to automate this process of executing instructions very efficiently, and the word 'computer' has acquired the meaning it has today. Initially these systems were able to perform simple logical and mathematical operations, but improvements to CMOS technology, archi-

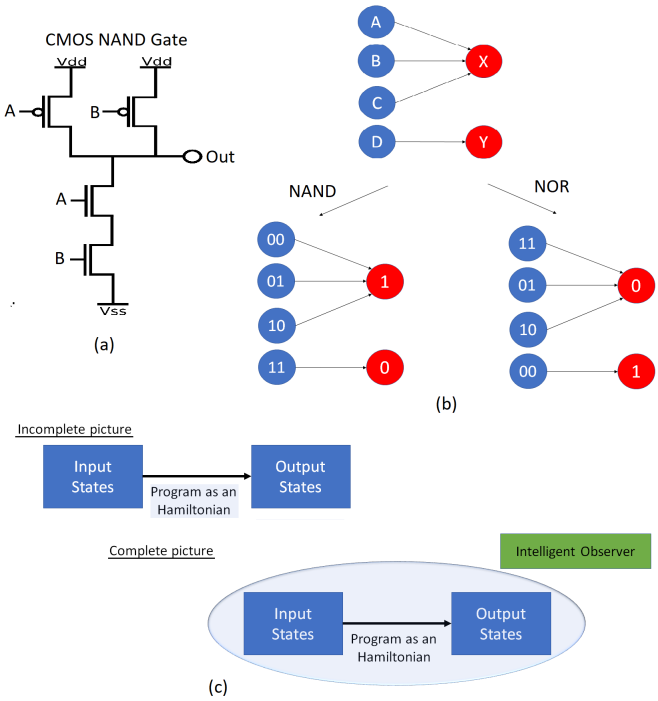

Figure 7.1. (a) Traditional CMOS circuit implementing a NAND gate with inputs A and B, and output Out. (b) A physical systems of 4 states A,B,C and D evolving into states X and Y is observer independent. Different interpretations of the input and output state encodings will realize different NAND and NOR operations. The computing achieved in this case is observer dependent. (c) Comparing the incomplete and complete picture of computing in our digital computers. The external observer who interprets the evolution of a system as a computation is often missed.

tectures and the programs themselves have enabled it to perform a wide range of sophisticated functions, and become vitally important to our everyday life. If we are go to beyond these traditional cases of computing, and look at broader examples like an apple falling from a tree or a pencil at rest on a desk, there are questions to be asked on whether these systems are computing their own time evolution functions? While these are important discussions to be had on the fundamental definitions of computing and their implications, they are beyond the scope of this dissertation. We will focus on computing from this practical useful sense, and use it in our discussion of computing from an *observer dependent* and *observer independent* sense [119].

Consider the CMOS circuit shown in Fig.(1a). There are two input nodes A and B, and an output node Out. From the figure, we can see that if logic 1's and 0's were encoded as voltage highs and lows respectively as it usually is in digital logic, the CMOS implementation behaves like a Boolean NAND gate. For the inputs 00,01 and 10, the system produces the output 1, and the output 0 for the input 11. It is immediately clear that if the encodings were to be modified, the same CMOS circuit can be used to achieve a logic NOR gate. Thus nothing about the physical circuit makes it a NAND or NOR gate Fig.(1b). It is our interpretation of the states of the system as inputs, outputs and determination of the encoding scheme that makes the physical CMOS circuit a system implementing a computation - logical NAND or NOR. While the time evolution of the physical system is an observer independent phenomenon, the computation achieved through the evolution is not. This observer dependence of computations achieved in artificial systems is very important, and often very easy to miss given the sophistication of these systems. In Fig.(1c), we can see that when we think of computing systems, we restrict ourselves to inputs, outputs and the process that produces the outputs from the inputs. However all of these are defined and interpreted by an observer, external to the system. When it comes to our modern day computers, nothing about physical system in on itself, imbues them with the property of computation. It completely comes down to the capabilities of this external observer to interpret the evolution of a physical system as an instantiation of a faithful computation. We are these observers who interpret the keystrokes on a keyboard as input, and the symbols on a monitor as outputs, thus making the processes that occur in the CPU to convert the inputs to outputs as computation. It is our interpretation that makes all these machines around us 'computers'. And our intelligence plays a significant role in not only interpreting the evolution of systems as computations, but also manipulating the physical structure of systems to achieve

the computations we desire (all of silicon based integrated circuits fall under this category).

In contrast, our own ability to perform computations is observer independent. We do not need an external observer to interpret and decide when we have or have not achieved a computation. This ability to perform observer independent computations is a vital characteristic of our 'true' intelligence, and fundamentally different from that exhibited by our computers [119], [120]. An important question that one should be asking immediately would be - what is it about our brains and our intelligence, that allows us to compute in an observer independent manner and also understand observer dependent computing?How do we understand and achieve this in artificial physical systems without resorting to observer dependent approaches again? One possible answer that has been suggested is that consciousness bestows ourselves with the property of observer independence. Consciousness is an extremely complex phenomenon, and often defined so broadly that attributing the observer independence property to it is simply *explaining away* the problem, without really addressing it. We would be explaining one phenomenon that we do not fully understand by attributing it to another phenomenon that we understand even less of. Rather than referring to the broad phenomenon of consciousness as a solution, I will instead use recent results on the thermodynamic conditions for intelligence citeNatesh to propose explanations of this unique ability.

7.3 Computational Approach of Machine Learning Algorithms

In the previous section, we introduced the distinction between observer dependent and independent computing. Before we move on to thermodynamic intelligence, it is important to understand where and how our current machine learning approaches fit in this picture. Our recent success with machine learning techniques have only served to reinforce the computational theory of the brain - that our intelligence can

be described as an algorithm, and achieved by implementing in digital computers. It is amazing how far we have come since the turn of the decade when the use of specialized hardware - graphic processing units (GPU), made the execution of resource intensive learning algorithms extremely viable. Significant improvements in image and object recognition by different research groups in 2011 and 2012 anchored the start of what is now known as the "deep learning revolution." In the five plus years since then, it is an understatement to say that the field of machine learning has grown exponentially. Artificial neural networks have significantly diversified and applied successfully in vastly different fields. AI has become quite the buzzword for both established companies and new startups looking to disrupt the market. While there continues to be large number of different approaches and architectures being developed rapidly, these techniques can be broadly classified as supervised, unsupervised and reinforcement learning [121]. The area of supervised learning is the most popular one, and where a significant amount of existing work has been carried out. Reinforcement learning has become extremely popular over the last couple of years, especially with the success of Google DeepMind's AlphaGo system. A good portion of unsupervised learning is currently carried out by converting it into a supervised learning problem. However moving forward, the general consensus is that unsupervised learning should form the bedrock of AI systems going forward, and the focus of the field needs to shift towards improving it.

The ability of these techniques to generate rich functionality through the training data has produced a general misunderstanding about the nature of these machine learning techniques. Machine learning software can be built by assembling networks of parameterized functional blocks, and by training them from examples using some form gradient based optimization. The end result of this process is very much similar to a regular program, except it is parameterized, automatically differentiated and trainable. And like regular programs, they can be implemented and compiled

using a programming language and compilers. In recognition of this, the field has rebranded themselves recently from deep learning to *differentiable programming*. Machine learning techniques are simply a sophisticated form of programming. They are not fundamentally any different from other traditional programs that have been run on computers over the years. Our current approach to AI can thus be summarized as *intelligence through observer dependent computing*. This has tremendous implications when it comes to analyzing the nature of intelligence produced by systems executing these programs.

7.3.1 Intelligence Through Computing

In order to understand the idea of intelligence through computing better, let us analyze the technological landscape [123], where we have been successful, and the techniques used to achieve that success. From Fig.(2), we can see that the area of most interest right now is on the top right-hand corner of the technology landscape, with systems that are capable of real time learning of information in a dynamic real world environment, use of old information to predict new ones and being able to transfer that knowledge across domains - characteristics of human general intelligence. The two areas to the left of the y-axis correspond to traditional computing applications, where static and dynamic modeling techniques have been extremely successful. The lower right quadrant has employed differentiable programming implemented in traditional hardware like GPUs to learn in a static offline manner, and represent the current state of the field. We have been using a mix of these existing techniques such as differential equations, search algorithms, deep learning, probability and statistical methods that have been successful in these other quadrants to the top right one, with limited success. The question we need to address is - are these techniques sufficient? If not, where will our new ideas come from?

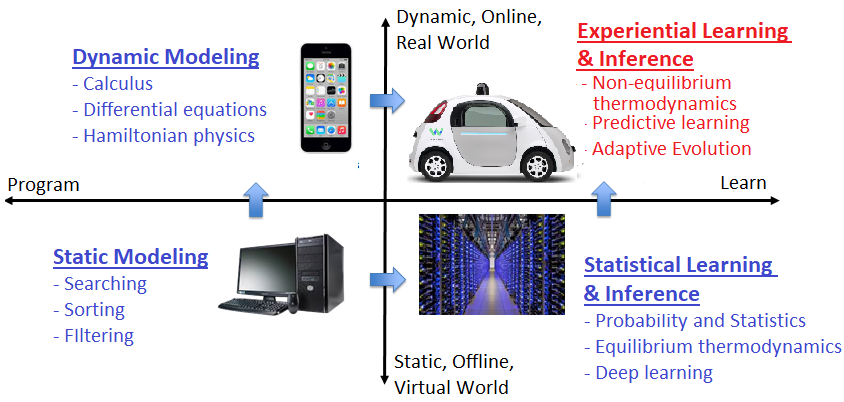

Figure 7.2. Our current computing applications can be divided according to a technological conceptual landscape. This landscape is divided between applications that are programmed and those which learn on the x-axis, and between online dynamic environment vs a static offline one on the y-axis. The conceptual landscape discusses the various foundational techniques that are used to achieve success in the corresponding area of the graph. The top right corner is the area of intense interest and ripe for exploration [123].

While discussing the difference between observer dependent versus observer independent intelligence, I will limit myself to the standard implementation of machine learning algorithms in traditional hardware like GPUs for the observer dependent case due to two major reasons - this particular implementation accounts for a significant proportion of AI systems achieved currently. Secondly, it is necessary to establish a very clear example of observer independent intelligence which is offered by this implementation. Once we have dealt with cases that are black and white, we can then move onto novel neuromorphic and self-organized computing systems which represent different shades of gray. The failure to account for the external intelligent observer produces major confusion about the nature of intelligence achieved by the learning programs from the lower right quadrant. Once the dependence of these computing systems on external intelligent observers is understood, it becomes immediately clear what the problem with computational algorithmic approaches to intelligence are. *We need observers who are themselves intelligent, to interpret the underlying physical processes as computing, in order for the overall system to be perceived as*

intelligent. The intelligence of digital computers implementing learning algorithms, and exhibiting appropriate input-output behavior are intelligent observer dependent by definition. Such systems lose their computing capability, and as consequence their 'intelligence' in the absence of an intelligent observer. All our current approaches to intelligence in artificial systems suffer from this fundamental condition. If the human race vanishes tomorrow, the flow of electrons through an integrated circuit or symbols appearing on a LED monitor will have no computational significance, let alone intelligence. It would be overly optimistic to expect observer independent intelligence to simply emerge from observer dependent computing, by increasing the computing resources or just implementing more complicated algorithms.

These systems can be thought of as mimicking human intelligence, albeit quiet poorly as it stands in the present. I would like to go back to the Turing test once again, in which the goal was to produce a computer program that can fool an human participant into believing that the program is human. By allowing for an intelligent human observer to participate, the test automatically allows for an observer-dependent intelligence to pass this test. However the test itself is not suitable to draw the distinction between a program mimicking our intelligence, and a system like us exhibiting observer independent intelligence. If an algorithm for general intelligence did exist, that when implemented on traditional computer hardware would respond to inputs in a way that is indistinguishable from a real human, then we cannot know the nature of system's intelligence - whether it is observer dependent or independent from studying it's input-output behavior alone. We would need knowledge on whether or not the underlying mechanisms inside the system employ observer dependent computing or not.

One is tempted to say that while these thought experiments are interesting, it does not really matter if our artificial systems exhibit observer independent intelligence like us, or merely mimic it in an observer dependent manner. This is a valid

objection to raise, and for benign applications like movie and music suggestions, simple text translation, etc. the difference does not matter as long as the system achieves the necessary input-output relationships. This is not the case for more serious applications of machine learning like self-driving cars, robotic surgery and healthcare insurance and management that we are attempting right now. There is a real danger that without solid definitions and understanding of these fundamental concepts, the entire field of AI will suffer from the hype produced by often spurious claims of consciousness, personhood [124], legal rights and attribution of human emotions [125], brought about by anthropomorphization of these systems. Artificial intelligence is a technology that is capable of revolutionizing every aspect of society and our lives. As with most groundbreaking technologies, it is accompanied by a large number of ethical and legal issues that needs to be studied and addressed before making important policy decisions in the public and private domains. Vital to these decisions will be a sound theoretical framework of intelligence - human and machine.

