AN INFORMATION-THEORETIC APPROACH TO CELLULAR DECISION-MAKING STRATEGIES:

How Rate Distortion Theory Provides an Optimal Method for Describing Binary Cellular Decision-Making Systems

by

Joshua Mangum

A Senior Honors Project Presented to the

Honors College

East Carolina University

In Partial Fulfillment of the

Requirements for

Graduation with Honors

by

Joshua Mangum

Greenville, NC

May 2014

Approved by:

Ioannis Gkigkitzis

Abstract

Rate distortion theory, a branch of information theory, was originally developed to help improve the efficiency of data transmission in telecommunications. It's currently being used as a major modeling method to provide a quantitative description for analyzing biological signaling pathways. Rate distortion theory provides a way to compute probability functions that describe how cells should respond given various stimuli or environmental changes, independent of the mechanism responsible for these decisions. In this thesis, mathematical models describing binary cell decisions will be studied and analyzed within the framework of rate distortion theory.

In this project we discuss the history, terminology, mathematical structure, and major aspects of rate distortion theory. These aspects of the theory will then provide the foundation for how can be applied in a biological context. The principle elements of these models depict cellular decision-making strategies as conditional probabilities, where environmental stimuli such as temperature fluctuations or concentration gradients are considered to be the input. The decisions made in response to these changing stimuli are the output of the algorithm. A rate distortion function defines the average amount of "incorrect" decisions given a stimulus, and a rate distortion curve quantifies stochastically, the fate of the given cell, given the stimulation. A Blahut Arimoto algorithm is used to compute the rate distortion curve that provides the optimal decision-making pathways.

According to Perkins and Swain, (Perkins and Swain, 2009) cellular decision-making has the following main features: a cell must (1) estimate the state of its environment by sensing stimuli; (2) make a decision informed by the consequences of the alternatives; and (3) perform these functions in a way that maximizes the fitness of the population. The consistency of these axioms and the effort to investigate, explain, and interpret observable characteristics of cellular functions such as hysteresis, irreversibility, and random strategies will be discussed. This theory provides a method for explaining why cells partake in self-destructive behavior such as apoptosis in order to benefit the population of the cells.

Table of Contents

Chapter 1

Introduction

Rate distortion theory, a branch of Information Theory, was first proposed by Claude Shannon in 1948 to help increase the fidelity of data transmission. It has in recent years been increasingly applied to analyzing biological decision-making pathways by describing how distorted a cellular decision is given certain extracellular environments [1].

Chapter 2 of this paper will demonstrate how rate distortion theory was originally developed to transmit all types of data using the same method of encoding and decoding information as 1's and 0's, which allows information to be transmitted more efficiently. A brief account of its natural trajectory following its foundation will be outlined to show how it has been applied to various other fields such as biology.

Chapter 3 will briefly explain the difference between lossy and lossless data compression, and why lossy data compression is often more ideal in many circumstances. Also a summary of common notations will be given and explained to readers who are unfamiliar with rate distortion theory so they may become accustomed to how simple aspects of this field are utilized. The most important aspects of information theory such as entropy, conditional distributions, and average mutual information will be explained with analogous examples in some detail to help familiarize the readers with the language employed by lossy data compression. This section will also demonstrate how to calculate a rate distortion function for a discrete memoryless source (d.m.s.), and how the rate distortion curve can be found once the probability distribution and rate distortion

function are known for a given d.m.s. Lagrange multipliers are the methods used to determine how the rate distortion curve is conceived, so a brief account of how to use Lagrange multipliers will also be given; however, it is assumed that the reader knows how to take simple first derivatives of unchallenging problems. A proof is provided for both the rate distortion curve and the Blahut-Aritmoto algorithm, which should be understandable provided that the reader knows the definition of a derivative.

Finally, section 4 will show how biological systems can be described under rate distortion theory, and how various people have exploited the aspects outlined in section 3 to cellular decision-making strategies. It will be demonstrated how it's possible to use the rate distortion curve and the Blahut-Aritmoto algorithm to describe how optimal a cellular decision-making strategy is in any given situation. The benefits of using rate distortion theory over a more direct approach will be stated in the conclusion section.

Chapter 2

Origins and History of Information Theory

Prior to examining the terminology and notation that's frequently used in information theory, I believe it's important to show a brief account of its history. By reviewing the origins of the theory one can get a general realization of why it was developed and why it was so revolutionary in redefining how we transmit data in the majority of all electronic devices. There was relatively little knowledge regarding how to transmit messages before the foundation of this theory, and it currently serves as the blueprint for transmitting data in all modern-day electronic devices. By examining a brief account of its trajectory after its foundation, one can gain insight into how the logarithms and concepts of information originated in other unrelated disciplines such as molecular biology. This section serves to familiarize readers with why information theory was developed for use in telecommunications, how it differed from the traditional methods of transmission that were used prior to it, and why it's creation helped ease the process of data transmission. It also demonstrates how it found its way into other fields besides telecommunications.

2.1: Origins of Information Theory

Prior to 1948, there existed only a vague notion of what a message with respect to communication transmission was. Communication engineers possessed a general idea of how to transmit and receive a waveform over a wire, but they lacked the knowledge of how to go about transforming a message into a waveform that could then be transmitted. In 1948, Claude Shannon published his first paper, "A Mathematical Theory in Communication" in the *Bell*

Systems Technical Journal [5]. This formulated the suggestion of how to turn a message into a waveform. The paper demonstrated how all modes of communication, including telephone signals, texts, radio waves, pictures, etc., could all be encoded in bits that would allow them to be transmitted by quantifying the information with thorough accuracy [5].

There were four new, revolutionary concepts that were first proposed by Shannon in his 1948 paper that led him to become known as "The father of the digital age [5]." One of the most exalted of these concepts was the idea found in his Noisy Coding Theorem. Shannon was the first person to state that all channels that transmit waveforms have a "speed limit measured in binary bits per second called the capacity of the channel. He claimed that it was mathematically impossible to get error free communication above this limit using any scheme of coding [5]." However, it was possible to encode up to this limit, even if there's an abundance of static or electronic noise present or if the transmitting signal is faint. In order to transmit a signal in the presence of excessive noise, additional bits must be added so that the majority of bits will go through, and those that don't go through can be restored by those that did. Bits are units used to quantify information in data transmission. Of course adding additional bits implies that the span of the message would increase and consequently would cause a slower transmission, but this new idea of enabling an engineer to construct a communication system with a probability of error as low as desired was of paramount importance. "This noisy channel coding theorem gave rise to error-correcting codes-the process of introducing redundancy into the digital representation to protect against corruption [5]." This can be demonstrated by how a CD can have minor scratches but still produce music without skipping.

Shannon also created the blueprint for the architecture and design of the communication system that is capable of being separated into distinct components, each of which is capable of being regarded as its own discrete mathematical model. The majority of all modern day communication systems such as computers and the design of the telephone exchange are based on Shannon's original notation of how to design a communication system [5].

The purpose of these communication systems is essentially to communicate a set of pre-selected messages from the source to the user in the presence of noise; this system can be depicted by the block diagram in Figure 2.1. The source produces information that is to be sent over the channel and presented to the user. The encoder is an apparatus that turns the input into a waveform that can be transmitted over the channel, and the decoder then transforms the waveform back into a form that can be presented to the user at the destination. The channel is simply a tangible medium which connects the source and the user where the waveform is transmitted across [4]. Generally, when referring to designing a communication system the engineer is free to construct the encoder and decoder given a certain source-user combination and a channel; i.e. the engineer has no power in altering the channel, source, or user. Associated with each channel is a capacity first defined by Shannon as the maximum rate at which information can be sent over a channel with no error produced at the output. Any two channels of separate communication systems may be considered equivalent if they share the same capacity [4].