7.3.2 Machine Learning vs The Brain

It is important to understand that while artificial neural networks are based on a computational algorithmic description of the brain, they are not heavily inspired by it. The fundamental idea that have made neural networks, especially deep neural networks with large number of hidden layers successful and viable is gradient based optimization of task-dependent objective functions, using the backpropagation algorithm. The backpropagation algorithm allows for calculation of the required error gradients, starting from the output end and distribute it across the network layers to determine the appropriate weights for error minimization. While we might continue to produce good results in image and speech recognition, playing games, and or in limited domain scenarios - tasks in which humans have performed much better than our computers historically, the techniques used to achieve these results are not

based on actual processes that happen in our own brain. There is very little evidence that backpropagation like mechanisms play a vital role in our intelligence. At best, these techniques can be seen as very coarse descriptions of mechanisms that produce intelligence in our biological brains. As a result, these algorithms often suffer in terms of their performance across a wide variety of tasks and energy efficiency (There is growing work now to improve upon knowledge transfer between tasks, with some techniques looking to hippocampus-neocortex interactions in the brain for inspiration [126]). They usually need millions of labeled training examples to learn for even simple tasks, large amounts of compute power for training and often struggle to transfer knowledge between tasks. These systems are also extremely susceptible to adversarial examples leading to catastrophic failures. Given the vast complexity of some of these networks, with millions of nodes and weights, it often extremely hard to understand their data representation schemes, the exact functioning of these networks and the specific conditions under which they will fail. The race to build and deploy systems in as many areas as possible, has sometimes devolved into an exercise of rapid flag planting, without sufficient interest to build and/or test a solid theoretical framework to answer all these above questions only further exasperates the problem. However there is light at the end of the tunnel, as more researchers in this community recognize this problem, and are taking steps to address it [127], [128]

7.3.3 A Simulation-Emulation Scale

These shortcomings should encourage us to ask a fundamental question on whether the current computational approach is the optimal way to engineering artificial intelligence? This issue can be pictured on a simulation-emulation scale of implementations as indicated in Fig.(3), with complete emulation (biological brain) at one end of the scale, and an extremely detailed simulation on the other. The detailed simulation approach involves detailed models of the microscopic ion channels, flow of neurotrans-

mitters, expression of proteins, etc [129]. This becomes extremely resource intensive very quickly, and definitely not scalable for commercial purposes. While being extremely useful, especially in the study of brain disorders and testing the effects of pharmaceutical drugs, they will not be a viable alternative to current machine learning approaches to solve a wide variety of tasks. On the other end, total emulation will involve figuring out how to grow a biological working brain and using it to perform tasks that a general intelligence is capable of. While steady progress is being made on this front through the growth of mini brains [130], given the slow speed of biological neurons and the issue of their viability outside the human body (not to mention the ethical issues that arise in this scenario), this does not seem as a suitable engineering option in the short term. Our computational machine learning algorithms running on traditional hardware, will lie (somewhere near the middle) on this scale closer to the simulation end. Neuromorphic chips exploiting novel architectures and devices like memristors [131], would offer an efficiency boost and would also lie near the middle, closer to emulation end. Systems which use utilize biological neurons [132] on a chip for computing offer a very interesting prospect, and should be closer to the emulation end than the simulation one. We can immediately notice the significant increase in the computing resources required for achieving intelligence as we move from the human brain (total emulation), through neuromorphic chips and machine learning algorithms on traditional hardware, to detailed computational models (total simulation). In the next section, I will discuss the alternative thermodynamic picture for intelligence, with the goal of producing *thermodynamic computers*, that would be closer to the emulation end than any of our current systems.

7.4 Thermodynamic Intelligence

The broad nature of intelligence is overtaken only by consciousness itself. In order to present an alternative to the computataional view on intelligence, it is necessary

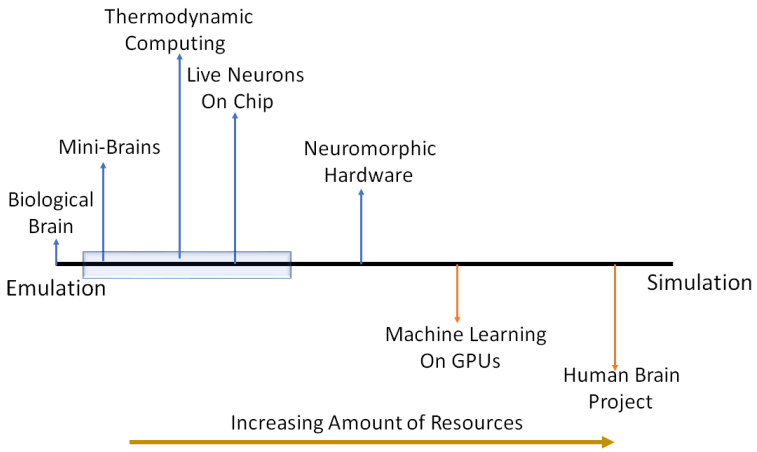

Figure 7.3. A Simulation-Emulation scale for various implementations of intelligent systems. The Human Brain Project, machine learning algorithms lie closer to the simulation end. Neuromorphic hardware, mini-brains and biological neurons on a chip lie towards the emulation end. Thermodynamic computing looks to produce viable systems closer to the emulation side of the scale.

to define as it pertains to this paper. We will use a prediction centric definition of intelligence - a system is intelligent if it can learn and predict future inputs based on past inputs. There is another important component to intelligence that we will refer to as *adaptation*, the thermodynamic conditions for which have been discussed in detail in [133]. Biology is incredibly diverse and complex, especially when it comes to the human brain. There are over a hundred known neurotransmitters, different types of neurons, neurogenesis and pruning, glial cells and different types of important plasticity mechanisms. While it is important to study these systems in detail, they do not offer a viable way to emulate intelligence in artificial systems. This is where these thermodynamic conditions offer significant advantages. These fundamental conditions will be applicable to biological systems, as well as provide an approach to building intelligent systems of the same nature with inorganic material. Note that while being useful, the ideas presented here are in its infancy and there is much work to be done, including identifying the temporal and spatial mechanisms in which these conditions are realized in biological systems. In order to establish the relationship

between thermodynamics, information theory and intelligence necessary for a theory of thermodynamic intelligence, we will employ the thermodynamic conditions from the previous chapter.

7.4.1 Dissipation, Homeostasis & Intelligence

In previous chapter, the author used the fact that since homeostasis is the maintenance of a macrostate I as the system is driven by an external field, it can be realized in the macrostate fluctuation theorem by using the relation $II = I$ in Eq(6.1). This allowed the derivation of a general thermodynamic condition for homeostasis in a complex self-organized system $\mathcal{S}$ ([138],[139]) driven by external signals $\mathcal{R}_0\mathcal{R}_1$. A special case of this physical condition called *reliable low dissipation*, was shown to be equivalent to the system implementing the solution to a constrained optimization called the Information Bottleneck [140], [141].

$$Maximize_{p(k_t^{\mathcal{S}}|i^{\mathcal{R}_0})}(\mathcal{I}^{\mathcal{R}_1\mathcal{S}} - \lambda(\mathcal{I}^{\mathcal{R}_0\mathcal{S}}))$$

where $p(k_t^{\mathcal{S}}|i^{\mathcal{R}_0})$ is the probability that the i-th past input $\mathcal{R}_0$ maps to the k-th state of the homeostatic system $\mathcal{S}$ at some time t, and quantifies the encoding of external signal in the states of the system. $\mathcal{I}^{\mathcal{R}_0\mathcal{S}}$ is the mutual information between the system $\mathcal{S}$ and the past inputs (history) $\mathcal{R}_0$, and is a measure of the system's finite complexity. $\mathcal{I}^{\mathcal{R}_1\mathcal{S}}$ is mutual information between $\mathcal{S}$ and the future input $\mathcal{R}_1$, and is a measure of prediction in the system. Thus homeostatic systems that satisfy the condition of reliable low dissipation achieve signal encoding that maximally predict the next driving input. This predictive encoding of external signals will form the basis of a predictive thermodynamic intelligence in the system.

The state transition probability $p(l_{t+1}^{\mathcal{S}}|k_t^{\mathcal{S}}, j^{\mathcal{R}_1})$ of a system satisfying the condition of reliable low dissipation are of the form given below

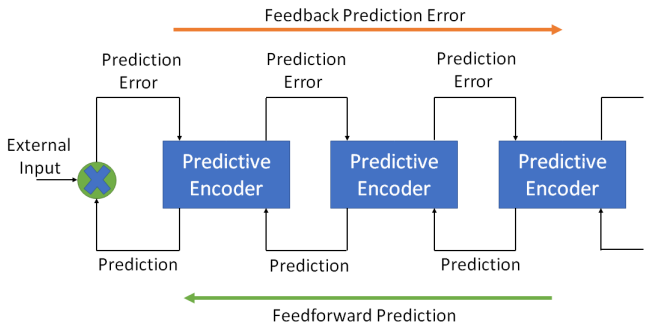

Figure 7.4. Hierarchical predictive coding architecture with the feedforward predictions moving from higher levels to lower levels, and feedback prediction errors moving in the opposite direction. The higher levels predicts the level below it, and the prediction of the external signal at the lowest level interacts with the external signal to general the prediction error. Propagating only the error signal in a feedback manner is more efficient.

$$p(l^{\mathcal{S}}_{t+1}|k^{\mathcal{S}}_{t}, j^{\mathcal{R}_1}) \propto e^{\frac{-1}{\lambda}\{D_{KL}[p(m^{\mathcal{R}_2}|i^{\mathcal{R}_0}j^{\mathcal{R}_1})|p(m^{\mathcal{R}_2}|l^{\mathcal{S}}_{t+1})]\}} \times e^{\frac{1}{\lambda'}\{D_{KL}[p(j^{\mathcal{R}_1}|i^{\mathcal{R}_0})|p(j^{\mathcal{R}_1}|k^{\mathcal{S}}_{t})]\}}$$

From the equation above, we can see that state transition probabilities $p(l^{\mathcal{S}}_{t+1}|k^{\mathcal{S}}_{t}, j^{\mathcal{R}_1})$ for a system satisfying reliable low dissipation has two components - a feedforward 'generative' component (first term on the right hand side) that continues to maximize prediction, and a feedback 'recognition' component minimizing errors (second term on the right). This form of feedforward-feedback architecture (Fig.(4)) is used in Helmholtz machines [143] and variational encoders [144] - popular unsupervised machine learning techniques. But unlike machine learning algorithms, the predictive intelligence that arise from the input encodings in the states of the system here is the inevitable result of satisfying the thermodynamic conditions. This predictive coding scheme has also been explored as a solution to the stability-plasticity problem [145] in [133].

These results are closely tied to a related framework called the free-energy principle [146] that looks to explain the underlying principle of the brain, and can be viewed as a variational Bayesian approach to understanding perception and action. It defines homeostasis, as a minimization of internal entropy of the system - which is a much more specialized condition than the non-equilibrium conditions discussed above. The free energy principle has been well developed over the years in neuroscience, and can be reduced to the Information Bottleneck approach as well [147]. Perception and action in the brain are studied under an hierarchical predictive coding architecture [148], as shown in Fig.(4). Learning and our intelligence is the combination of feedforward predictions of the incoming sensory signal, and feedback of the prediction error updating the system as indicated in the equations before. An extension of the results presented above to include for active agents is derived in [133] (and in the previous chapter) and the exploitation-exploration trade-offs studied.

7.4.2 Observer Independent Intelligence

In the previous sections, we discussed how intelligence was achieved through the use of observer dependent computing under a computational approach. We will now discuss how the thermodynamic conditions above could offer a possible explanation for observer independent intelligence in systems satisfying the conditions. In observer dependent computing and intelligence, the evolution of a physical system and it's interaction with external signals become an instantiation of specific computations when the states of the system are interpreted as inputs and outputs. For implementation of machine learning algorithms, this would require an external observer to choose what the encoding of the inputs that needs to be learned into the states of the system, and interpreting the evolution of the system as learning. This would make the intelligence exhibited by such systems observer dependent. One should expect an ob-

server independent intelligent system to achieve this without the need for an external observer.

Imagine a system satisfying the thermodynamic conditions of reliable high and low dissipation when driven by external signals. Only those signals interacting with the system to satisfy the above conditions either become 'memory' or predicted and corrected for by the system depending upon which condition they satisfy, and thus effectively learned. Say the system $\mathcal{S}$ is driven by an external signal $\mathcal{R}_A\mathcal{R}_B$, in which the system satisfies the thermodynamic conditions only with respect to the signal $\mathcal{R}_A$. In this case, the system makes predictions $\mathcal{R}_A$, and not $\mathcal{R}_B$. If it satisfies the reliable high dissipation condition with respect to $\mathcal{R}_B$, then the system has increased it's correlations with $\mathcal{R}_B$ and hence learns a memory of it. An observer independent intelligent system doesn't choose the signal that it wants to learn and predict, and then satisfies the corresponding thermodynamic conditions for with respect to the chosen signal(s). *Those signals that satisfy the thermodynamic condition of reliable high and low dissipation, appears as being chosen by the system $\mathcal{S}$ to become the inputs to be learned and predicted by the system - forming the basis for independence from an external observer.* To an external intelligent observer who is studying the system, it will appear as though the system is exhibiting predictive intelligence with respect to input $\mathcal{R}_A$, and the choice of input was made by the system, and thus dependent on the system only. Note that, while only those signals that interact with any physical system can be considered as inputs that affect computing, those that satisfy the specific conditions would become the inputs for learning in observer independent manner. It is necessary to state that the human brain is a vastly more complex system than what we have discussed here, resulting in extremely rich behaviors. There are broad range of phenomenon like awareness, attention and subjective experiences related to consciousness that an observer independent intelligence like ourselves are capable

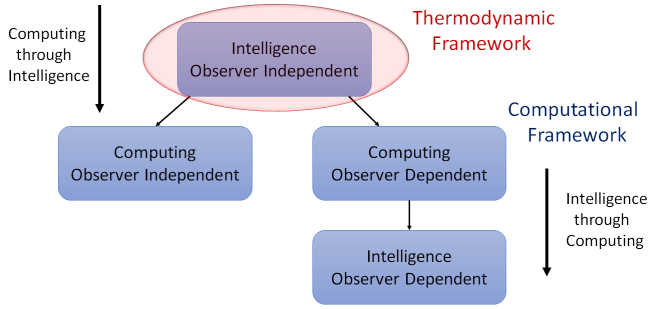

Figure 7.5. The figure indicates the bigger picture noting how observer dependent/independent computing and intelligence are related together, as well as the distinction between *computing through intelligence* and *intelligence through computing*. The thermodynamic framework looks to address observer independent intelligence, while the computational approaches produce observer dependent intelligence.

of. While the ideas presented here does not capture all this richness, the underlying concepts provide a strong base to build upon for future work.