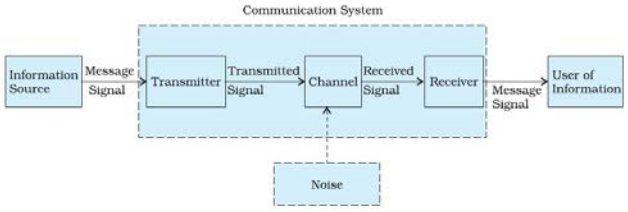

Figure 2.1: The block diagram is a depiction of any communication system. The encoder turns the input form into 1's and 0's so that it can be transmitted over the channel to the decoder. The decoder then reconverts it back into a form that the user can understand at the destination. The channel is a physical medium that connects the source and the user [3-5].

The third of Shannon's remarkable ideas was that the content of the transmitted message was not dependent upon its transmission. All forms of communication such as texts, sounds, pictures, audio, and video may conceivably be regarded as 1's and 0's [5]. Once these forms of communication are represented as bits digitally, they may be sent over the communication channel. Prior to this idea, most communication engineers worked in their own separate, respected fields, because each field had varying methods of how to transmit the type of

information they worked with as an electromagnetic waveform over a wire. Audio and imaging transmission had nothing to do with each other prior to Shannon's discourse [5].

Finally, Shannon developed the postulation of source coding (better known as data compression) to better the efficiency of data depiction by eliminating redundancy in the information. A demand to convey information at increased speeds may generally necessitate that information is sent over a communication channel at a rate that transcends the capacity of the channel. Consequently, some amount of distortion is inevitable in such situations. Too keep this distortion to a minimum, one must prioritize the data produced by the source in conformity with its importance, and the less pertinent information should either be compacted or deleted preceding transmission. Data compression algorithms are the strategies that are designed to obtain the more important information from the output of a source and to eradicate the more superfluous data. Rate distortion theory is the deductive field that treats data compression from the perspective of information theory. In his treatise Shannon states, "The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point [4]." This primary problem can be summarized into two constituent questions: "(1) What information should be transmitted? and (2) How should this information be transmitted? [4]" The preponderance of Shannon's exposition is directed towards addressing the second question. However, according to Berger, there exists a dichotomy that the two questions cannot be separated from one another. A quandary is often encountered when developing intricate encoding and decoding techniques to transmit data at dependable rates that approach the capacity of the channel. That is, after figuring out how to transmit the data, it's not always apparent what information should be sent. The ramification is that a considerable amount of the data transmitted could either be redundant or unimportant. The majority of the literature regarding information theory is directed to addressing Shannon's second question, and there's still much more research to be done on question 1 [4].

2.2 History of Information Theory

The proposals brought forth by Shannon quickly acquired an immense amount of attention shortly after their publication because of the new notion that something as obscure as information could be reduced to a mathematical model to allow its transmission. Despite his intentions of directing his paper to communication engineers, his theory quickly found its way into the mass media by the 1950's [5]. However, it should be noted that information theory isn't entirely a result of Shannon's 1948 paper. Although most popular assertions attribute Shannon as the founder of this theory, Kline states that it should be regarded as the product of a decade worth of parley between mathematicians, physicists, engineers, social scientists, and humanists [6]. They debated over issues such as the principle interpretation of "information" and attempted to designate the fields where it could be properly or feasibly used and those where it couldn't.

As early as the 1950's, social and natural scientists related this theory to several diverse disciplines including experimental and cognitive psychology, linguistics, physiology, and molecular biology [6]. At first, many scientists believed this new theory underpinned all foundational circumstances regarding communication. It wasn't until the following decade of the 1960's where it was successfully applied to space and computer communication. Prior to these applications of the theory, many engineers remained skeptical about its use regardless of being Shannon's target audience [6]. The ultimate result was a debate regarding what "information" actually was and what its proper applications were throughout the 1950's. The argument was naturally resolved in the 60's as the theory was adopted by several fields and modified accordingly to suite the needs of these fields. It was proven to be appurtenant to some disciplines but useless to others. Debates regarding the boundaries of this theory helped to expand its applications where proven beneficial; the theory was eventually excluded from areas where it was deemed unrelated from people speaking out against its misuse in *IEEE* [6].

Throughout the duration of the 50's and 60's the term "information theory" and "information" referred to a wide variety of ideations [6]. In 1953, an electrical engineer at MIT

named Robert Fano claimed there were four primary interpretations regarding what "information" actually consists of; three of these interpretations were closely related to Shannon's ideas and the fourth involved "miscellaneous philosophical speculations on broad communication problems [6]." One interpretation was proposed in Shannon's paper. He found a correlation between the uncertainty in information and the signal transmitted based off this information. He concluded that, "The greater the uncertainty about which symbols a discrete source would select from a set, the more information the source produced [6]." The subject matter or content of the symbols was unimportant to Shannon, because he had discovered that all types of media, languages, and images could be conveyed in the same manner. He described the channel capacity and amount of information as positive entropy to construct a code that could transform data into signals that allow it to be sent over a noisy channel with a pre-determined amount of error [6]. Thus, "information" according to Shannon is what represents entropy, which reduces the uncertainty in transmitting symbols in order to decrease the amount of error in a communications system. It effectively provided a correlation between the complexity of the communication system and the fidelity of its data transmission.

Wiener's similar interpretation symbolized messages from noisy signals, these messages he regarded as "information." Weiner derived a value of negative entropy to help filter signals and represent messages in the presence of noise. Shannon had positive entropy to code messages into signals that could be transmitted in the presence of noise. Both values of entropies are similar in notation with the exception of the negative sign for Weiner's value. The sign difference may be attributed to the different approaches in the use of entropy [6]. Shannon and Weiner both published their research independently from each other in the 1940's. Despite claiming he was influenced in part by Weiner's research, Shannon approached a comparable problem with slightly different methods, and he referred to his work as "information theory" in his talks and works early on. Weiner used the term "information theory" in a more general sense [6].

Ronald Fisher prior to World War II gave a third interpretation of "information". He derived a similar correlation between entropy and information as Shannon and Wiener. In his theory when there are unknown statistical parameters, the "amount of information" that can be expected may be found from a certain number of observations. Fisher's interpretation was used to analyze waveforms in a wire of a communications system by Denis Gabor in 1946 [6].

Just a couple of years after the publication of "A Mathematical Theory of Communication", Shannon's single paper encouraged various universities to offer seminars on information theory which quickly became classes including one taught by Shannon himself at MIT called "Advanced Topics in Information Theory [5]." In 1951, just 3 years after Shannon's first paper, a journal titled *The IRE Transactions of Information Theory* was launched whose principal concern was to promote education. Symposia were also created as well that searched for the most prestigious papers related to the field; these conferences helped to augment the scope of the theories. In the first conference held in 1950, a physicist named Donald MacKay attempted to unify all the interpretations of "information." He declared the expansive field referred to as information theory as "the making of representation" in science, communication, and the arts [6]. MacKay believed that the different concepts of "information" should be clarified by their proper objectives, and he coined phrases such as *scientific information theory*, *structural information*, and *metrical information* to refer to the various applications of "information theory" [6]. By the third of such symposia's held in 1956, the breadth of the field had grown so vast that it contained participants in disciplines such as animal welfare, anthropology, and political theory, which clearly had no basis to employ any ideas proposed by Shannon, Weiner, or Fisher. Persons in specialties like neurophysiology were trying to describe how birds communicated songs to each other in the presence of noise using information theory. Conversely, there were many disciplines at these symposia's such as statistics, computer science, and physics that could facilitate the ideas of these theories in an applicable way [5].

Despite gaining immense popularity, many engineers remained doubtful over many uses regarding "information." In the 1952 symposium Robert Fano publically announced his disproval over the emerging research regarding "information theory" [6]. Fano stated in a letter that the confusion in the four broad areas regarding "information theory" was responsible for his view of the misunderstandings. He decided to limit the term "information theory" to Shannon's work; others would eventually do the same [6].