7.4.3 Computing Through Intelligence

Once the observer independent intelligence of a system is established, it is straightforward to see how the system is capable of performing observer independent computing. The thermodynamic conditions above imply predictive learning in a system as it continues to be driven by external signals. A system with sufficient memory can learn numbers, simple logical and mathematical functions as a temporal prediction task. The ability to observe and learn these functions like patterns allows the system to calculate outputs, without having to employ the same algorithms that our digital computers do. This form of *computing through observer independent intelligence* will be observer independent. Performing addition is a pattern recognition task of learning numbers, the addition symbol, and being able to recognize and predict patterns in the results of the addition operation, independent of an external observer. Notice that it is easier to compute large calculations in your brain, when the inputs allow for specific patterns in the outputs that the system can exploit to predict. Multi-

ples of 5000 are a lot simpler to calculate than 4857. Of course, intelligent systems like ourselves are also capable observer dependent computing - by manipulating and interpreting external systems to be used as computational tools - from rudimentary forms like a piece of pen and paper, an abacus to sophisticated supercomputers. The information learned by the system, directly affects the observer-independent computation that the system performs, as well as the interpretation of external systems as performing a computation. It is not possible for a person to add or multiply two numbers without knowing what numbers or addition and multiplication are. For example, say the system $\mathcal{S}$ that has only learned the AND function. It will only ever be able to interpret the evolution of a system with 4 states, with 3 mapping into one output as a AND implementation. This would be in contrast with a system that has knowledge of both the AND and OR logical functions, which can then choose to interpret the same physical evolution of the system in one of two ways.

In this section, we introduced the framework of thermodynamic intelligence using the fluctuation theorem conditions for predictive intelligence in an homeostatic systems. These conditions were then used to discuss how systems satisfying these conditions can exhibit observer independent intelligence, and be capable of achieving observer independent computing through this intelligence as shown in Fig.(5). Now that we have established the fundamental distinction between our intelligence and that exhibited by current learning algorithms, I will discuss a new engineering paradigm that will seek to leverage the conditions for thermodynamic intelligence to build more efficient learning systems.

7.5 Engineering Thermodynamic Intelligence

In order to discuss the new paradigm of thermodynamic computing, as a more effective approach to engineering intelligence in our artificial systems, it is important to first look at the differences between top-down designed and bottom-up intrinsic com-

putation. As discussed in detail before, our current approaches to building intelligent machines are heavily rooted in a computational picture using learning algorithms. They represent a top-down strategy of designing systems to perform the computations required for learning and intelligence. Designed computation refers to strategies that encompasses a significant portion of our current approaches, especially in the transistor paradigm where we build digital switches using the precise patterning of lithographic techniques. In a top-down approach, we start with the big picture, an overview of the system is formulated, specifying, but not detailing, any first-level subsystems. Each subsystem is then refined in yet greater detail, sometimes in many additional subsystem levels, until the entire specification is reduced to base elements. Top-down design is also characterized by large amount of control of the underlying microstructures by the engineer. When it comes to computing, this translates to deciding on the higher level input-output behavior to be achieved, and then figuring out the underlying architecture to do so. This process continues to lower register and gate levels, before reaching that of the elemental devices which are usually logic switches achieved using CMOS. Top-down design of computing systems produces systems that realize observer dependent computing and intelligence, since the designer is the observer manipulating and interpreting the physical system as implementing computing.

Intrinsic computation refers to leveraging a system's internal spatial and temporal dynamics to achieve useful information processing. The challenge to overcome here would be to understand the non-linear dynamics of the system used for intrinsic computing at both an individual device and network level in order to use them effectively. These systems are characterized by a bottom-up approach to their design. Bottom-up approach to design is less precise than top-down, and achieved through the self-organization of a large number of interacting basic building blocks, by piecing together of systems to give rise to more compex systems, thus making the original

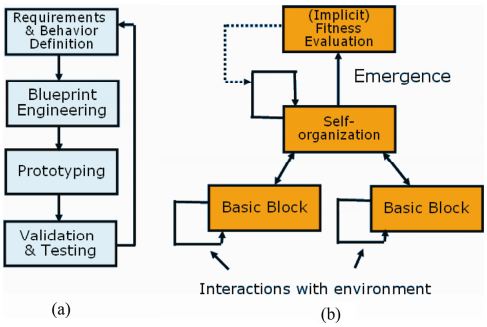

Figure 7.6. Steps in a (a) Top-down design process. (b) Bottom-up design process.

blocks, subsystems of the emergent systems. The goal of the engineer would be to direct the self-organizing processes towards systems with useful temporal dynamics that can be leveraged for the task at hand, without exercising significant control of the underlying microstructures. Detailed discussion of these ideas are available in [149], [150].

Bottom-up approaches can be used to achieve systems capable of both observer dependent and independent computing. In reservoir computing for example, we utilize complex self-organized systems and interpret their input-output behavior to realize useful observer dependent computing. In the previous section, we discussed that the dynamics of a complex self-organized system that satisfies the condition of reliable low dissipation condition can exhibit predictive learning. In [**?**], the author discussed the condition of *reliable high dissipation* that allows for the system to exhibit *adaptation* as well. Bottom-up engineering of systems that satisfy these conditions, which we will call thermodynamic computing will be capable of observer independent intelligence, and by extension observer independent computing. Instead of a top-down approach of describing and implementing intelligence as a computational process, thermodynamic computing will look to engineer a non-equilibrium system - a *thermodynamic computer* in a bottom-up manner to satisfy specific macrostate conditions of thermodynamic intelligence, that will allow the system to exhibit intelligent behavior.

As stated before, these ideas are still in their early stages, and there lies a number of major challenges and opportunities ahead of us. A significant one is that of translating these thermodynamic conditions into design principles that can be physically realized and observed in self-organizing systems. While we will require to make a major effort towards interdisciplinary research in materials, fabrication techniques and network dynamics to produce systems of interest, there is nothing in principle that would prevent us from engineering such systems. In order to improve the efficiency of machine learning implementations, we have moved away from the traditional von Neumann architectures towards memristive crossbar architectures that have become the standard for neuromorphic computing. We must be now willing to look past these crossbar systems, and towards self-organized dendritic and small world networks that might offer even greater efficiency. The need to precisely control these systems arises fundamentally from the fact that we are trying to produce intelligence through observer dependent computing in a top-down manner, and the control necessitates well designed architectures. If general intelligence is the goal, we must be willing to trade on control of underlying microstructures to achieve the required system level behavior through bottom-up approaches. *If the brain's architecture can be convoluted and 'messy', why should our AI systems not be?*. There are already existing technology realizations that can provide an excellent base to build our thermodynamic computers on. A very promising option would be self-organizing atomic switch networks with memristive properties [151], [152] (Fig. 7.7), [153] possessing high device density, rich non-linear properties and critical network behavior, and have already been used for neuromorphic and reservoir computing applications. Successful realization of the conditions of thermodynamic intelligence using such networks will not only experimentally validate the larger framework, they will also signify a massive step towards the goal of achieving observer-independent intelligence like our own in an artificial system.

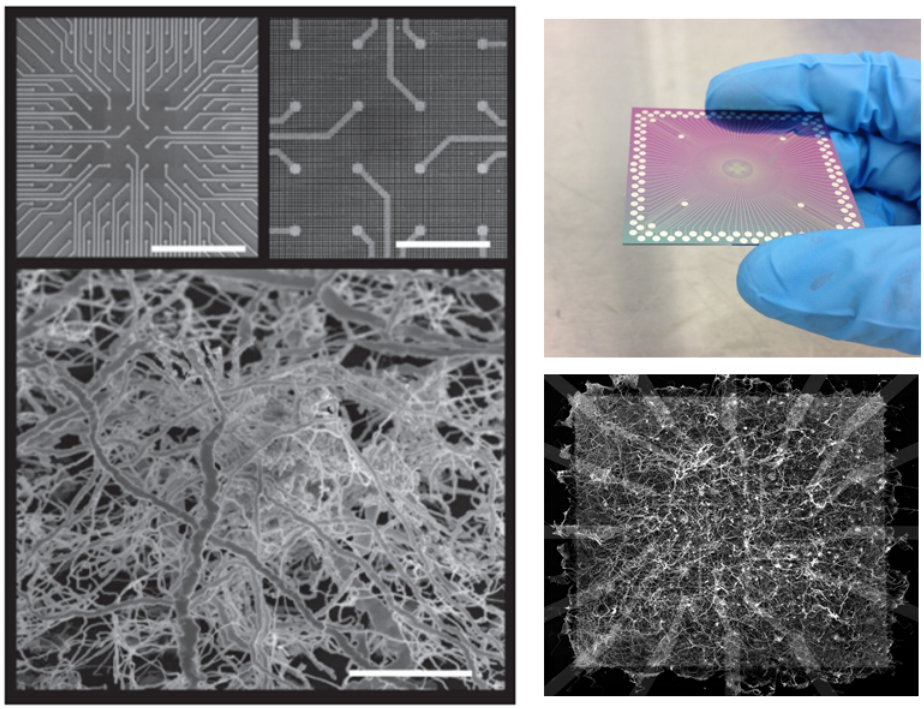

Figure 7.7. Neuromophic atomic switch networks made using Ag nanowires, fabricated using a combination of top-down and bottom-up self-organizing processes. These networks exhibit edge of chaos behavior and used for simple time-series prediction tasks [151], [152].

7.6 Summary & Conclusion

Given the limitations of our current intelligent systems, the goal of this chapter was to analyze the fundamental ideas of computing underlying these approaches in order to identify the type of intelligence produced by such systems. The intelligence through computing approach would realize an observer dependent intelligence, that is fundamentally different from the observer independent intelligence that humans are capable of. I then presented an alternative framework of thermodynamic intelligence drawing on recent results from the field of complex systems, non-equilibrium thermodynamics and predictive learning. The condition of reliable low dissipation was used to propose explanation of observer independent intelligence that systems satisfying these conditions would be able to exhibit. Building off this idea, observer independent

computing through intelligence was explored. These technical results were then used to discuss a bottom-up approach to realizing thermodynamic intelligence in systems called thermodynamic computing, that might represent a more optimal strategy to building these systems than our current approaches. As the focus of the computing industry continues to shift towards intelligent systems and with the end of Moore's law upon us, we are presented with the unique challenge of needing to look beyond traditional ideas that have served us well for long and the unique opportunity to explore novel methodologies across the stack.

In addition to the changes in our engineering approach that is brought about by the distinction in intelligence produced by current machine learning systems and ourselves, there are also significant legal and ethical implications that were hinted at before. I will conclude the chapter here, by quickly mentioning those and leave a detailed discussion for future work. Our current artificial intelligence systems no matter how advanced a machine learning algorithm it is implementing, are observer dependent and lose their intelligence sans our interpretation. This implies that they are very different from us, and very much like our laptops, calculators and other machines. If we cannot consider personhood and the rights that come along with it to our calculators, then we cannot do so to any of our current machine learning systems, even if they possess sophisticated input-output behavior. This also implies that these machines cannot be held responsible and liable when they make catastrophic errors. Just as we do not attribute the responsibility of an accident caused by a badly designed brake pad to the car itself, the same should be true of the machine learning algorithms employed in self-driving cars. Like other systems designed by engineers, the ultimate responsibility of these observer dependent intelligent systems lies with the system designer. The complex nature of these systems, and the ambiguity in the nature of intelligence produced by them cannot serve as cover for poor engineering practices. A

solid framework of intelligence is necessarily required in order to make sound policy decisions on these AI systems that are being deployed all around us.

CHAPTER 8

SUMMARY AND FUTURE WORK

The goal of this dissertation was to study the fundamental limits of dissipation associated with state transitions in finite state automata, and then use these limits to understand the type of state transitions and functional mappings that minimize dissipation in physical systems. FSAs are powerful abstract tools of computing and can be used to characterize different information processing operations. In Chapter 4, I started with a physical description of a deterministic irreducible FSAs under the referential approach to physical information theory in section 4.1 and quantified the dissipation bound for steady state transitions in terms of the information loss about the external driving inputs in section 4.2. Section 4.2 also included a discussion of irreversibility in FSA. The analysis was further extended to account for output generation in Moore and Mealy machines in sections 4.3 and 4.4 respectively. I then derived the dissipation bounds for broader class of probabilistic FSAs in section 4.5 and introduced computational efficacy measures for FSAs inspired by the same for $\mathcal{L}$-machines in section 4.6. The chapter concludes with derivation of the lower bound on dissipation in FSAs with temporally correlated inputs in section 4.7, a result that will continue to be used throughout the dissertation.

The lower bounds developed for these finite state automata were extended to neural networks implementing learning algorithms, to understand the limits of dissipation associated with learning in chapter 5. Learning systems are an important focus of the computing industry at the moment, and these algorithms have to ultimately be implemented in physical systems. It is thus crucial to understand the

limits of dissipation for these systems, and the parameters upon which they depend on. We characterized the weight changes in the neural network according to a learning rule as state transitions in determinsitic FSAs, and used this model to determine the lower bound on dissipation associated in the training testing phase of a simple perceptron learning a classification task in section 5.3. The effect of learning rates and training data distribution on the dissipation bound were also analyzed in the same section. The analysis using the FSA formulation was then extended to study the dissipation associated in recurrent neural networks, specifically with the use of Hopfield and Boltzmann networks as content associative memories in section 5.4. A dissipation bound for simulated annealing was derived and analyzed for different annealing schedules in the same section. We also proposed formulating a learning algorithm with the energy dissipation of network as the sole learning objective function in section 5.5.

The insight gained from understanding the limits of dissipation in neural networks allowed exploration of the very fundamental connections between thermodynamic quantities such as energy dissipation and entropy, with the emergence of learning in physical systems in chapter 6. In the recent past there has been increased interest in using non-equilibrium thermodynamics to characterize complex biological systems. In the same vein, we used the macrostate fluctuation theorems to present the reliable low dissipation condition that quantifies the relationship between self-organization, homeostasis, minimal dissipation and predictive inference for passive agents in section 6.5. In section 6.6, we used recent results on the reliable high dissipation condition to discuss adaptive learning in physical results. The strength of thermodynamic conditions of intelligence - predictive inference and adaptive learning - lie in their independence of any specific realization, and hence offer a perfect base to build a theory of intelligence in artificial systems to be based upon. These results were also extended to active agents in section 6.7 to study exploration-exploitation trade-

offs in the optimal policy. Initial work on the relationship between the reliable low dissipation condition and catastrophic forgetting is discussed in section 6.8.

Chapter 7 focuses on the fundamental assumption that underlie our current computing paradigm and the computational approach to intelligence in the brain. Understanding some of the philosophical basis of computing is necessary, since they have determine our design thinking and engineering methodologies. In section (7.2), the distinctions between observer dependent and independent computing is drawn and our ability to perform the latter is discussed. The status of the current machine learning approaches with respect to these issues is explored in section (7.3). Based on the thermodynamic conditions of intelligence from the previous chapter, a new bottom-up framework of physical intelligence called thermodynamic intelligence is proposed in section (7.4). There is also detailed discussion on why thermodynamic computing can produce observer independent computing. I conclude the chapter with a discussion on the necessary changes to devices, network architectures and design philosophies that is required moving forward to build a thermodynamic computer in section (7.5).

The future is indeed very bright for the field of computing as these novel emerging technologies mature and new application spaces open up. We are at a very unique moment, given the opportunity to be able to define new paradigms as we seek to create another technological revolution. The dissertation reflects the author's own intellectual journey from characterizing the limits of dissipation in computing to questioning the very fundamentals of it in order to understand the important connections between thermodynamics and intelligence. There is much left to be done moving forward with a lot of exciting research questions left to be answered.

The following are important questions that need to be addressed for us to make progress in building efficient intelligent systems. I leave this for future work, and will not be addressed as part of this dissertation.

(a) Improving the non-equilibrium fluctuation theorems based on existing work on thermodynamic lengths and information geometry to describe biological systems more accurately.

(b) Mapping of the thermodynamic conditions into physical observables that can be experimentally tested and verified on suitable test-bed systems. The neuromorphic atomic switch networks provide such a suitable test-bed system.

(c) Understanding the different spatial and temporal scales at which the thermodynamic conditions apply at, thus understanding the relevant scales at which intelligence is realized in the brain.

(d) Identifying optimal devices and architectures necessary for a thermodynamic computer. Mapping the thermodynamic conditions into design constraints.

(e) The effect of physical temperature and noise on learning, and determining the optimal temperature for intelligence.

(f) Understanding the role of emergence in complex systems.

(g) Discussing the role of consciousness in intelligence, and understanding if energy efficient realization of human level intelligence in physical structures also inevitably produces consciousness.

(h) Improved understanding and characterization of information in physical systems, the role of observers and their capabilities when it comes to generating information.

APPENDIX A

TECHNICAL BACKGROUND

A.1 Information Bottleneck

The information bottleneck can be viewed as a rate distortion problem, with a distortion function that measures how well Y is predicted from a compressed representation Z, compared to its direct prediction from X. For the compressed variable Z, the bottleneck is represented as the following constraint optimization problem

$$Minimize_{p(z|x)} I(X;Z) - \beta I(Z;Y)$$

where $I(Z;Y)$ and $I(X;Z)$ the mutual information between Z and Y, and X and Z represent accuracy and complexity respectively. β is the Lagrange trade-off parameter. In order to solve this optimization problem, we construct the Lagrangian $\mathcal{L}_g = I(X;Z) - \beta I(Z;Y) - \sum_{x,z} \lambda(x) p(z|x)$, and differentiate it with respect to $p(z|x)$ and equate it to zero.

$$\frac{d\mathcal{L}}{dp(z|x)} = 0$$

where β is the Langrangian parameter for the information constraint and $\lambda(x)$ is normalization of the conditional distributions at each x. In order to calculate the above equation, we need the following important identities

$$p(y|z) = \sum_x p(y|x) p(x|z)$$

$$p(z) = \sum_x p(z|x) p(x)$$

$$p(z|y) = \sum_x p(z|x)p(x|y)$$

and their derivatives with respect to $p(z|x)$ are given as

$$\frac{\delta p(z)}{\delta p(z|x)} = p(x)$$

$$\frac{\delta p(z|y)}{\delta p(z|x)} = p(x|y)$$

Now starting with the Lagrangian $\mathcal{L}_g$

$$\mathcal{L}_g = I(X;Z) - \beta I(Z;Y) - \sum_{x,z} \lambda(x)p(z|x)$$

$$\begin{aligned}
\mathcal{L}_g &= \sum_{x,z} p(z|x)p(x) log\left[\frac{p(z|x)}{p(x)}\right] - \beta \sum_{z,y} p(z,y) log\left[\frac{p(z|y)}{p(z)}\right] \\
&- \sum_{x,z} \lambda(x)p(z|x)
\end{aligned}$$

Taking the derivatives with respect to $p(z|x)$ for a given x and z, we get

$$\begin{aligned}
\frac{\delta \mathcal{L}_g}{\delta p(z|x)} &= p(x)[1 + \log p(z|x)] - \frac{\delta p(z)}{\delta p(z|x)}[1 + \log p(z|x)] \\
&-\beta \left(\sum_y \frac{\delta p(z|y)}{\delta p(z|x)} p(y)[1 + log p(z|y)] - \frac{\delta p(z)}{\delta p(z|x)}[1 + \log p(z)] \right) \\
&-\lambda(x)
\end{aligned}$$

Substituting the identities from earlier in the section and rearranging the equations

$$\frac{\delta \mathcal{L}_g}{\delta p(z|x)} = p(x) \left(\log \left[\frac{p(z|x)}{p(x)}\right] - \beta \sum_y p(y|x) \log \left[\frac{p(y|z)}{p(y)}\right] - \frac{\lambda(x)}{p(x)} \right)$$

Adding and subtracting $\sum_y p(y|x) \log \left(\frac{p(y|x)}{p(y)}\right)$, and defining $\bar{\lambda}(x)$ to be

$$\bar{\lambda}(x) = \frac{\lambda(x)}{p(x)} - \beta \sum_y p(y|x) \log \left[\frac{p(y|x)}{p(y)} \right]$$

We substitute this into the equation for $\frac{\delta \mathcal{L}_g}{\delta p(z|x)}$ and equating it to 0, we have

$$\begin{aligned} \frac{\delta \mathcal{L}_g}{\delta p(z|x)} &= p(x) \left(\log \left[\frac{p(z|x)}{p(x)} \right] + \beta \sum_y p(y|x) \log \left[\frac{p(y|x)}{p(y|z)} \right] - \bar{\lambda}(x) \right) \\ &= 0 \end{aligned}$$

Solving this equation for $p(z|x)$, we get the solution

$$p(z|x) = \frac{p(z)}{Z(x,\beta)} exp\left(-\beta D_{KL}[p(y|x)|p(y|z)]\right)$$

with

$$Z(x,\beta) = exp[\beta \bar{\lambda}(x)] = \sum_z p(z) exp\left(-\beta D_{KL}[p(y|x)|p(y|z)]\right)$$

where $Z(x,\beta)$ is the normalization partition function.

It is important to note that the Kullback-Leibler divergence, $D_{KL}[p(y|x)|p(y|z)]$, emerged as the relevant "effective distortion measure" from our variational principle and is not assumed otherwise. It is therefore natural to consider it as the "correct" distortion $D(x,z) = D_{KL}[p(y|x)|p(y|z)]$ for quantization in the information bottleneck setting. The following three equations are solved self-consistently to obtain the desired distributions for $p(z)$ and $p(z|x)$

$$p(y|z) = \sum_x p(y|x) p(x|z)$$

$$p(z) = \sum_x p(z|x) p(x)$$

$$p(z|x) = \frac{p(z)}{Z(x,\beta)} exp\left(-\beta D_{KL}[p(y|x)|p(y|z)]\right)$$

A.2 Landauer's Principle - Entropic & Energy Cost of Information Processing

Consider a closed composite system consisting of an "information bearing" subsystem $\mathcal{RS}$ and environment $\mathcal{B}$. Let the states of $\mathcal{R}$ and $\mathcal{S}$ be initially correlated and assume that $\mathcal{RS}$ is initially isolated from $\mathcal{B}$. Initial state of the global system is

$$\hat{\rho} = \hat{\rho}^{\mathcal{RS}} \otimes \hat{\rho}^{\mathcal{B}}$$

The quantum mutual information between $\mathcal{R}$ and $\mathcal{S}$ is given by

$$I(\hat{\rho}^{\mathcal{R}}; \hat{\rho}^{\mathcal{S}}) = \mathcal{I}^{\mathcal{RS}} = S(\hat{\rho}^{\mathcal{R}}) + S(\hat{\rho}^{\mathcal{S}}) - S(\hat{\rho}^{\mathcal{RS}}).$$

The initial total entropy is given by

$$S_{tot}(\hat{\rho}) = k_B \ln(2)[S(\hat{\rho}^{\mathcal{RS}}) + S(\hat{\rho}^{\mathcal{B}})$$

A.2.1 Information Processing

An operation processing information about $\mathcal{R}$ which is encoded in $\mathcal{S}$ is given as an unitary evolution $\hat{\mathcal{U}}$ of $\mathcal{RSB}$ that involves only interactions between $\mathcal{S}$ and $\mathcal{B}$.

$$\hat{\rho}' = \hat{\mathcal{U}}\hat{\rho}\hat{\mathcal{U}}^{\dagger}$$

where $\hat{\mathcal{U}} = \hat{\mathcal{U}}^{\mathcal{R}} \otimes \hat{\mathcal{U}}^{\mathcal{S}}$. The interactions between $\mathcal{S}$ and $\mathcal{B}$ will generally decrease the correlations between $\mathcal{R}$ and $\mathcal{S}$. Thus information about $\mathcal{R}$ is lost in $\mathcal{S}$ during the operation. Final quantum mutual information between $\mathcal{R}$ and $\mathcal{S}$ is

$$I(\hat{\rho}^{\mathcal{R}}; \hat{\rho}^{\mathcal{S}'}) = \mathcal{I}^{\mathcal{RS}'} = S(\hat{\rho}^{\mathcal{R}}) + S(\hat{\rho}^{\mathcal{S}'}) - S(\hat{\rho}^{\mathcal{RS}'}).$$

The final total entropy is

$$S_{tot}(\hat{\rho}') = k_B \ln(2)[S(\hat{\rho}^{\mathcal{RS}'}) + S(\hat{\rho}^{\mathcal{B}'})]$$

A.2.2 Information Loss and Change in Entropy

The change in total entropy during the information processing operation

$$\Delta S = S_{tot}(\hat{\rho}') - S_{tot}(\hat{\rho})$$

The change in quantum mutual information is given by

$$\Delta I = I(\hat{\rho}^{\mathcal{R}}; \hat{\rho}^{\mathcal{S}'}) - I(\hat{\rho}^{\mathcal{R}}; \hat{\rho}^{\mathcal{S}})$$

Thus we can see that

$$\Delta S \geq -k_B \ln(2) \Delta I$$

This gives us the entropic form of Landauer's principle - the entropy increase is lower bounded at $k_B \ln(2)$ per bit of information that is lost during the information processing operation. We will now build upon this entropic bound and obtain the energetic form of Landauer's principle as well.

A.2.3 Information Loss and Energy Flow

In order to study the energy costs of operations that discard information, like irreversible logical operations, it is assumed that the environment is initially a thermal bath at temperature T. Thus the initial state of $\mathcal{B}$ is described by the canonical density operator

$$\hat{\rho}^{\mathcal{B}} = Z^{-1} exp\left(-\frac{\hat{H}_{\mathcal{B}}}{k_B T}\right)$$

where $\hat{H}_{\mathcal{B}}$ is the Hamiltonian of the bath, T is the bath temperature and Z is the partition function given by

$$Z = Tr\left[exp\left(-\frac{\hat{H}_{\mathcal{B}}}{k_B T}\right)\right]$$

The expected energy increase in the environment is given by

$$\Delta\langle E^{\mathcal{B}}\rangle \geq \langle E^{\mathcal{B}'}\rangle - \langle E^{\mathcal{B}}\rangle = Tr[\hat{\rho}^{\mathcal{B}'}\hat{H}_{\mathcal{B}}] - Tr[\hat{\rho}^{\mathcal{B}}\hat{H}_{\mathcal{B}}]$$

Consider the quantity $\Delta\langle E^{\mathcal{B}}\rangle - T\Delta S^{\mathcal{B}}$ where $\Delta S^{\mathcal{B}} = S(\hat{\rho}^{\mathcal{B}'}) - S(\hat{\rho}^{\mathcal{B}})$. Following the derivation in [13], and using

$$\ln \hat{\rho}^{\mathcal{B}} = -\frac{\hat{\rho}^{\mathcal{B}}}{k_B T} - \ln Z$$

we get $\Delta\langle E^{\mathcal{B}}\rangle - T\Delta S^{\mathcal{B}} = k_B T(Tr[\hat{\rho}^{\mathcal{B}'} \ln \hat{\rho}^{\mathcal{B}'}] - Tr[\hat{\rho}^{\mathcal{B}} \ln \hat{\rho}^{\mathcal{B}}])$, which is the relative entropy between initial and final environment states. Since relative entropy is non-negative for any two density operators, we obtain the inequality

$$\Delta\langle E^{\mathcal{B}}\rangle \geq T\Delta S^{\mathcal{B}} \tag{A.1}$$

From the entropic derivation of Landauers Principle, we have that

$$\Delta S = \Delta S^{\mathcal{RS}} + \Delta S^{\mathcal{B}} \geq -k_B \ln(2)\Delta I$$

$$\Delta S^{\mathcal{B}} \geq -k_B \ln(2)[\Delta I + \Delta S^{\mathcal{RS}}]$$

Substituting into Eq.(A.1), we get

$$\Delta\langle E^{\mathcal{B}}\rangle \geq -k_B T \ln(2)[\Delta I + \Delta S^{\mathcal{RS}}] \tag{A.2}$$

Since we know $\Delta I + \Delta S^{\mathcal{RS}} = \Delta S^{\mathcal{S}} = S(\hat{\rho}^{\mathcal{S}'}) - S(\hat{\rho}^{\mathcal{S}})$, this gives us

$$\Delta\langle E^{\mathcal{B}}\rangle \geq -k_B T \ln(2)\Delta S^{\mathcal{S}}$$

APPENDIX B

DISSIPATION BOUNDS IN FSA

B.1 Dissipation Bounds for Deterministic Irreducible FSA in Steady State

Theorem: For physical FSA $\mathcal{F}_P = \left\{\mathcal{S}, \mathcal{R}, \{\hat{\sigma}^{\mathcal{S}}\}, \{\hat{x}^{\mathcal{R}}\}, \{\tilde{\mathcal{L}}\}\right\}$ and input pmf $\{q\}$, the input-averaged amount of energy dissipated to a thermal environment $\mathcal{B}$ on each state transition is lower bounded in steady state as

$$\Delta\langle E^{\mathcal{B}}\rangle \geq k_B T \ln(2) \sum_j q_j \left(\mathcal{I}^{\mathcal{R}_0\mathcal{S}} - \mathcal{I}_j^{\mathcal{R}_0\mathcal{S}'}\right) \tag{B.1}$$

where k_B is the Boltzmann constant, T is the environment temperature, and $\mathcal{I}^{\mathcal{R}_0\mathcal{S}} - \mathcal{I}_j^{\mathcal{R}_0\mathcal{S}'}$ is, for a state transition induced by input $\hat{x}_j^{\mathcal{R}_1}$, the reduction in quantum mutual information between the state of the register system $\mathcal{S}$ and the sequence of all past inputs physically instantiated in referent system $\mathcal{R}_0$. The bound is rigorously derived in [138] and is available in the Appendix here.