It was discovered that the concepts of information theory had several attractive qualities to governments as well. Shannon published his first paper shortly after World War II. During this time period the essence of long-distance communication had evolved from electromagnetic waveforms to data, so computers became the primary means of receiving information and logarithms were needed [5]. The launch of Sputnik also generated interest because of the desire for dependable and accurate communication in the presence of excessive noise that's encountered in space. The field began to increase rapidly as more coding theorems were created for a myriad of applications in communications, and the data-transmission rate grew exponentially [5]. By the end of the 1960's, the research between Shannon and Wiener had been separated into two groups within the electrical engineering community. Shannon's work was referred to as "information theory," and Wiener's work was referred to as "statistical communication theory" [6]. The IEEE accepted specification of these titles. Instead of agreeing on a ubiquitous definition of "information," researchers could now create boundaries between the various concepts of "information" and use the appropriate interpretations that now had distinct titles with their own formulas. It ended the dispute over the definition of "information" in that it concluded there are many definitions each with their own applications. It provided a closure for the scientific debate, and researchers could now apply these separate theories to MacKay's taxonomy. That is they could apply the theories to others fields besides communication without having to argue over competing interpretations of "information" [6].

Towards the end of the 1960's the apparent need for applying the use of information theory as proposed by Shannon to new areas began to dwindle. As it progressed along its natural trajectory it became a social phenomenon, almost like a fad at first [5, 6]. His ideas were inspired by the works of Weiner and Fisher, which consequently had competing interests proceeding the publication of his first paper. The decade long debate that followed concluded with Shannon's work ultimately being referred to as information theory. Throughout the duration of this debate his theory was increasingly applied to a plethora of separate fields along with the theories of Weiner and Fisher. Arguments followed rather Shannon's theory was actually applicable to these fields or not, and the fields were it wasn't deemed permissible were convinced to stop publishing research using his ideas. Shannon's principle concepts that separated him from his counterparts was the notion of transforming a message into a waveform that could be transmitted over the wire and the entropy associated with the uncertainty of these messages to achieve a certain fidelity criterion. While his notion of entropy closely corresponds to that of Wiener's, it was Shannon's concept of entropy that would prove to be the most valuable and most used. Today information theory is used to help quantify things such as cellular decision-making strategies. Applications such as this can be attributed in part to these early symposiums where the concepts of information being used in various fields were discussed, debated, and analyzed by many several contributors [6].

Chapter 3

How to Use Rate Distortion Theory

Rate distortion theory is a branch of information theory where the designer of the communication system must settle for some amount of distortion between the input and the output. It is assumed that the reader has no formal knowledge regarding how to calculate rate distortion functions or use the common notations and terminology of the theory. Before learning how to use aspects of rate distortion theory, it's first important to learn how it differs from the rest of information theory. After becoming aware of how rate distortion theory can be more applicable in designing communications systems in real-life scenarios, then it's feasible to acquire the skills necessary to do so. One must become familiarized with the common notions and what each variable represents before learning how to perform calculations with them. Afterwards, it's important to understand other alternative methods for performing these calculations that are often quicker, easier, and more intuitive.

This chapter overviews the difference between rate distortion theory (also known as lossy data compression) and lossless data compression. It then provides the theoretical framework that underpins rate distortion theory starting with explicit examples of the common terminology with analogies to help comprehend them. Afterwards proofs are given to show how the rate distortion function and rate distortion curve are derived. Finally alternative methods for calculating these functions are presented.

3.1: Lossless Data Compression Vs. Lossy Data Compression

Shannon forged the concept of lossless and lossy data compression in his innovating paper "A Mathematical theory of communication [5]." He designated lossless data compression as a verbatim replication of the original data; the data produced at the output is a perfectly accurate duplication of the original data transmitted from the input. Shannon hypothesized that there's an intrinsic limit to lossless data compression called the entropy source *H*, which depends upon the statistical nature of the source. If the entropy source is less than the channel capacity *C*, then no distortion should be expected. It's conceivable to compress the data with a compression rate close to *H*; however, it's mathematically impossible to transmit information at a rate that surpasses *H* without some amount of distortion *D* [7].

With lossy data compression (more commonly referred to as rate distortion theory) some amount of distortion is tolerated between the source and the receiver. Shannon formulated that if there is an amount of tolerable distortion, then there exists a rate distortion function, represented by $R(D)$, which yields the best possible compression rate. The rate distortion function may be contrived if there is a given source with all its statistical properties known and a given distortion measure, or a mathematical apparatus that specifies how close the distortion of the receiver is to the original data transmitted by the source [7]. If no distortion is present, then $D=0$ and $R(0)=H$, which represents lossless data compression where the best possible compression rate is simply the entropy source. Lossy data compression is merely an expansion of this concept where the communication system moves away from no distortion ($D=0$) to some amount of distortion ($D>0$) [7]. However, decreasing the distortion in lossy data compression requires an increase in the complexity of the communication system. Therefore, it generally costs more money or requires more resources to construct a communication system that has lower levels of distortion. The primary question posed with rate distortion theory is then, "How to balance the cost of intricate communication systems with the benefit of accurate transmission of the data?" [1, 4] The rate distortion curve answers this question by giving the minimal amount of mutual information

needed between the input and the output to transmit data with an upper limit of expected distortion. This curve effectively shows the correlation between how much distortion is tolerable (how efficiently to transmit data) versus how much mutual information is required to achieve this distortion, which corresponds to the resources required to produce it.

3.2: Common Notations and Terminology Used in Rate Distortion Theory

It's conventional to denote a finite set $\{a_1, a_2, \ldots a_M\}$ as an alphabet; the letters refer to the elements of the alphabet. An alphabet is considered to be of size *M* if it contains *M* discrete letters, and the letters may represent any type of information. In information theory an alphabet of size *M* should be defined as [4]:

$$A_M=\{0, 1, 2, 3, \ldots, M\text{-}1\}. \tag{3.2.1}$$

A probability distribution is a function the projects the alphabet A_M into the space [0, 1]. It meets the requirement that the sum of all probabilities of picking any letter at random equals 100% [4]. This is represented as

$$\sum_{j=0}^{M-1} P(j)=1. \tag{3.2.2}$$

Another way of imagining equation 3.2.2 is by a six-sided dice. The probability of rolling any number, for instance the number 2, is $\frac{1}{6}$. This equation gives the sum of all possible probabilities. That is, the sum of the probabilities for rolling each number is $\frac{1}{6}+\frac{1}{6}+\frac{1}{6}+\frac{1}{6}+\frac{1}{6}+\frac{1}{6}$=1, or 100%. A finite ensemble is characterized by (A_M, *P*) [4].

"In statistics random variables are numerical quantities of a probability experiment, so its value is determined by chance, and they are often denoted using letters such as *X [8]."* For example, if a six-sided dice is rolled four times and the random variable *X* represents the number of times the dice landed on one, then the possible values of *X* are 0, 1, 2, 3, or 4. With respect to

information theory, a random variable is any function defined on the alphabet A_M of an ensemble (A_M, P); a random variable is generally abbreviated as r.v. in information theory [4]. An r.v. may contain a finite number of discrete values or it may contain an infinite number of values over a range on a number line [8]. If the r.v. has a range over a real line, it's referred to as a real r.v. A real r.v. denoted by f that's correlated with the ensemble (A_M, P) has the expected value [4]:

$$E[f] = \sum_{j=o}^{M-1} P(j) f(j). \tag{3.2.3}$$

To illustrate equation 3.2.3, suppose you wanted to find the probability of selecting a person who is 6 feet tall that lives in Chicago. Then $f(j)$ would be the real r.v., in this case the height of a person randomly selected that lives in Chicago. $P(j)$ is the probability of selecting a person who corresponds to the given r.v., here it's the probability of selecting a person who is 6 feet tall that lives in Chicago. The expectation $E[f]$ represents the average height of all the people that live in Chicago and is given from sum of the probability distribution (probability of selecting a person of a certain height in Chicago) times the r.v. (the height of that person).