Proof: Since von Neumann entropy $S(\hat{\rho})$ is invariant under unitary similarity transformations, and since $\mathcal{R}_1\mathcal{S}\mathcal{B}$ evolves unitarily on each step, we have:

$$S(\hat{\rho}^{\mathcal{R}_1\mathcal{S}\mathcal{B}}) = S(\hat{\rho}^{\mathcal{R}_1\mathcal{S}\mathcal{B}'}). \tag{B.2}$$

From (4.11) and the additivity of von Neumann entropy for separable (tensor product) states, the initial state entropy is

$$S(\hat{\rho}^{\mathcal{R}_1\mathcal{S}\mathcal{B}}) = S(\hat{\rho}^{\mathcal{R}_1}) + S(\hat{\rho}^{\mathcal{S}}) + S(\hat{\rho}^{\mathcal{B}}_{th}).$$

From (4.12) and the joint entropy theorem[1], the entropy of the final state $\hat{\rho}^{\mathcal{R}_1\mathcal{S}\mathcal{B}'} = Tr_{\mathcal{R}_0}[\hat{\rho}^{\mathcal{R}\mathcal{S}\mathcal{B}'}]$ is

$$S(\hat{\rho}^{\mathcal{R}_1\mathcal{S}\mathcal{B}'}) = S(\hat{\rho}^{\mathcal{R}_1}) + \sum_j q_j S(\hat{\rho}^{\mathcal{S}\mathcal{B}'}_j)$$

where, in terms of the steady state occupation probabilities π_k,

$$\hat{\rho}^{\mathcal{S}\mathcal{B}'}_j \equiv \sum_k \pi_k \hat{\rho}^{\mathcal{S}\mathcal{B}'}_{kj}.$$

Substitution of these initial and final entropies into (B.8) yields

$$S(\hat{\rho}^{\mathcal{S}}) + S(\hat{\rho}^{\mathcal{B}}_{th}) = \sum_j q_j S(\hat{\rho}^{\mathcal{S}\mathcal{B}'}_j). \tag{B.3}$$

The sum on the right-hand side is upper bounded as

$$\sum_j q_j S(\hat{\rho}^{\mathcal{S}\mathcal{B}'}_j) \leq \sum_j q_j S(\hat{\rho}^{\mathcal{S}'}_j) + S\left(\sum_j q_j \hat{\rho}^{\mathcal{B}'}_j\right)$$

by the subadditivity and concavity of the von Neumann entropy, so

$$S(\hat{\rho}^{\mathcal{S}}) + S(\hat{\rho}^{\mathcal{B}}_{th}) \leq \sum_j q_j S(\hat{\rho}^{\mathcal{S}'}_j) + S(\hat{\rho}^{\mathcal{B}'})$$

where

$$\hat{\rho}^{\mathcal{B}'} \equiv \sum_j q_j \hat{\rho}^{\mathcal{B}'}_j.$$

---

[1]The joint entropy theorem states that, for a composite system $\mathcal{U}\mathcal{V}$, a set $\left\{|x^{\mathcal{U}}_j\rangle\langle x^{\mathcal{U}}_j|\right\}$ of orthogonal pure states of $\mathcal{U}$, a set $\{\hat{\rho}^{\mathcal{V}}_j\}$ of general density operators on $\mathcal{V}$, and a set of probabilities $\{q_j\}$,

$$S\left(\sum_i p_i |x^{\mathcal{U}}_i\rangle\langle x^{\mathcal{U}}_i| \otimes \hat{\rho}^{\mathcal{V}}_i\right) = S(\hat{\rho}^{\mathcal{V}}) + \sum_i p_i S(\hat{\rho}^{\mathcal{V}}_i)$$

where $S(\hat{\rho}^{\mathcal{V}}) = H(\{q_j\}) = -\sum_j q_j \log_2 q_j$ is the Shannon entropy of $\{q_j\}$. See [**?**] for an extensive discussion of (generally quantum mechanical) entropy and its properties.

Rearranging terms, we obtain the lower bound

$$\Delta S^{\mathcal{B}} \geq S(\hat{\rho}^{\mathcal{S}}) - \sum_j q_j S(\hat{\rho}_j^{\mathcal{S}'}) \tag{B.4}$$

on the entropy change $\Delta S^{\mathcal{B}} = S(\hat{\rho}^{\mathcal{B}'}) - S(\hat{\rho}_{th}^{\mathcal{B}})$ of the bath for the state transition.

Since the states of $\mathcal{R}_1\mathcal{S}$ and $\mathcal{B}$ are initially separable, the dynamical evolution of $\mathcal{R}_1\mathcal{S}\mathcal{B}$ is unitary, and $\mathcal{B}$ is initially in a thermal equilibrium state, the expected increase in the energy of the bath on each step is lower bounded by Partovi's inequality [52] as

$$\Delta\langle E^{\mathcal{B}}\rangle \geq k_B T \ln(2) \Delta S^{\mathcal{B}}. \tag{B.5}$$

With (B.16), this is

$$\Delta\langle E^{\mathcal{B}}\rangle \geq k_B T \ln(2) \left(S(\hat{\rho}^{\mathcal{S}}) - \sum_j q_j S(\hat{\rho}_j^{\mathcal{S}'}) \right). \tag{B.6}$$

In terms of the mutual information quantities

$$I^{\mathcal{R}_0\mathcal{S}} = S(\hat{\rho}^{\mathcal{S}}) - \sum_k \pi_k S(\hat{\sigma}_k^{\mathcal{S}}), \qquad I_j^{\mathcal{R}_0\mathcal{S}'} = S(\hat{\rho}_j^{\mathcal{S}'}) - \sum_k \pi_k S(\hat{\rho}_{kj}^{\mathcal{S}'})$$

which can again be rewritten as

$$\Delta\langle E^{\mathcal{B}}\rangle \geq k_B T \ln(2) \left(I^{\mathcal{R}_0\mathcal{S}} + \sum_k \pi_k S(\hat{\sigma}_k^{\mathcal{S}}) - \sum_j q_j \left[I_j^{\mathcal{R}_0\mathcal{S}'} + \sum_k \pi_k S(\hat{\rho}_{kj}^{\mathcal{S}'}) \right] \right).$$

Noting that

$$\sum_k \pi_k S(\hat{\sigma}_k^{\mathcal{S}}) = \sum_j q_j \sum_k \pi_k S(\hat{\rho}_{kj}^{\mathcal{S}'})$$

in steady state, the above bound simplifies to (B.1) and the theorem is proved.

B.2 Dissipation Bound for a Mealy Machine Over a Cycle

Theorem. For physical FSA $\mathcal{F}_P = \{\mathcal{S}, \mathcal{O}, \mathcal{R}, \{\hat{\sigma}^{\mathcal{S}}\}, \{\hat{\varrho}^{\mathcal{O}}\}, \{\hat{x}^{\mathcal{R}}\}, \{\bar{\mathcal{L}}\}, \{\bar{\mathcal{V}}\}\}$ and input pmf $\{q\}$, the input averaged amount of energy dissipated to a thermal environment $\mathcal{B}$ over one *Mealy machine cycle* is lower bounded in steady state as

$$\Delta\langle E^{\mathcal{B}}\rangle_{cycle} \geq k_B T \ln(2) \left(\sum_j q_j \left[I_j^{\mathcal{R}_0 \mathcal{O}\mathcal{S}} - I_j^{\mathcal{R}_0 \mathcal{O}\mathcal{S}'} \right] \right.$$
$$\left. + \mathcal{I}^{\mathcal{R}_1 \mathcal{O}\mathcal{S}'} + \sum_j q_j S(\hat{\rho}_j^{\mathcal{O}\mathcal{S}'}) - \sum_{j'} q_{j'} S(\hat{\rho}_{j'}^{\mathcal{O}\mathcal{S}''}) \right) \tag{B.7}$$

where k_B is the Boltzmann constant and T is the environment temperature. $I_j^{\mathcal{R}_0 \mathcal{O}\mathcal{S}} - I_j^{\mathcal{R}_0 \mathcal{O}\mathcal{S}'}$ is, for a state transition induced by input $\hat{x}_j^{\mathcal{R}_1}$, the reduction in quantum mutual information between the joint state of the output and automata register $\mathcal{O}\mathcal{S}$ and the sequence of all past inputs physically instantiated in referent system $\mathcal{R}_0$. $\mathcal{I}^{\mathcal{R}_1 \mathcal{O}\mathcal{S}'}$ is amount of quantum mutual information between the joint state $\mathcal{O}\mathcal{S}$ at the intermediate state and the referent $\mathcal{R}_1$. $S(\hat{\rho}_j^{\mathcal{O}\mathcal{S}'})$ and $S(\hat{\rho}_{j'}^{\mathcal{O}\mathcal{S}''})$ are the self entropies of the states of $\mathcal{O}\mathcal{S}$ associated with the j-th input of $\mathcal{R}_1$ and the j'-th input of $\mathcal{R}_2$, at the intermdeiate and final stages respectively.

Proof: Since von Neumann entropy $S(\hat{\rho})$ is invariant under unitary similarity transformations, and $\mathcal{R}_1 \mathcal{O}\mathcal{S}\mathcal{B}$ evolves unitarily at the start of the *cycle*, we have:

$$S(\hat{\rho}^{\mathcal{R}_1 \mathcal{O}\mathcal{S}\mathcal{B}}) = S(\hat{\rho}^{\mathcal{R}_1 \mathcal{O}\mathcal{S}\mathcal{B}'}). \tag{B.8}$$

Using the joint entropy theorem, we have the initial and final state entropies as

$$S(\hat{\rho}^{\mathcal{R}_1 \mathcal{S}\mathcal{B}}) = S(\hat{\rho}^{\mathcal{R}_1}) + \sum_j q_j S(\hat{\rho}_j^{\mathcal{O}\mathcal{S}}) + S(\hat{\rho}_{th}^{\mathcal{B}}),$$

$$S(\hat{\rho}^{\mathcal{R}_1 \mathcal{S}\mathcal{B}'}) = S(\hat{\rho}^{\mathcal{R}_1}) + \sum_j q_j S(\hat{\rho}_j^{\mathcal{O}\mathcal{S}\mathcal{B}'}).$$

Substituting for initial and final entropies into (B.8), and from subadditivity and concavity of von Neumann entropy, we have

$$\sum_j q_j S(\hat{\rho}_j^{\mathcal{O}\mathcal{S}}) + S(\hat{\rho}_{th}^{\mathcal{B}}) \leq \sum_j q_j S(\hat{\rho}_j^{\mathcal{S}'}) + S(\hat{\rho}^{\mathcal{B}'}).$$

Rearranging terms, we obtain the lower bound on the on the entropy change of the bath $\Delta S^{\mathcal{B}}$ for the state transition

$$\Delta S^{\mathcal{B}} = S(\hat{\rho}^{\mathcal{B}'}) - S(\hat{\rho}^{\mathcal{B}}_{th}) \geq \sum_j q_j \left[S(\hat{\rho}_j^{\mathcal{OS}}) - S(\hat{\rho}_j^{\mathcal{OS}'}) \right]. \tag{B.9}$$

The expected increase in the energy of the bath on each step is lower bounded by Partovi's inequality [52] as $\Delta\langle E^{\mathcal{B}}\rangle \geq k_B T \ln(2) \Delta S^{\mathcal{B}}$. With (B.16), this is

$$\Delta\langle E^{\mathcal{B}}\rangle \geq k_B T \ln(2) \sum_j q_j \left[S(\hat{\rho}_j^{\mathcal{OS}}) - S(\hat{\rho}_j^{\mathcal{OS}'}) \right]. \tag{B.10}$$

For the case of steady state in the FSA, the bound can be written in terms of mutual information quantities as

$$\Delta\langle E^{\mathcal{B}}\rangle_1 \geq k_B T \ln(2) \sum_j q_j \left[I_j^{\mathcal{R}_0\mathcal{OS}} - I_j^{\mathcal{R}_0\mathcal{OS}'} \right]. \tag{B.11}$$

The lower bound on the energy dissipation associated with the unitary evolution of the system $\mathcal{R}_2\mathcal{OSB}$ to produce the final states, is calculated using a similar set of steps.

$$\begin{aligned} \Delta\langle E^{\mathcal{B}}\rangle_2 \quad \geq \quad & k_B T \ln(2) \left[\mathcal{I}^{\mathcal{R}_1\mathcal{OS}'} + \sum_j q_j S(\hat{\rho}_j^{\mathcal{OS}'}) \right. \\ & \left. - \sum_{j'} q_{j'} S(\hat{\rho}_{j'}^{\mathcal{OS}''}) \right] \end{aligned} \tag{B.12}$$

The total dissipation over one *cycle*, $\Delta\langle E^{\mathcal{B}}\rangle_{cycle} = \Delta\langle E^{\mathcal{B}}\rangle_1 + \Delta\langle E^{\mathcal{B}}\rangle_2$. Adding (B.11) and (B.12) gives us the bound in (B.7).