A very important r.v. in information theory is self-information defined by Berger as, "The information one receives upon being told that the r.v. X has assumed the value of j." It can be represented by [4]:

$$i(j) = -\log P(j). \tag{3.2.4}$$

The concept of self-information can be easily demonstrated with a simple example by examining figure 3.2.1. The x-axis represents the probability of an event occurring, and the y-axis represents the surprise one encounters upon the occurrence of that particular event. If the probability of an event is 1 or 100%, then there is 0 surprise associated with that event, and it's impossible to have an event with a probability greater than 1. As the probability decrease from 1, the surprise associated with that event increases and approaches infinity asymptotically. Too demonstrate, imagine a student sitting in class before the class begins. The probability of the teacher walking

into the classroom before class starts has a very high probability near 1; consequently, it has little surprise associated with it. If the teacher didn't walk into the room and failed to show up for class, then there would be more surprise associated with that event, because the probability is lower. However, if the president or the Pope walked into the classroom prior to class, then there would be an immense amount of surprise associated with the event because of the low probability of that occurring. This straightforward example is analogous to the measure of information one receives upon being told the r.v. *X* has assumed the value of *j*. In other words, if the probability of *j* is very high, then there is little information provided to the user upon deducing that *X* has taken that value, and the opposite is true if the probability of *j* is low.

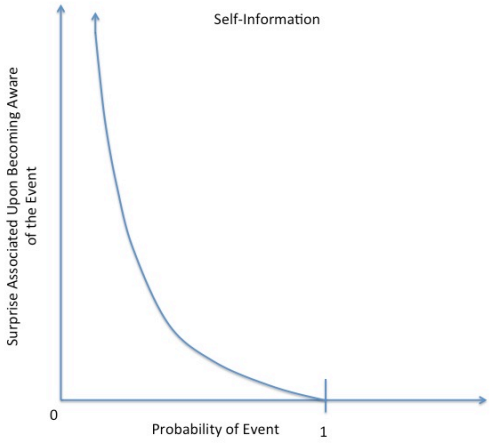

Figure 3.2.1: A graph demonstrating the self-information as the information one receives upon being informed that *X* has taken the value *j*.

The entropy of a source is often defined as the self-information of an r.v. It's described as [4]:

$$H(X) = -\sum_{j=0}^{M-1} P(j)\log P(j). \tag{3.2.5}$$

Equation 3.2.3 gave the expected (or average) value of a random variable *f* when the sum of all possible values of that r.v. were multiplied by the probability of having that value. Similarly, equation 3.2.5 shows the sum of the product between the r.v. that represents self-information when *X*=*j* and the probability of attaining that value of *j*. Therefore, the entropy can be regarded

"as a measure of the average a priori uncertainty regarding which value X will assume, or the information one receives upon being told what value has been assumed by the r.v. $X(j)=j$[4]." The entropy $H(X)$ is what Shannon originally defined as the inherent limit to lossless data compression as well. The graph $H(X)$ versus P is shown in figure 3.2.2. It demonstrates that the maximum value of $H(X)$=1 bit is obtained when P=1/2. This makes sense because the uncertainty is maximum when the probability is 50%; it's equally likely for an event to occur or not to occur. Thus by assuming that the probability is 50%, you get the least amount of a priori knowledge regarding which value X will assume [3]. Also $H(X)$ is 0 when P is either 0 or 1. This also obeys intuition because when P=1 or 0, the variable is not random and no uncertainty exists. It's inevitable for the event to either occur or not to occur respectively, so there's no uncertainty regarding which value X will assume [3].

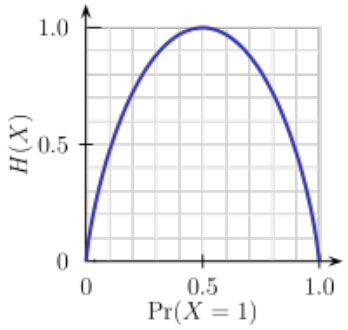

Figure 3.2.2: This graph represents the entropy of the source demonstrating that $H(X)$ is highest when P=1/2 and lowest when P=0 or P=1 [3].

The concept of self-information and entropy must now be expanded to include two alphabets in order to describe how the reproduced data differs from the original data when traversing a communication system. Given two alphabets A_M and A_N, we let $j \in A_M$ and $k \in A_N$, then $P(j,k)$ is called the joint distribution and is defined on the product space A_{MN} [4]. For example, imagine an alphabet A_M having letters {1, 2, 3, 4,} and an alphabet A_N having letters {red, green, yellow, purple}. The joint distribution $P(j,k)$ would refer to the probability

distribution associated with picking a letter j from alphabet A_M times the probability of picking a letter k from the alphabet A_N, when presented with discrete alphabets. The probability of selecting $P(2, \text{yellow})=P(j, k)=P(j)Q(k)=(1/4)(1/4)=1/16$. The concept of joint distributions may be represented in the following table where the columns are probabilities associated with picking the letter j from the alphabet A_M, and the rows represent probabilities associated with picking the letter k from the alphabet A_N.

j/k	−1	0	1
10	.2	.1	.05
20	.3	.15	.2

Figure 3.2.3: This table shows an example of a joint probability distribution $P(j, k)$. It represents all the possible combinations of picking the letter j from alphabet A_M and picking letter k from alphabet A_N. The sum of all probabilities is 1, or 100%.

This table represents the probability of every combination of picking the letter j from the alphabet {-1, 0, 1} and picking k from the alphabet {10, 20}. All probabilities for every combination of (j,k) must also add to 100%. Here the probability of picking (-1, 20) is 20% [4]. You can obtain the marginal distribution of $P(j)$ represented by equation 3.2.2 from the joint distribution as well by adding across a column for a given value of j. This can be depicted by the following table.

j/k	−1	0	1	*%total*
10	.2	.1	.05	.35
20	.3	.15	.2	.65
%total	.5	.25	.25	1

Figure 3.2.4: This table is the same as figure 3.2.2. with the sum of all probabilities added in the rows and the columns. It demonstrates that the sum of the rows, or picking k from alphabet A_N adds to 1. The sum of the columns or picking j from alphabet A_M also adds to 1.

This table shows that the total probability of picking j as -1 is 50%; it's obtained from adding across the rows for all k values associated with j=-1. A similar marginal distribution for k may be obtained by adding across the columns for all j values associated with a given k value. The sum of all values for j would then be 100% in accordance with equation 3.2.2. Also, the sum of all k values must add to 100%, which is also demonstrated by the table. The marginal distributions can also be represented by the following equations [4]:

$$P(j)=\sum_{k=0}^{N-1}P(j,k) \tag{3.2.6}$$

and

$$Q(k)=\sum_{j=0}^{M-1}P(j,k) \tag{3.2.7}$$

Conditional distributions can also be obtained from joint distributions. They can be expressed by asking what does assuming the letter *j* tell us about the probability of picking the letter *k*, or vice versa. This can be found be dividing the joint distribution by the marginal distribution that we are conditioning by. Too demonstrate, if we take the first table where we defined the joint distribution, what is the probability of picking *k*=20 if we assume *j*=0?

j/k	−1	0	1	*%total*
10	.2	.1	.05	.35
20	.3	.15	.2	.65
%total	.5	.25	.25	1

The table demonstrates that the probability of receiving 20 from the *k* alphabet when *j*=0 is: $Q(k|j)=\frac{P(j,k)}{P(j)}=\frac{.15}{.25}=\frac{3}{5}$ or 60%. Conditional probabilities are defined as [4]:

$$P(j|k)=\frac{P(j,k)}{Q(k)} \tag{3.2.8}$$

and

$$Q(k|j)=\frac{P(j,k)}{P(j)}. \tag{3.2.9}$$

The conditional distribution in equation 3.2.9 represents the probability of the occurrence of $Q(k)$ if the event of $P(j)$ is already known. Equation 3.2.8 represents the probability of the occurrence of $P(j)$ if the event of $Q(k)$ is already known [4].

One of the most important real r.v.'s encountered in information theory that's defined over joint ensembles is the conditional self-information defined by [4]:

$$i(j|k)=-\log P(j|k). \tag{3.2.10}$$

If the occurrence $Y=k$ has already taken place then the conditional self-information represents the information received upon being informed of the occurrence of $X=j$. It's similar to the concept of self-information explained earlier.

By combining 3.2.4 with 3.2.10 you obtain the mutual information, which is a measure of the amount of information one r.v. contains about another [4]:

$$i(j;k) = i(j) - i(j|k). \tag{3.2.11}$$

Mutual information can also be described qualitatively as the amount of information that the occurrence of $X=j$ communicates to someone ignorant of what value Y has assumed minus the information that's communicated to someone who already knows of the occurrence of $Y=k$ [4].

Accordingly, the conditional entropies are a measure of the average a priori uncertainty concerning which value X has assumed if one already knows how Y has been identified and vice versa. They are described as [4]:

$$H(X|Y) = -\sum_{j,k} P(j,k) \log P(j|k) \tag{3.2.12}$$

and

$$H(Y|X) = -\sum_{j,k} P(j,k) \log Q(k|j). \tag{3.2.13}$$

The average mutual information may be obtain by subtracting 3.2.12 from 3.2.5 to yield $I(X;Y) = H(X) - H(X|Y)$. This can also be expressed as [4]:

$$I(X;Y) = \sum_{j,k} P(j,k) \log\left(\frac{P(j,k)}{P(j)Q(k)}\right). \tag{3.2.14}$$

"The average mutual information expresses the average information supplied about X by specification of Y and is equivalent to the average a priori uncertainty in X minus the average uncertainty in X that still remains after Y is specified [4]." The relationship between $H(X)$, $H(Y)$, $H(X|Y)$, $H(X, Y)$, $H(Y|X)$, and $I(X; Y)$ can be shown by the Venn diagram in figure 3.2.5. As demonstrated by the picture $I(X;Y)=I(Y; X)$. In other words, the knowledge supplied about X by

specification of *Y* is equal to the knowledge supplied about *Y* by specification of *X*. This can also be described as the intersection of information between *X* and *Y [3]*. The average mutual information is directly associated with the rate distortion curve as stated in section 3.1. Higher mutual information leads to lower rates of distortion but requires increased complexity in the communication system. When designing a communication system, the engineer ordinarily seeks to maximize mutual information as much as possible at rates that don't exceed the budget for creating it, and the rate distortion curve provides the methods for achieving this.

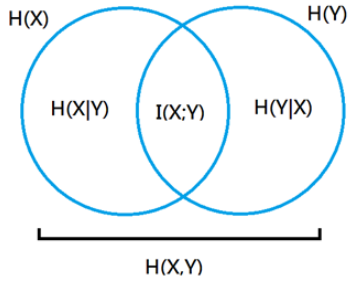

Figure 3.2.5: The Venn diagram demonstrates the relationship between entropy, conditional entropy, and average mutual information. It shows that the knowledge of *Y* by specification of *X* is the same as the knowledge of *X* by specification of *Y* [3, 4].

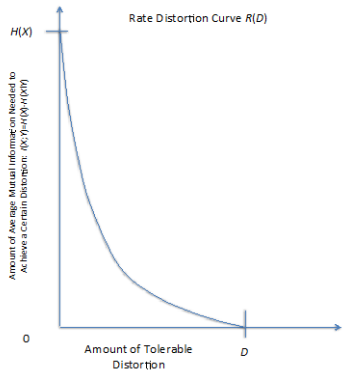

Figure 3.2.6: This graph shows the rate distortion curve, *R*(*D*). It provides the means for balancing the fidelity of a communication system with the complexity of creating it by providing a correlation between the amount of expected distortion and the average mutual information required to achieve it.

Suppose we have a discrete memoryless channel (d.m.c.) with an input alphabet A_M and an output alphabet A_N. The capacity of the channel can then be defined as [4]:

$$C = \max I(X;Y) = \max \sum_{j,k} P(j,k) \log\left(\frac{P(j,k)}{P(j)Q(k)}\right). \qquad (3.2.15)$$

The maximum is taken with regards to viable options of the input distribution [4]. $I(X; Y)$ is also defined by the following three constraints: $\sum_j P(j) = 1$ and $P(j) \geq 0,\ 0 \leq j \leq M-1$. Shannon's noisy coding theorem then states, "Let a d.m.c. have a capacity *C* and a discrete stationary source have entropy rate *H*. If $H \leq C$, where both are measured in nats per source letter, then the output of

the source can be encoded for transmission over the channel with an arbitrarily small frequency of error [4]." However, in practices often encountered the communication engineer must settle for some amount of distortion when building a device to transmit data and rate distortion theory must then be used.

A graph of the rate distortion curve is shown in figure 3.2.6; $R(D)$ is a decreasing function of D. The x-axis represents the upper limit of expected or tolerable distortion $E[d]=D$. The y-axis gives the minimum amount of average mutual information required to achieve this expected level of distortion. As previously stated in section 3.1, it's shown that the amount of average mutual information required to achieve no distortion is simply the entropy source $H(X)$, because $H(X|Y)=0$ in lossless data compression. A point on the $R(D)$ curve is represented as $(D, I(X;Y))$ or as (D_s, R_s), designating the distortion D and the bit-rate R respectively. The subscript s refers to the slope of the rate-distortion curve at the given point on the rate distortion function. The rate distortion curve can give two types of optimization problems given certain constraints. One possibility is minimizing the expected distortion given a certain bit-rate R. This corresponds to maximizing the fidelity of data transmission given a certain amount of resources to construct the communications system. The other possibility is minimizing the bit-rate R given a certain tolerable distortion D. This coincides to minimizing the complexity of the communication system as much as possible, yet still achieving a certain upper limit of tolerable distortion. In other words, using the least amount of resources and/or energy required to achieve a certain fidelity requirement. Two graphs are provided to help depict these concepts.

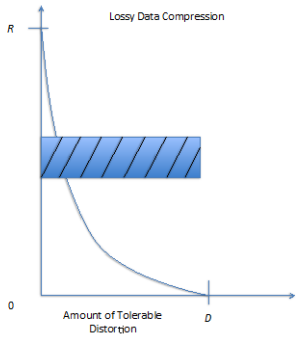

Figure 3.2.7: This graph depicts minimizing the distortion *D*, given a certain bit-rate *R*. This is analogous maximizing the fidelity of data transmission given a certain amount of resources to construct a communications system

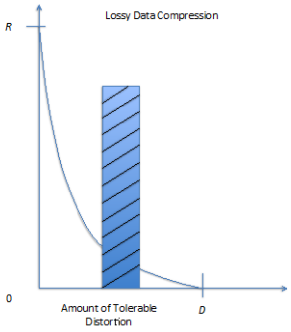

Figure 3.2.8: This graph depicts minimizing the bit-rate *R*, given a certain expected distortion *D*. This is analogous to using the least resources possible to achieve a certain fidelity criterion of data transmission.