B.3 Dissipation Bounds for Probabilistic FSAs

Theorem: For a physical probabilistic FSA $\mathcal{F}_P = \left\{\mathcal{S}, \mathcal{R}, \{\hat{\sigma}^{\mathcal{S}}\}, \{\hat{x}^{\mathcal{R}}\}, \{\tilde{\mathcal{L}}\}\right\}$ and input pmf $\{q\}$, the input-averaged amount of energy dissipated to a thermal environment $\mathcal{B}$ on each state transition is lower bounded in steady state as

$$\Delta\langle E^{\mathcal{B}}\rangle \geq k_B T \ln(2) \left[\sum_j q_j \left(\mathcal{I}^{\mathcal{R}_0\mathcal{S}} - \mathcal{I}_j^{\mathcal{R}_0\mathcal{S}'}\right) + \sum_i p_i (H(\{\pi_{k|i}^{(n-1)}\}) - \sum_{(j)} q_j H(\{\pi_{k'|(i,j)}^{(n)}\}))\right] \tag{B.13}$$

where k_B is the Boltzmann constant, T is the environment temperature, and $\mathcal{I}^{\mathcal{R}_0\mathcal{S}} - \mathcal{I}_j^{\mathcal{R}_0\mathcal{S}'}$ is, for a state transition induced by input $\hat{x}_j^{\mathcal{R}_1}$, the reduction in quantum mutual information between the state of the register system $\mathcal{S}$ and the sequence of all past inputs physically instantiated in referent system $\mathcal{R}_0$. $H(\{\pi_{k|i}^{(n-1)}\})$ is the Shannon entropy of $\{\pi_{k|i}^{(n-1)}\}$, the probability that the i-th input maps to the k-th state of the FSA before the $(n-1)$-th transition. $H(\{\pi_{k|(i,j)}^{(n)}\})$ is Shannon entropy of the distribution $\{\pi_{k|(i,j)}^{(n)}\}$, the probability that the (i,j)-th inputs maps to the k' state after the state transition.

Proof: Since von Neumann entropy $S(\hat{\rho})$ is invariant under unitary similarity transformations, and since $\mathcal{R}_1\mathcal{S}\mathcal{B}$ evolves unitarily on each step, we have:

$$S(\hat{\rho}^{\mathcal{R}_1\mathcal{S}\mathcal{B}}) = S(\hat{\rho}^{\mathcal{R}_1\mathcal{S}\mathcal{B}'}). \tag{B.14}$$

From (4.11) and the additivity of von Neumann entropy for separable (tensor product) states, the initial state entropy is

$$S(\hat{\rho}^{\mathcal{R}_1\mathcal{S}\mathcal{B}}) = S(\hat{\rho}^{\mathcal{R}_1}) + S(\hat{\rho}^{\mathcal{S}}) + S(\hat{\rho}_{th}^{\mathcal{B}}).$$

From (4.12) and the joint entropy theorem, the entropy of the final state $\hat{\rho}^{\mathcal{R}_1\mathcal{S}\mathcal{B}'} = Tr_{\mathcal{R}_0}[\hat{\rho}^{\mathcal{R}\mathcal{S}\mathcal{B}'}]$ is

$$S(\hat{\rho}^{\mathcal{R}_1\mathcal{S}\mathcal{B}'}) = S(\hat{\rho}^{\mathcal{R}_1}) + \sum_j q_j S(\hat{\rho}_j^{\mathcal{S}\mathcal{B}'})$$

where,

$$\hat{\rho}_j^{\mathcal{SB}'} \equiv \sum_i p_i \hat{\rho}_{ij}^{\mathcal{SB}'}.$$

Substitution of these initial and final entropies into (B.8) yields

$$S(\hat{\rho}^{\mathcal{S}}) + S(\hat{\rho}_{th}^{\mathcal{B}}) = \sum_j q_j S(\hat{\rho}_j^{\mathcal{SB}'}). \tag{B.15}$$

The sum on the right-hand side is upper bounded as

$$\sum_j q_j S(\hat{\rho}_j^{\mathcal{SB}'}) \leq \sum_j q_j S(\hat{\rho}_j^{\mathcal{S}'}) + S\left(\sum_j q_j \hat{\rho}_j^{\mathcal{B}'}\right)$$

by the subadditivity and concavity of the von Neumann entropy, so

$$S(\hat{\rho}^{\mathcal{S}}) + S(\hat{\rho}_{th}^{\mathcal{B}}) \leq \sum_j q_j S(\hat{\rho}_j^{\mathcal{S}'}) + S(\hat{\rho}^{\mathcal{B}'})$$

where

$$\hat{\rho}^{\mathcal{B}'} \equiv \sum_j q_j \hat{\rho}_j^{\mathcal{B}'}.$$

Rearranging terms, we obtain the lower bound

$$\Delta S^{\mathcal{B}} \geq S(\hat{\rho}^{\mathcal{S}}) - \sum_j q_j S(\hat{\rho}_j^{\mathcal{S}'}) \tag{B.16}$$

on the entropy change $\Delta S^{\mathcal{B}} = S(\hat{\rho}^{\mathcal{B}'}) - S(\hat{\rho}_{th}^{\mathcal{B}})$ of the bath for the state transition.

Since the states of $\mathcal{R}_1\mathcal{S}$ and $\mathcal{B}$ are initially separable, the dynamical evolution of $\mathcal{R}_1\mathcal{SB}$ is unitary, and $\mathcal{B}$ is initially in a thermal equilibrium state, the expected increase in the energy of the bath on each step is lower bounded by Partovi's inequality [52] as

$$\Delta\langle E^{\mathcal{B}}\rangle \geq k_B T \ln(2) \Delta S^{\mathcal{B}}. \tag{B.17}$$

With (B.16), this is

$$\Delta\langle E^{\mathcal{B}}\rangle \geq k_B T \ln(2) \left(S(\hat{\rho}^{\mathcal{S}}) - \sum_j q_j S(\hat{\rho}_j^{\mathcal{S}'})\right). \tag{B.18}$$

In terms of the mutual information quantities

$$I^{\mathcal{R}_0\mathcal{S}} = S(\hat{\rho}^{\mathcal{S}}) - \sum_i p_i S(\hat{\rho}_i^{\mathcal{S}}), \qquad I_j^{\mathcal{R}_0\mathcal{S}'} = S(\hat{\rho}_j^{\mathcal{S}'}) - \sum_i p_i S(\hat{\rho}_{ij}^{\mathcal{S}'})$$

this can again be rewritten as

$$\Delta\langle E^{\mathcal{B}}\rangle \geq k_B T \ln(2) \left(I^{\mathcal{R}_0\mathcal{S}} + \sum_i p_i S(\hat{\rho}_i^{\mathcal{S}}) - \sum_j q_j \left[I_j^{\mathcal{R}_0\mathcal{S}'} + \sum_i p_i S(\hat{\rho}_{ij}^{\mathcal{S}'}) \right] \right). \quad \text{(B.19)}$$

Noting that since

$$\hat{\rho}_i^{\mathcal{S}} = \sum_k \pi_{k|i}^{(n-1)} \hat{\sigma}_k^{\mathcal{S}}, \qquad \hat{\rho}_{ij}^{\mathcal{S}} = \sum_{k'} \pi_{k|(i,j)}^{(n)} \hat{\sigma}_{k'}^{\mathcal{S}'}$$

where $\pi_{k|i}^{(n-1)}$ and $\pi_{k|(i,j)}^{(n)}$ are the probabilities that the initial state of the FSA maps onto the k-th and k'-th distinguishable state for the i-th and (i,j)-th inputs before the $(n-1)$ and n-th transitions respectively. We thus get

$$S(\hat{\rho}_i^{\mathcal{S}}) = H(\{\pi_{k|i}^{(n-1)}\}) + \sum_k \pi_{k|i}^{(n-1)} S(\hat{\sigma}_k^{\mathcal{S}})$$

$$S(\hat{\rho}_{ij}^{\mathcal{S}'}) = H(\{\pi_{k'|(i,j)}^{(n)}\}) + \sum_{k'} \pi_{k'|(i,j)}^{(n)} S(\hat{\sigma}_{k'}^{\mathcal{S}'})$$

In steady state we have that $\pi^{(n-1)} = \pi^{(n)}$ and

$$\sum_k \pi^{(n-1)} S(\hat{\sigma}_k^{\mathcal{S}}) = \sum_{k'} \pi^{(n)} S(\hat{\sigma}_{k'}^{\mathcal{S}'})$$

This can be rewritten as

$$\sum_i p_i \sum_k \pi_{k|i}^{(n-1)} S(\hat{\sigma}_k^{\mathcal{S}}) = \sum_{(i,j)} p_i q_j \sum_{k'} \pi_{k|(i,j)}^{(n)} S(\hat{\sigma}_{k'}^{\mathcal{S}'})$$

Thus we get

$$\sum_i p_i \left[S(\hat{\rho}_i^{\mathcal{S}}) - \sum_j q_j S(\hat{\rho}_{ij}^{\mathcal{S}'}) \right] = \sum_i p_i \left[H(\{\pi_{k|i}^{(n-1)}\}) - \sum_{(j)} q_j H(\{\pi_{k'|(i,j)}^{(n)}\}) \right]$$

Substituting the above equality into (B.19) to give us the lower bound on dissipation for a probabilistic FSA in steady state.

$$\Delta\langle E^{\mathcal{B}}\rangle \geq k_B T \ln(2) \left[\sum_j q_j \left(\mathcal{I}^{\mathcal{R}_0\mathcal{S}} - \mathcal{I}_j^{\mathcal{R}_0\mathcal{S}'} \right) + \sum_i p_i (H(\{\pi_{k|i}^{(n-1)}\}) - \sum_{(j)} q_j H(\{\pi_{k'|(i,j)}^{(n)}\})) \right]$$

B.4 Dissipation in FSA with Correlated Inputs

Theorem: For physical FSA $\mathcal{F}_P = \left\{\mathcal{S}, \mathcal{R}, \{\hat{\rho}^{\mathcal{S}}\}, \{\hat{x}^{\mathcal{R}}\}, \{\tilde{\mathcal{L}}\}\right\}$ and input pmf $\{q\}$, the input-averaged amount of energy dissipated to a thermal environment $\mathcal{B}$ on each state transition is lower bounded as

$$\Delta\langle E^{\mathcal{B}}\rangle \geq k_B T \ln(2) \left(\sum_j q_j [S(\hat{\rho}_j^{\mathcal{S}}) - S(\hat{\rho}_j^{\mathcal{S}'})] \right) \tag{B.20}$$

where k_B is the Boltzmann constant and T is the environment temperature. For a state transition induced by input $\hat{x}_j^{\mathcal{R}_1}$, $\hat{\rho}_j^{\mathcal{S}}$ and $\hat{\rho}_j^{\mathcal{S}'}$ are the density operators associated with the j-th input before and after the state transition. This can be rewritten in information theoretic terms as

$$\Delta\langle E^{\mathcal{B}}\rangle \geq k_B T \ln(2) \left[-\Delta S^{\mathcal{S}} + \Delta \mathcal{I}^{\mathcal{R}_1 \mathcal{S}} \right] \tag{B.21}$$

where $-\Delta S^{\mathcal{S}}$ is the reduction in von Neumann entropy of the system $\mathcal{S}$ over the transition, and $\Delta \mathcal{I}^{\mathcal{R}_1 \mathcal{S}} = \mathcal{I}^{\mathcal{R}_1 \mathcal{S}'} - \mathcal{I}^{\mathcal{R}_1 \mathcal{S}}$ is the change in quantum mutual information between the system $\mathcal{S}$ and the latest input $\mathcal{R}_1$. The quantum mutual information between $\mathcal{S}$ and $\mathcal{R}_1$ before the state transition $\mathcal{I}^{\mathcal{R}_1 \mathcal{S}}$, can be seen as a measure of prediction of the next input $\mathcal{R}_1$ by the system $\mathcal{S}$. In the next few sections, we will describe a simple machine trying to learn the external input pixel values and calculate the lower bound on dissipation for both temporally correlated and independent inputs.

Proof: The statistical state of the composite $\mathcal{RSB}$ is given by the density operator

$$\hat{\rho}^{\mathcal{RSB}} = \hat{\rho}^{\mathcal{R}_0 \mathcal{R}_1 \mathcal{S}} \otimes \hat{\rho}_{th}^{\mathcal{B}}$$

where $\hat{\rho}^{\mathcal{R}_0 \mathcal{R}_1 \mathcal{S}} = \sum_i p_i \left(\hat{\rho}_i^{\mathcal{R}_0} \otimes \hat{\rho}_i^{\mathcal{S}} \otimes \sum_j q_{j|i} \hat{\rho}_j^{\mathcal{R}_1} \right)$. $\{q_{j|i}\}$ is the conditional probability distribution of the j-th input of $\mathcal{R}_1$, given the i-th input string of $\mathcal{R}_0$.