3.3: Determining Distortion Functions

As stated previously in section 3.1, the rate distortion cure can be found if a source with all its statistical properties and a fidelity criterion are both given. For the source along with it's statistical properties we will use a discrete memoryless source (abbreviated d.m.s) that has input alphabet A_M and output alphabet A_N. A d.m.s. may be described by saying there exists a conditional probability $Q(k|j)$, where there's a probability that the letter *k* is produced at the output when the input letter is *j* for every ordered pair in the product space A_{MN} [4]. Discrete simply refers to both alphabets being discrete or having a determined number of letters. Memoryless means the channel operates so that it produces each new letter of an input word independently from each other. In other words, the output of the next letter isn't contingent upon the previous letter or any other letter that was generated before it [4]. In some situations a certain letter could create a higher or lower probability that the next successive letter will have a determined value, but this is not the case with memoryless sources. A d.m.s. is also stationary, or time independent. This means that the expected entropy $H(X)$ would be the same whenever the output was required to reproduce information received from the input, regardless of when or how many times it's required to reproduce this information. The letters that are produced at the output

are called independent identically distributed discrete r.v.'s (abbreviated i.i.d.) [4]. The d.m.s. described here is denoted as $\{X_t, P\}$. It can be represented mathematically as [4]:

$$P(\vec{x}) = \prod_{t=1}^{n} P(x_t). \tag{3.3.1}$$

A method is now needed to measure how distorted each source word is from the output of the d.m.s. $\{X_t, P\}$. In information theory, a word distortion measure assigns an estimation of the distortion value for every possible source word. "It's a non negative cost function, denoted by $\rho_n(\vec{x}, \vec{y})$, that specifies the penalty charged for reproducing the source word **x** by the vector **y** [4]." The word distortion measure for a d.m.s. is given as:

$$\rho_n(\vec{x}, \vec{y}) = \frac{1}{n}\sum_{t=1}^{n} \rho(x_t, y_t) \tag{3.3.2}$$

.

A succession of word distortion measures defined on the product space $[0, \infty)$ is called the fidelity criterion. It's represented as [4]:

$$F_\rho = \{\rho_n(\vec{x}, \vec{y}), 1 \le n < \infty\}. \tag{3.3.3}$$

The fidelity criterion for a d.m.s. is appropriately called a single letter fidelity criterion. A single letter fidelity criterion is nothing more than an arithmetic average of the distortion between the correlated letters at the input and the output. For example, suppose the source word is 7 at the input but the output produced 13. Using equation 3.3.2, n would be 1 greater than the difference between 13 and 7 because t=1. This would give:

$$\frac{1}{7}\sum_{t=1}^{n} \rho(7,13) = \frac{1}{7}(7+8+9+10+11+12+13) = \frac{1}{7}(70) = 10.$$

The arithmetic average of 7 and 13 is given by $\frac{7+13}{2}$=10, which is the same value obtained when using the word distortion measure from equation 3.3.2. This single letter fidelity criterion demonstrates how reproducing the number 77 at the output when the source word was 67 is more

serious than reproducing the number 66 at the output when the source word was 67. This is analogous to how reproducing wrong information in communication systems may have different levels of severity depending on what false information has been reproduced.

Now that the single-letter fidelity criterion F_ρ and the d.m.s. $\{X_t, P\}$ are both given, a corresponding joint distribution will be given for each probability assignment $Q(k|j)$ over the product space. The joint distribution is given by equation 3.2.9 and can be rearranged to yield $P(j,k) = P(j)Q(k|j)$ [4]. The single-letter fidelity criterion of equation 3.3.2 that produces F_ρ transforms into an r.v. over the joint ensemble whose expected value is contingent upon the selection of $Q(k|j)$. It's represented as [4]:

$$d(Q) = \sum_{j,k} P(j)Q(k|j)\rho(j,k). \tag{3.3.4}$$

Recall that the expectation value $E[f]$ in the example following equation 3.2.3 represented the average height of the population of Chicago given the sum of a probability distribution times the r.v. (the height of the population). Similarly, the expectation value of equation 3.3.4, represented by $d(Q)$, delineates the average distortion associated with Q given the sum of the conditional probability distribution and the r.v. that represents the distortion between j and k. "$Q(k|j)$ is said to be D-admissible if and only if the expectation value $d(Q)$ is less than or equal to D (if $d(Q) \leq D$) [4]." The group containing all D-admissible conditional probability assignments is defined as [4]:

$$Q_D = \{Q(k|j) : d(Q) \leq D\}. \tag{3.3.5}$$

For each conditional probability assignment it should be fairly intuitive that an average mutual information is present in addition to the average distortion. This average mutual information then gives rise to the rate distortion curve $R(D)$ for any fixed value of acceptable distortion D. The

average mutual information and rate distortion curve are given by equations 3.3.6 and 3.3.7 respectively [4]:

$$I(Q)=\sum_{j,k}P(j)Q(k|j)\log\left(\frac{Q(k|j)}{Q(k)}\right) \tag{3.3.6}$$

and

$$R(D)=\min_{Q\in Q_D} I(Q). \tag{3.3.7}$$

In this example we can say that $R(D)$ of $\{X_t, P\}$is given in compliance with F_ρ. In equation 3.3.7 it is the source and not the channel that's provided [4]. This is similar to the example provided by figure 3.2.8 where the designer of the communications system wishes to reduce the bit-rate (mutual information) as much as possible to achieve a given fidelity requirement. Also known as using the least amount of resources possible to achieve a prespecified level of *D*, which is why it's shown as a minimum and not a maximum. Often in rate distortion theory the designer wishes to maximize the mutual information as much as possible to increase accuracy. However, if the tolerable rate of distortion is specified then you want to use the least resources possible to achieve that fidelity requirement as in equation 3.3.7 [4].

There exist several other distortion functions that will generate varying values of expected distortion for the rate distortion curve. Two of the most common distortion functions used for continuous alphabets is the squared error distortion function and the hamming distortion function [3]:

$$d\left(x,\hat{x}\right)=\left(x-\hat{x}\right)^2 \tag{3.3.8}$$

and

$$d\left(x,\hat{x}\right)=\begin{cases} 0 & if \quad x=\hat{x} \\ 1 & if \quad x\neq\hat{x} \end{cases} \tag{3.3.9}$$

Because there's little information regarding more complicated fidelity criterion, most of the discourse surrounding them are addressed towards single-letter fidelity criteria [4]. Each

distortion function defines the goals of the communication system by quantifying how distorted the output is from the input; it's represented as $\rho(j,k)$ in equation 3.3.4. Depending on how the engineer wishes to transmit data will ultimately determine which type of distortion function is appropriate to use.

For any source, the rate distortion region is defined as the total of the set of achievable rate distortion pairs (*R*, *D*). The rate distortion function is defined as an infimum of rates *R* such that a point (*R*, *D*) is in the rate distortion region of the source for a given value of *D* [3]. This is what was calculated for equation 3.3.7. It's analogous to the example discussed in Figure 3.29. "The distortion rate function $D(R)$, is the infimum of all distortions *D* such that (*R*, *D*) is in the distortion region of the source for a given rate *R* [3]." The distortion rate function provides another way of analyzing the rate distortion region to find the boundary defined by rate distortion pairs; it can be considered equivalent to the rate distortion function [3]. With the distortion rate function you are provided with the source and the capacity of the channel over which it must be sent, so the problem is to define the least amount of distortion as in Figure 3.2.7. I thought it would be worth mentioning to demonstrate alternative real-life problems in constructing communication systems. However, information theory has evolved to primarily use rate distortion functions, not distortion rate functions [4]. For this reason, I choose only to derive various values of $R(D)$ even though the use of $D(R)$ should be mentioned. When examining rate-distortion theory in its biological context we will only need to use the rate-distortion function.

3.4: Lagrange Multipliers

Lagrange multipliers are generally used in math to find the maxima and minima of a function $f(x, y)$ subject to the constraint $g(x, y)$=c, where c is a constant [7]. The method of Lagrange multipliers can of course be expanded to three are more variables such that the critical points of $f(x_1, x_2, \ldots, x_n)$ can be found given the constraint $g(x_1, x_2, \ldots x_n)$=c. In figure 3.4.1,

you can find the extreme of $f(x, y)$ when the point (x, y) is limited to the level curve $g(x, y)=k$. It's represented by the bold line where the constant here is defined by k.

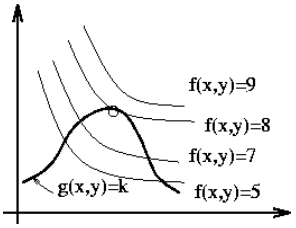

Figure 3.4.1: The graph shows how the extreme of $f(x, y)$ can be found when the point (x, y) is constrained to the level curve $g(x, y)=k$.