Since von Neumann entropy $S(\hat{\rho})$ is invariant under unitary similarity transformations and $\mathcal{R}_1\mathcal{SB}$ evolves unitarily on each step, we have:

$$S(\hat{\rho}^{\mathcal{R}_1\mathcal{SB}}) = S(\hat{\rho}^{\mathcal{R}_1\mathcal{SB}'}). \tag{B.22}$$

And using the joint entropy theorem we have the initial entropy of $\hat{\rho}^{\mathcal{R}_1\mathcal{SB}}$ to be

$$S(\hat{\rho}^{\mathcal{R}_1\mathcal{SB}}) = S(\hat{\rho}^{\mathcal{R}_1}) + S(\hat{\rho}_j^{\mathcal{S}}) + S(\hat{\rho}_{th}^{\mathcal{B}}).$$

Using the entropy theorem again, the entropy of the final state $\hat{\rho}^{\mathcal{R}_1\mathcal{SB}'} = Tr_{\mathcal{R}_0}[\hat{\rho}^{\mathcal{RSB}'}]$ is

$$S(\hat{\rho}^{\mathcal{R}_1\mathcal{SB}'}) = S(\hat{\rho}^{\mathcal{R}_1}) + \sum_j q_j S(\hat{\rho}_j^{\mathcal{SB}'})$$

Substitution of these initial and final entropies into Eq.(B.22) yields

$$\sum_j q_j S(\hat{\rho}_j^{\mathcal{S}}) + S(\hat{\rho}_{th}^{\mathcal{B}}) = \sum_j q_j S(\hat{\rho}_j^{\mathcal{SB}'}). \tag{B.23}$$

The sum on the right-hand side is upper bounded as

$$\sum_j q_j S(\hat{\rho}_j^{\mathcal{SB}'}) \leq \sum_j q_j S(\hat{\rho}_j^{\mathcal{S}'}) + S\left(\sum_j q_j \hat{\rho}_j^{\mathcal{B}'}\right)$$

Using the subadditivity and concavity of the von Neumann entropy, so

$$S(\hat{\rho}^{\mathcal{S}}) + S(\hat{\rho}_{th}^{\mathcal{B}}) \leq \sum_j q_j S(\hat{\rho}_j^{\mathcal{S}'}) + S(\hat{\rho}^{\mathcal{B}'})$$

where we have

$$\hat{\rho}^{\mathcal{B}'} \equiv \sum_j q_j \hat{\rho}_j^{\mathcal{B}'}.$$

Rearranging terms, we obtain the lower bound

$$\Delta S^{\mathcal{B}} \geq \sum_j q_j \left(S(\hat{\rho}_j^{\mathcal{S}}) - S(\hat{\rho}_j^{\mathcal{S}'}) \right) \tag{B.24}$$

on the entropy change $\Delta S^{\mathcal{B}} = S(\hat{\rho}^{\mathcal{B}'}) - S(\hat{\rho}_{th}^{\mathcal{B}})$ of the bath for the state transition.

The states of $\mathcal{R}_1\mathcal{S}$ and $\mathcal{B}$ are initially separable, the dynamical evolution of $\mathcal{R}_1\mathcal{S}\mathcal{B}$ is unitary, and $\mathcal{B}$ is initially in a thermal equilibrium state. Thus the expected increase in the energy of the bath on each step is lower bounded by Partovi's inequality [52] as

$$\Delta\langle E^{\mathcal{B}} \rangle \geq k_B T \ln(2) \Delta S^{\mathcal{B}}. \tag{B.25}$$

With Eq.(B.24), this is

$$\Delta\langle E^{\mathcal{B}} \rangle \geq k_B T \ln(2) \left(\sum_j q_j [S(\hat{\rho}_j^{\mathcal{S}}) - S(\hat{\rho}_j^{\mathcal{S}'})] \right). \tag{B.26}$$

In terms of the mutual information quantities

$$I^{\mathcal{R}_1\mathcal{S}} = S(\hat{\rho}^{\mathcal{S}}) - \sum_j q_j S(\hat{\rho}_j^{\mathcal{S}}), \qquad I^{\mathcal{R}_1\mathcal{S}'} = S(\hat{\rho}^{\mathcal{S}'}) - \sum_j q_j S(\hat{\rho}_j^{\mathcal{S}'})$$

where $I^{\mathcal{R}_1\mathcal{S}}$ and $I^{\mathcal{R}_1\mathcal{S}'}$ is the quantum mutual infomation between the system $\mathcal{S}$ and incoming input $\mathcal{R}_1$ before and after the state distribution himself. By substituting Eq.(B.24), this can again be rewritten as

$$\Delta\langle E^{\mathcal{B}} \rangle \geq k_B T \ln(2) \left[-\Delta S^{\mathcal{S}} + \Delta I^{\mathcal{R}_1\mathcal{S}} \right].$$

where $\Delta I^{\mathcal{R}_1\mathcal{S}} = I^{\mathcal{R}_1\mathcal{S}'} - I^{\mathcal{R}_1\mathcal{S}}$ as is the change in the correlations between $\mathcal{R}_1$ and $\mathcal{S}$. Of course in steady state, we have $\Delta S^{\mathcal{S}} = 0$ and the bound reduces to

$$\Delta\langle E^{\mathcal{B}} \rangle \geq k_B T \ln(2) \Delta I^{\mathcal{R}_1\mathcal{S}}.$$

The lower bounds on dissipation for a FSA with correlated inputs have been derived in this section. It is clear from the lower bound that $I^{\mathcal{R}_1\mathcal{S}}$ is similar to a predictive component in the dissipation bound. This is extremely relevant especially given the increase in interesting of learning topics. In the next section, we will introduce a simple learning machine that learns the correlated data inputs, and calculate the lower bound on it's dissipation.

BIBLIOGRAPHY

[1] Moore, Gordon E., "Cramming more components onto integrated circuits.", (1965): 114-117.

[2] Teuscher, Christof, "The weird, the small, and the uncontrollable: Redefining the frontiers of computing," *Computer*, 50.8 (2017): 52-58.

[3] Hill, Mark D., and Michael R. Marty, "Amdahl's law in the multicore era", *Computer*, 41.7, 33-38, 2008.

[4] Fuchs, Adi, and David Wentzlaff, "The Accelerator Wall: Limits of Chip Specialization", *2019 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, IEEE, 2019.

[5] Dennard, Robert H., et al., "Design of ion-implanted MOSFET's with very small physical dimensions.", *IEEE Journal of Solid-State Circuits*, 9.5 (1974): 256-268.

[6] Keyes, Robert W, "Physical limits of silicon transistors and circuits", *Reports on Progress in Physics*, 68.12 (2005): 2701.

[7] Ganesh, Natesh, "Thermodynamic Intelligence, a Heretical Theory," *2018 IEEE International Conference on Rebooting Computing (ICRC)*, McLean, VA, USA, 2018, pp. 1-10.

[8] OpenAI, "AI and Compute", *OpenAI blog*, https://openai.com/blog/ai-and-compute/, 2018.

[9] Anderson, Neal G., "ECE697 PT Physical Information Theory," UMass Spring-2014.

[10] Cover, Thomas M., and Thomas, Joy A., "Elements of information theory", *John Wiley & Sons*, 2012.

[11] Shannon, Claude Elwood, "A mathematical theory of communication.", *Bell System Technical Journal*, 27.3 (1948): 379-423.

[12] Tishby, Naftali, Fernando C. Pereira, and William Bialek, "The Information Bottleneck Method," *arXiv preprint physics*/0004057 (2000).

[13] Blahut, Richard, "Computation of channel capacity and rate-distortion functions.", *IEEE Transactions on Information Theory* , 18.4 (1972): 460-473.

[14] Winograd, Shmuel, and Cowan, Jack D., "Reliable computation in the presence of noise.", No. 22. *Cambridge, Mass.: Mit Press*, 1963.

[15] Kondepudi, Dilip, and Ilya Prigogine. Modern thermodynamics: from heat engines to dissipative structures. John Wiley & Sons, 2014.

[16] Schement, Jorge R. and Ruben, Brent D, "Information Theory and Communication", *Transaction Publishers*, Volume 4, pp 43,53, 1993.

[17] Kondepudi, Dilip, and Ilya Prigogine,*Modern thermodynamics: from heat engines to dissipative structures*, (2014).

[18] Landauer, Rolf, "Irreversibility and heat generation in the computing process", *IBM Journ. of Res. and Dev.*, 5.3 (1961): 183-191.

[19] Bennett, Charles H, "Notes on Landauer's principle, reversible computation, and Maxwell's Demon." *Studies In History and Philosophy of Science Part B: Studies In History and Philosophy of Modern Physics*, 34.3 (2003): 501-510.

[20] Lent, C.S., Orlov, A.O., Porod, W., Snider G., "Energy Limits in Computation: A Review of Landauers Principle, Theory and Experiments", *Springer International Publishing*, 2019.

[21] Anderson, Neal G., "ECE697 PT Physical Information Theory," UMass Spring-2014.

[22] Anderson, Neal G., "On the physical implementation of logical transformations: Generalized L-machines," *Theoretical Computer Science*, 411.48 (2010): 4179-4199.

[23] Anderson, Neal G., "Overwriting information: Correlations, physical costs, and environment models", *Physics Letters A*, 376.17 (2012): 1426-1433.

[24] Anderson, Neal G., "Information erasure in quantum systems," *Physics Letters A*, 372.34 (2008): 5552-5555.

[25] Anderson, Neal G., "Irreversible information loss: fundamental notions and entropy costs", *International Journal of Modern Physics: Conference Series*, Vol. 33. 2014.

[26] Anderson, Neal G., "Conditioning, correlation and entropy generation in Maxwells Demon.", *Entropy*, 15.10 (2013): 4243-4265.

[27] Anderson, Neal G, "Information as a physical quantity", *Information Sciences*, 415, 397-413, 2017.

[28] Nielsen, Michael A., and Isaac L. Chuang, "Quantum computation and quantum information," *Cambridge University Press*, (2000).

[29] Wehrl, Alfred, "General properties of entropy," *Reviews of Modern Physics*, 50.2 (1978): 221.

[30] Nielsen, Michael A., and Isaac L. Chuang, "Quantum computation and quantum information." *Phys. Today*, 54 (2001): 60-2.

[31] Holevo, Alexander Semenovich, "Bounds for the quantity of information transmitted by a quantum communication channel.", *Problemy Peredachi Informatsii*, 9.3 (1973): 3-11.

[32] Schumacher, Benjamin, Michael Westmoreland, and William K. Wootters, "Limitation on the amount of accessible information in a quantum channel.", *Physical review letters*, 76.18 (1996): 3452.

[33] Ladyman, James, et al., "The connection between logical and thermodynamic irreversibility," *Studies In History and Philosophy of Science Part B: Studies In History and Philosophy of Modern Physics*, 38.1 (2007): 58-79.

[34] Ladyman, James, "What does it mean to say that a physical system implements a computation?," *Theoretical Computer Science*, 410.4-5 (2009): 376-383.

[35] Landauer, Rolf, "Irreversibility and heat generation in the computing process", *IBM Journ. of Res. and Dev.*, 5.3 (1961): 183-191.

[36] Landauer, Rolf, "Irreversibility and heat generation in the computing process", *IBM Journ. of Res. and Dev.*, 5.3 (1961): 183-191.

[37] Anderson, Neal G, "On the physical implementation of logical transformations: Generalized L-machines.", *Theoretical Computer Science*, 411.48 (2010): 4179-4199.

[38] Anderson, Neal G., "Overwriting information: Correlations, physical costs, and environment models", *Physics Letters A*, 376.17 (2012): 1426-1433.

[39] Minsky, Marvin Lee, "Computation." *Englewood Cliffs: Prentice-Hall*, 1967.

[40] Hennie, Frederick C, "Finite-state models for logical machines.", *Wiley New York*, 1968.

[41] Matherat, Philippe, and Marc-Thierry Jaekel, "Logical Dissipation of Automata Implements-Dissipation of Computation.", *arXiv preprint quant-ph/9805018*, 1998.

[42] Wiesner, Karoline, et al., "Information-theoretic lower bound on energy cost of stochastic computation.", *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 468.2148 (2012): 4058-4066.

[43] Ganesh, Natesh & Neal G. Anderson, "Irreversibility and dissipation in finite-state automata", *Physics Letters A* 377.45 (2013): 3266-3271.

[44] DeBenedictis, Erik P., et al., "A path toward ultra-low-energy computing", *2016 IEEE International Conference on Rebooting Computing (ICRC)*, IEEE, 2016.

[45] Anderson, Neal G., Ilke Ercan, and Natesh Ganesh, "Toward nanoprocessor thermodynamics," *IEEE Transactions on Nanotechnology*, 12.6 (2013): 902-909.

[46] Pin, Jean-Eric, "On reversible automata.', *Latin American Symposium on Theoretical Informatics*, Springer, Berlin, Heidelberg, 1992.

[47] Anderson, Neal G., "Information erasure in quantum systems.", *Physics Letters A*, 372.34 (2008): 5552-5555.

[48] Ladyman, James, "What does it mean to say that a physical system implements a computation?," *Theoretical Computer Science*, 410.4-5 (2009): 376-383.

[49] Ercan, Ilke, and Neal G. Anderson. "Heat dissipation in nanocomputing: lower bounds from physical information theory," *IEEE Transactions on Nanotechnology*, 12.6 (2013): 1047-1060.

[50] Ladyman, James, et al., "The connection between logical and thermodynamic irreversibility," *Studies In History and Philosophy of Science Part B: Studies In History and Philosophy of Modern Physics*, 38.1 (2007): 58-79.

[51] Wehrl, Alfred, "General properties of entropy," *Reviews of Modern Physics*, 50.2 (1978): 221.

[52] Partovi, M. Hossein, "Quantum thermodynamics," *Physics Letters A*, 137.9 (1989): 440-444.

[53] Williams, Stanley & DeBenedictis, Erik P.,"OSTP NanotechnologyInspired Grand Challenge: Sensible Machines," *IEEE Rebooting Computing*, 2015.

[54] Suri, M., Bichler, O., Querlioz, D., Traor, B., Cueto, O., Perniola, L., Sousa, V., Vuillaume, D., Gamrat, C. & DeSalvo, B., "Physical aspects of low power synapses based on phase change memory devices", *Journal of Applied Physics*, 112(5), p.054904.

[55] Roy, K., Fan, D., Fong, X., Kim, Y., Sharad, M., Paul, S., Chatterjee, S., Bhunia, S. & Mukhopadhyay, S., "Exploring spin transfer torque devices for unconventional computing", *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 5(1), pp.5-16.

[56] Indiveri, G., Linares-Barranco, B., Legenstein, R., Deligeorgis, G. & Prodromakis, T., "Integration of nanoscale memristor synapses in neuromorphic computing architectures", *Nanotechnology*, 24(38), p.384010.

[57] George J. K., Nejadriahi H., & Sorger V. J., "Towards On-Chip Optical FFTs for Convolutional Neural Networks," *2017 IEEE International Conference on Rebooting Computing (ICRC)*, Washington, DC, USA, 2017, pp. 1-4.

[58] Rojas, Raul,"Neural Networks: a systematic introductio", 2009.

[59] Jain, Anil K., Jianchang Mao, and K. M. Mohiuddin, "Artificial neural networks: A tutorial.", *Computer*, 3 (1996): 31-44.

[60] Landauer, Rolf, "Irreversibility and heat generation in the computing process", *IBM Journ. of Res. and Dev.*, 5.3 (1961): 183-191.