It is clear from the diagram that the extreme here is a maximum. In order to maximize the following function $f(x, y)$ given the constraint $g(x, y)$=k, you can simply find the largest k value where the level curve $f(x, y)$ intersects with the constraint $g(x, y)$=k. Due to the definition of a maximum, at this point where the two curves just touch, they share a common tangent line. This means their gradient vectors must be parallel, and this association yields the connection $\nabla f(x, y)=\lambda \nabla g(x, y)$. λ is a scalar quantity referred to as the "Lagrange multiplier." It serves as a dummy variable to help find the maximum and minimum values of $f(x, y)$ subject to the constraint $g(x, y)$=k. This is accomplished by finding all the values of x, y, and λ which satisfy the conditions $\nabla f(x, y)=\lambda \nabla g(x, y)$ and $g(x, y)$=k. The largest value found by evaluating $f(x, y)$ at these points is the maximum and the smallest is the minimum. If more than one constraint is present, then extra Lagrange multipliers should be added to help determine the values of the variables of the function.

Before showing how Lagrange multipliers are used to determine the rate distortion curve let's look at the following trivial example to reiterate how they're used prior to tackling a more difficult example. What are the extrema of the function $f(x, y)=3x-y+1$ given the constraint $g(x, y)=3x^2 + y^2=9$?

$\nabla f = \lambda \nabla g$ gives the following two relationships:

(1) $\frac{df}{dx}=\lambda\frac{dg}{dx}\Rightarrow 3=\lambda 6x\Rightarrow x=\frac{1}{2\lambda}$

and (2) $\frac{df}{dy}=\lambda\frac{dg}{dy}\Rightarrow -1=\lambda 2y\Rightarrow y=-\frac{1}{2\lambda}$

Now use these two values and plug them into the constraint to solve for λ. After attaining all possible values for λ, plug these values back into the equations derived in (1) and (2). The largest value found by evaluating $f(x, y)$ at these points is the maximum, and the smallest is the minimum, as stated above.

$3x^2 + y^2=9$

$3\left[\frac{1}{2\lambda}\right]^2+\left[-\frac{1}{2\lambda}\right]^2=9$

$\frac{3}{4\lambda^2}+\frac{1}{4\lambda^2}=9$

$\frac{1}{\lambda^2}=9$

$\lambda^2=\frac{1}{9}$

$\lambda=\pm\frac{1}{3}$ so we have

(1) $3=\lambda 6x\Rightarrow 3=\left(\frac{1}{3}\right)6x\Rightarrow x=\frac{3}{2}$ and $3=\lambda 6x\Rightarrow 3=\left(-\frac{1}{3}\right)6x\Rightarrow 3=-2x\Rightarrow x=-\frac{3}{2}$

(2) $-1=\lambda 2y\Rightarrow -1=\left(\frac{1}{3}\right)2y\Rightarrow y=-\frac{3}{2}$ and $-1=\lambda 2y\Rightarrow -1=\left(-\frac{1}{3}\right)2y\Rightarrow y=\frac{3}{2}$

Now evaluate $f(x, y)$ at these points

$3\left(\frac{3}{2}\right)-\left(-\frac{3}{2}\right)+1=\frac{9}{2}+\frac{3}{2}+\frac{2}{2}=\frac{14}{2}=7$ and

$3\left(-\frac{3}{2}\right)-\frac{3}{2}+1=-\frac{9}{2}-\frac{3}{2}+\frac{2}{2}=-\frac{10}{2}=-5$

Thus, there exists a maximum 7 at $\left(\frac{3}{2},-\frac{3}{2}\right)$ and a minimum -5 at $\left(-\frac{3}{2},\frac{3}{2}\right)$ for the function $f(x, y)=3x-y+1$ given the constraint $g(x, y)=3x^2 + y^2=9$. Now let's look at how Lagrange multipliers are used to help determine the rate distortion curve.

The problem posed in section 3.3 of finding the rate distortion function is that of minimizing the average mutual information denoted as [4]:

$$I(Q)=\sum_{j,k} P_j Q_{k|j} \log\left(\frac{Q_{k|j}}{Q_k}\right). \tag{3.4.1}$$

Equation 3.4.1 should be minimized as to choose a value of the conditional probability $Q_{k|j}$ obeying the following three constraints[1]:

1. $$Q_{k|j} \geq 0 \tag{3.4.2}$$

2. $$\sum_k Q_{k|j} = 1 \tag{3.4.3}$$

and 3. $$\sum_{j,k} P_j Q_{k|j} \rho_{j,k} = D. \tag{3.4.4}$$

The inequality constraint represented in equation 3.4.2 produces an obstacle in the methodical calculation of $R(D)$. For more information see Berger's Rate Distortion Theory [4], but for now if we disregard this constraint, then the problem becomes a direct computation using Lagrange multipliers. The expanded function may be solved as [4]:

$$J(Q)=I(Q)-\sum_j \mu_j \sum_k Q_{k|j} - s\sum_{j,k} P_j Q_{k|j} \rho_{j,k}. \tag{3.4.5}$$

In equation 3.4.5 μ_j and s have been chosen as the two Lagrange multipliers. With the requirement that

[1] In the majority of the literature regarding information theory probability distributions are denoted as P_j. Similarly, all other functions describing random variables or alphabet letters use subscripts for the j and k values. I have therefore introduced this notation here and continue to use it throughout the remainder of the paper.

$$\frac{dJ}{dQ_{k|j}} = 0. \tag{3.4.6}$$

We can solve the unconstrained problem in the following proof where we set $\log \lambda_j = \frac{\mu_j}{P_j}$ and set the indices *j* and *k* equal to *m* and *n* in **step 2** and **step 3**. Before analyzing the proof realize that the probability of *Q*(*k*) is dependent upon the conditional probability $Q_{k|j}$ by looking at the correlation between equations 3.2.9 and 3.2.7. These equations rearrange to give the following relation:

$$Q_k = \sum_j P_j Q_{k|j}. \tag{3.4.7}$$

This effectively means that the probability of receiving the letter *k* from the alphabet A_N is contingent upon the conditional probability of receiving the letter *k* at the output when the letter transmitted by the input is *j*, as intuition would suggest. The proof for computing the *R*(*D*) function is shown as follows [4].

Proof:

Step 1. $J(Q) = I(Q) - \sum_j \mu_j \sum_k Q_{k|j} - s \sum_{j,k} P_j Q_{k|j} \rho_{j,k}$

Step 2. $J(Q) = \sum_{j,k} P_j Q_{k|j} \log\left(\frac{Q_{k|j}}{Q_k}\right) - \sum_j \log(\lambda_j) P_j \sum_k Q_{k|j} - s \sum_{j,k} P_j Q_{k|j} \rho_{j,k}$

Step 3. $J(Q) = \sum_{m,n} P_m Q_{n|m} \log\left(\frac{Q_{n|m}}{Q_n}\right) - \sum_m \log(\lambda_m) P_m \sum_n Q_{n|m} - s \sum_{m,n} P_m Q_{n|m} \rho_{m,n}$

Step 4. $J(Q) = \sum_m P_m \sum_n Q_{n|m} \left[\log\left(\frac{Q_{n|m}}{Q_n}\right) - \log \lambda_m - s\rho_{m,n} \right]$

Step 5. Now define $\beta_{m,n}$ as $\beta_{m,n} = Q_{n|m}\left[\log\left(\frac{Q_{n|m}}{\lambda_m Q_n}\right) - s\rho_{m,n}\right]$ and take the derivative of $\frac{dJ}{dQ_{k|j}}$

and $\frac{dJ}{d\beta_{m,n}}$.