[61] Theis, Thomas N & Wong Phillip H. S., "The End of Moore's Law: A New Beginning for Information Technology", *IEEE Computing in Science and Engineering*, Vol. 19, Issue 2, Mar - Apr 2017.

[62] Ganesh, Natesh & Neal G. Anderson, "Irreversibility and dissipation in finite-state automata", *Physics Letters A* 377.45 (2013): 3266-3271.

[63] Jo, S.H., Chang, T., Ebong, I., Bhadviya, B.B., Mazumder, P. & Lu, W., "Nanoscale memristor device as synapse in neuromorphic systems", *Nano letters*, 10(4), pp.1297-1301.

[64] Ganesh, Natesh, and Neal G. Anderson, "Dissipation in neuromorphic computing: Fundamental bounds for feedforward networks", *2017 IEEE 17th International Conference on Nanotechnology (IEEE-NANO)*, IEEE, 2017.

[65] Zeiler, Matthew D., "ADADELTA: an adaptive learning rate method", *arXiv preprint*, arXiv:1212.5701 (2012).

[66] Tishby, Naftali & Noga Zaslavsky, "Deep learning and the information bottleneck principle", *IEEE Information Theory Workshop (ITW)*, 2015.

[67] Slonim, Noam & Naftali Tishby, "Document clustering using word clusters via the information bottleneck method", *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2000.

[68] Ganesh, Natesh, "A Thermodynamic Treatment of Intelligent Systems," *2017 IEEE International Conference on Rebooting Computing (ICRC)*, Washington, DC, USA, 2017, pp. 1-4.

[69] Storkey, Amos, & Romain Valabregue, "Hopfield learning rule with high capacity storage of time-correlated patterns.", *IEEE Electronics Letters*, 33.21 (1997): 1803-1804.

[70] Hertz, John A, "Introduction to the theory of neural computation.", *CRC Press*, 2018.

[71] Street, Sterling, "Neurobiology as information physics," *Frontiers in Systems Neuroscience*, 10 (2016).

[72] Collell, Guillem, and Jordi Fauquet, "Brain activity and cognition: a connection from thermodynamics and information theory," *Frontiers in Psychology*, 6 (2015).

[73] Halley, Julianne, and David A. Winkler., "Consistent concepts of selforganization and selfassembly," *Complexity*, 14.2 (2008): 10-17.

[74] Fry, Robert L., "Physical Intelligence and Thermodynamic Computing," *Entropy*, 19.3 (2017): 107.

[75] Hopfield, John J., "Neural networks and physical systems with emergent collective computational abilities," *Proceedings of the National Academy of Sciences*, 79.8 (1982): 2554-2558.

[76] Pathria, R K., "Statistical Mechanics," *Oxford: Butterworth-Heinemann*, 1996.

[77] Halley, Julianne, and David A. Winkler., "Consistent concepts of selforganization and selfassembly," *Complexity*, 14.2 (2008): 10-17.

[78] Crooks, G. E., "Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences", *Phys. Rev. E* 60, 27212726 (1999)

[79] England, Jeremy L., "Dissipative adaptation in driven self-assembly" *Nature Nanotechnology*, 10.11 (2015): 919-923.

[80] Perunov, Nikolay, Robert A. Marsland, and Jeremy L. England., "Statistical physics of adaptation," *Physical Review X* 6.2 (2016): 021036.

[81] Ganesh, Natesh, and Neal G. Anderson, "Irreversibility and dissipation in finite-state automata," *Physics Letters A* 377.45 (2013): 3266-3271.

[82] Still, Susanne, et al., "Thermodynamics of prediction," *Physical Review Letters* 109.12 (2012): 120604.

[83] Kinosita, Kazuhiko, et al., "A rotary molecular motor that can work at near 100% efficiency," *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 355.1396 (2000): 473-489.

[84] Conant, Roger C., and W. Ross Ashby. "Every good regulator of a system must be a model of that system." *Intl. Journ. of Systems Science,* 1.2 (1970): 89-97.

[85] Valiant, Leslie, *Probably Approximately Correct: Nature's Algorithms for Learning and Prospering in a Complex World*, Basic Books (AZ), 2013.

[86] Baez, John, & Blake Pollard, "Relative entropy in biological systems.", *Entropy*, 18.2 (2016): 46.

[87] Watson, Richard A., and Ers Szathmry. "How can evolution learn?." Trends in ecology & evolution 31.2 (2016): 147-157.

[88] Harper, Marc, "The replicator equation as an inference dynamic.", *arXiv preprint*, arXiv:0911.1763 (2009).

[89] Baez, John, "Diversity, Entropy and Thermodynamics," *Exploratory Conference in the Mathematics of Biodiversity*, Centre de Recerca Matemtica, Barcelona, 2012.

[90] Cressman, Ross and Yi Tao, "The replicator equation and other game dynamics.", *Proceedings of the National Academy of Sciences*, 111.Supplement 3 (2014): 10810-10817.

[91] Turrigiano, Gina G., and Sacha B. Nelson. "Hebb and homeostasis in neuronal plasticity," *Current Opinion in Neurobiology,* 10.3 (2000): 358-364.

[92] Still, Susanne, "Information bottleneck approach to predictive inference," *Entropy* 16.2 (2014): 968-989.

[93] Still, Susanne, "Thermodynamic cost and benefit of data representations," *arXiv preprint*, arXiv:1705.00612 (2017).

[94] Dayan, Peter, et al., "The Helmholtz Machine," *Neural Computation*, 7.5 (1995): 889-904.

[95] Hinton, Geoffrey E., and Richard S. Zemel, "Autoencoders, Minimum description length and Helmholtz free energy," *Advances in Neural Information Processing Systems*, 1994.

[96] Rao, Rajesh PN, and Dana H. Ballard, "Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects," Nature neuroscience 2.1 (1999): 79.

[97] Seth, Anil K, "The cybernetic Bayesian brain.", *Open MIND*, Frankfurt am Main: MIND Group, 2014.

[98] Ahmad Subutai, & Jeff Hawkins, "Properties of sparse distributed representations and their application to hierarchical temporal memory.", *arXiv preprint*, arXiv:1503.07469 (2015).

[99] Davies, Mike, et al., "Loihi: A neuromorphic manycore processor with on-chip learning.", *IEEE Micro*, 38.1 (2018): 82-99.

[100] Friston, Karl, "The free-energy principle: a rough guide to the brain?" *Trends in Cognitive Sciences*, 13.7 (2009): 293-301.

[101] Thrun Sebastian B, "Efficient exploration in reinforcement learning." (1992).

[102] Still, Susanne, "Information-theoretic approach to interactive learning.", *EPL (Europhysics Letters)*, 85.2 (2009): 28005.

[103] Sutton, Richard S., & Andrew G. Barto, *Reinforcement learning: An introduction*, MIT press, 2018.

[104] Chentanez, Nuttapong, Andrew G. Barto, & Satinder P. Singh, "Intrinsically motivated reinforcement learning", *Advances in neural information processing systems*, 2005.

[105] McCloskey, Michael, and Neal J. Cohen, "Catastrophic interference in connectionist networks: The sequential learning problem," *Psychology of learning and motivation*, Vol. 24. Academic Press, 1989. 109-165.

[106] Kirkpatrick, James, et al. "Overcoming catastrophic forgetting in neural networks," *Proceedings of the National Academy of Sciences*, 114.13 (2017): 3521-3526.

[107] Zhang, Byoung-Tak, "Information-Theoretic Objective Functions for Lifelong Learning," *AAAI Spring Symposium: Lifelong Machine Learning*, 2013.

[108] Creutzig, Felix, Amir Globerson, and Naftali Tishby, "Past-future information bottleneck in dynamical systems," *Physical Review E*, 79.4 (2009): 041925.

[109] Ulieru, Mihaela, and Rene Doursat, "Emergent engineering: a radical paradigm shift," *International Journal of Autonomous and Adaptive Communications Systems*, 4.1 (2010): 39-60.

[110] Teuscher, Christof, "The weird, the small, and the uncontrollable: Redefining the frontiers of computing," *Computer*, 50.8 (2017): 52-58.

[111] Avizienis, Audrius V., et al., "Neuromorphic atomic switch networks," *PloS One*, 7.8 (2012): e42772.

[112] Bose, Saurabh K., et al, "Stable Self-Assembled Atomic-Switch Networks for Neuromorphic Applications," *IEEE Transactions on Electron Devices*, 64.12 (2017): 5194-5201.

[113] Lancaster, Madeline A., et al., "Cerebral organoids model human brain development and microcephaly," *Nature*, 501.7467 (2013): 373.

[114] Nguyen, Michael, and Suriyanarayanan Vaikuntanathan, "Design principles for nonequilibrium self-assembly," *Proceedings of the National Academy of Sciences*, 113.50 (2016): 14231-14236.

[115] Website of the *IEEE Rebooting Computing Initiative*, https://rebootingcomputing.ieee.org/about

[116] Turing, Alan, "Computing machinery and intelligence," *Parsing the Turing Test*, Springer, Dordrecht, 2009. 23-65.

[117] Turing, Alan, "Intelligent machinery, a heretical theory (c. 1951)," *Jack Copeland B*, (2004): 465.

[118] Pinker, Steven, "How the mind works," *Annals of the New York Academy of Sciences* 882.1 (1999): 119-127.

[119] Searle, John R., "Minds, brains, and programs," *Behavioral and brain sciences*, 3.3 (1980): 417-424.

[120] Searle, John R., and S. Willis, "Consciousness and language," Cambridge University Press, 2002.

[121] Domingos, Pedro, "A few useful things to know about machine learning," *Communications of the ACM*, 55.10 (2012): 78-87.

[122] *A Beginner's Guide to Differentiable Programming*,

[123] Hylton, Todd, "Thermodynamics and the Future of Computing," *IEEE International Conference on Rebooting Computing*, 2017.

[124] The EU is trying to decide whether to grant robots personhood,

[125] Everything you need to know about Sophia, the world's first robot citizen,

[126] Kirkpatrick, James, et al. "Overcoming catastrophic forgetting in neural networks," *Proceedings of the National Academy of Sciences*, 114.13 (2017): 3521-3526.

[127] Marcus, Gary, "Deep Learning: A Critical Appraisal," *arXiv preprint*, arXiv:1801.00631 (2018).

[128] Rahimi, Ali, "Test of Time award - acceptance speech," *Neural Information Processing Systems*, 2017,

[129] Shepherd, Gordon M., et al, "The Human Brain Project: neuroinformatics tools for integrating, searching and modeling multidisciplinary neuroscience data," *Trends in neurosciences*, 21.11 (1998): 460-468.

[130] Lancaster, Madeline A., et al., "Cerebral organoids model human brain development and microcephaly," *Nature*, 501.7467 (2013): 373.

[131] Wang, Zhongrui, et al., "Fully memristive neural networks for pattern classification with unsupervised learning," *Nature Electronics*, 1.2 (2018): 137.

[132] *Koniku - Intelligence is Natural*,

[133] Ganesh, Natesh, "Intelligence is Physical," *Under preparation*, 2018.

[134] Pathria, R K., "Statistical Mechanics," *Oxford: Butterworth-Heinemann*, 1996.

[135] Crooks, G. E., "Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences", *Phys. Rev. E* 60, 2721 - 2726 (1999).

[136] England, Jeremy L., "Statistical physics of self-replication," *The Journal of chemical physics*, 139.12 (2013): 09B623.

[137] Turrigiano, Gina G., and Sacha B. Nelson. "Hebb and homeostasis in neuronal plasticity," *Current Opinion in Neurobiology,* 10.3 (2000): 358-364.

[138] Ganesh, Natesh, "A Thermodynamic Treatment of Intelligent Systems," *IEEE International Conference on Rebooting Computing (ICRC)*, IEEE, 2017.

[139] Halley, Julianne, and David A. Winkler., "Consistent concepts of selforganization and selfassembly," *Complexity*, 14.2 (2008): 10-17.

[140] Tishby, Naftali, Fernando C. Pereira, and William Bialek, "The Information Bottleneck Method," *arXiv preprint physics*/0004057 (2000).

[141] Still, Susanne, "Thermodynamic cost and benefit of data representations," *arXiv preprint*, arXiv:1705.00612 (2017).

[142] Rao, Rajesh PN, and Dana H. Ballard, "Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects," Nature neuroscience 2.1 (1999): 79.

[143] Dayan, Peter, et al., "The Helmholtz Machine," *Neural Computation*, 7.5 (1995): 889-904.

[144] Hinton, Geoffrey E., and Richard S. Zemel, "Autoencoders, Minimum description length and Helmholtz free energy," *Advances in Neural Information Processing Systems*, 1994.

[145] McCloskey, Michael, and Neal J. Cohen, "Catastrophic interference in connectionist networks: The sequential learning problem," *Psychology of learning and motivation*, Vol. 24. Academic Press, 1989. 109-165.

[146] Friston, Karl, "The free-energy principle: a rough guide to the brain?" *Trends in Cognitive Sciences*, 13.7 (2009): 293-301.

[147] Karl, Friston, "A free energy principle for biological systems," *Entropy* 14.11 (2012): 2100-2121.

[148] Friston, Karl, and Stefan Kiebel, "Predictive coding under the free-energy principle," *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364.1521 (2009): 1211-1221.

[149] Ulieru, Mihaela, and Rene Doursat, "Emergent engineering: a radical paradigm shift," *International Journal of Autonomous and Adaptive Communications Systems*, 4.1 (2010): 39-60.

[150] Teuscher, Christof, "The weird, the small, and the uncontrollable: Redefining the frontiers of computing," *Computer*, 50.8 (2017): 52-58.

[151] Avizienis, Audrius V., et al., "Neuromorphic atomic switch networks," *PloS One*, 7.8 (2012): e42772.

[152] Scharnhorst, Kelsey S., et al., "Atomic switch networks as complex adaptive systems" *Japanese Journal of Applied Physics*, 57.3S2 (2018): 03ED02.

[153] Bose, Saurabh K., et al, "Stable Self-Assembled Atomic-Switch Networks for Neuromorphic Applications," *IEEE Transactions on Electron Devices*, 64.12 (2017): 5194-5201.