Step 6. $$\frac{dJ}{dQ_{k|j}} = \sum_j P_j \sum_k Q_{k|j}\log\left(Q_{k|j}\right) - \sum_j P_j \sum_k Q_{k|j}\log\left(\lambda_j \sum_j P_j Q_{k|j}\right) - \sum_j P_j \sum_k Q_{k|j} s\rho_{j,k}$$

Step 7. $$\frac{dJ}{dQ_{i|j}} = \frac{\partial}{\partial Q_{i|j}}\left\{\begin{array}{l} \sum_j P_j Q_{1|j}\log(Q_{1|j}) - \sum_j PQ_{1|j}\log\left(\lambda_j \sum_j P_j Q_{1|j}\right) - \sum_j P_j Q_{1|j} s\rho_{j,1} \\ +\sum_j P_j Q_{2|j}\log(Q_{2|j}) - \sum_j PQ_{2|j}\log\left(\lambda_j \sum_j P_j Q_{2|j}\right) - \sum_j P_j Q_{2|j} s\rho_{j,2} \\ +\sum_j P_j Q_{i|j}\log(Q_{i|j}) - \sum_j PQ_{i|j}\log\left(\lambda_j \sum_j P_j Q_{i|j}\right) - \sum_j P_j Q_{i|j} s\rho_{j,i} \end{array}\right\}$$

Step 8. $$\frac{dJ}{dQ_{i|j}} = \frac{\partial}{\partial Q_{i|j}}\left[\sum_j P_j Q_{i|j}\log\left(Q_{i|j}\right) - \sum_j P_j Q_{i|j}\log\left(\lambda_j \sum_j P_j Q_{i|j}\right) - \sum_j P_j Q_{i|j} s\rho_{j,i}\right]$$

Step 9.

$$\frac{dJ}{dQ_{k|j}} = \frac{\partial}{\partial Q_{i|j}}\left[\sum_j P_j Q_{i|j}\log\left(Q_{i|j}\right) - \left[\sum_j P_j Q_{i|j}\log\left(\lambda_j\right) + \sum_j P_j Q_{i|j}\log\left(\sum_j P_j Q_{i|j}\right)\right] - \sum_j PQ_{i|j} s\rho_{j,i}\right]$$

Step 10. Use product rule to differentiate the function:

$$\begin{aligned} \frac{dJ}{dQ_{i|j}} &= \sum_j P_j \frac{\partial}{\partial Q_{i|j}}\left[Q_{i|j}\right]\log\left(Q_{i|j}\right) + \sum_j P_j Q_{i|j}\frac{\partial}{\partial Q_{i|j}}\left[\log\left(Q_{i|j}\right)\right] - \sum_j P_j \frac{\partial}{\partial Q_{i|j}}\left[Q_{i|j}\right]\log\left(\lambda_j\right) \\ &- \sum_j P_j \frac{\partial}{\partial Q_{i|j}}\left[Q_{i|j}\right]\log\left(\sum_j P_j Q_{i|j}\right) - \sum_j P_j Q_{i|j}\frac{\partial}{\partial Q_{i|j}}\left[\log\left(\sum_j P_j Q_{i|j}\right)\right] - \sum_j P_j \frac{\partial}{\partial Q_{i|j}}\left[Q_{i|j}\right] s\rho_{j,i} \end{aligned}$$

Step 11.

$$\frac{dJ}{dQ_{i|j}} = \sum_j P_j \log\left(Q_{i|j}\right) + \sum_j P_j Q_{i|j}\left(\frac{1}{Q_{i|j}}\right) - \sum_j P_j \log\left(\lambda_j\right) - \sum_j P_j \log\left(\sum_j P_j Q_{i|j}\right) - \sum_j P_j Q_{i|j}\left(\frac{1}{\sum_j P_j Q_{i|j}}\right) - \sum_j P_j s\rho_{j,i}$$

Step 12. $$\frac{dJ}{dQ_{i|j}} = \sum_j P_j \left\{ \log\left(Q_{i|j}\right) - \left[\log\left(\lambda_j\right) + \log\left(\sum_j P_j Q_{i|j} \right) \right] + 1 - s\rho_{j,2} - \frac{Q_{i|j}}{\sum_j P_j Q_{i|j}} \right\}$$

Step 13. $$\frac{dJ}{dQ_{i|j}} = \sum_j P_j \left\{ \log\left(Q_{i|j}\right) - \log\left(\lambda_j \sum_j P_j Q_{i|j} \right) + 1 - s\rho_{j,i} - \frac{Q_{i|j}}{\sum_j P_j Q_{i|j}} \right\}$$

Step 14. $$\frac{dJ}{dQ_{k|j}} = \sum_j P_j \left[\log\left(\frac{Q_{k|j}}{\lambda_j \sum_j P_j Q_{k|j}} \right) - s\rho_{j,k} + 1 - \frac{Q_{k|j}}{\sum_j P_j Q_{k|j}} \right]$$

Step 15. $$\frac{dJ}{dQ_{k|j}} = \sum_j P_j \left[\log\left(\frac{Q_{k|j}}{\lambda_j Q_k} \right) - s\rho_{j,k} + 1 - \frac{Q_{k|j}}{Q_k} \right] = 0$$

,

Step 16. $$\frac{dJ}{d\beta_{m,n}} = \sum_m P_m \frac{\partial}{\partial \beta_{m,n}} \left(\beta_{m,n}\right)$$

Step 17.

$$\frac{d\beta_{m,n}}{dQ_{k|j}} = \frac{\frac{dJ}{dQ_{k|j}}}{\frac{dJ}{d\beta_{m,n}}} = \left(\frac{dJ}{dQ_{k|j}} \right) \left(\frac{d\beta_{m,n}}{dJ} \right) = \frac{\sum_j P_j \left[\log\left(\frac{Q_{k|j}}{\lambda_j \sum_j P_j Q_{k|j}} \right) - s\rho_{j,k} + 1 - \frac{Q_{k|j}}{\sum_j P_j Q_{k|j}} \right]}{\sum_m P_m} = 0$$

Step 18. $$\frac{d\beta_{m,n}}{dQ_{k|j}} = \begin{cases} 0 & if \quad (n \neq k) \\ -\frac{P_j Q_{k|m}}{Q_k} & if \quad (n = k, m \neq j) \\ \log\left(\frac{Q_{k|j}}{\lambda_j Q_k} \right) - s\rho_{j,k} + 1 - \frac{P_j Q_{k|j}}{Q_k} & if \quad (n = k, m = j) \end{cases}$$

Step 19. Combining step 6-17 with $\frac{dJ}{dQ_{k|j}} = 0$ we get: $\frac{dJ}{dQ_{k|j}} = \sum_j P_j \left[\log\left(\frac{Q_{k|j}}{\lambda_j Q_k} \right) - s\rho_{j,k} \right] = 0.$

Step 20. $\log\left(\frac{Q_{k|j}}{\lambda_j Q_k} \right) - s\rho_{j,k} = 0$

Step 21. $\log\left(\frac{Q_{k|j}}{\lambda_j Q_k} \right) = s\rho_{j,k}$

Step 22. $\frac{Q_{k|j}}{\lambda_j Q_k} = e^{s\rho_{j,k}}$

Step 23. $Q_{k|j} = \lambda_j Q_k e^{s\rho_{j,k}}$; this is called the stationary point.

Step 24. $\sum_k \left[Q_{k|j} = \lambda_j Q_k e^{s\rho_{j,k}} \right]$; using the second constraint

Step 25. $1 = \lambda_j Q_k e^{s\rho_{j,k}}$

Step 26. $\lambda_j = \left(Q_k e^{s\rho_{j,k}} \right)^{-1}$

Step 27. Combing step 22 with step 25 we get:

$$\left[Q_{k|j} = \frac{Q_k e^{s\rho_{j,k}}}{\sum_l Q_l e^{s\rho_{j,l}}} \right] \frac{\sum_j P_j}{Q_k} \Rightarrow \frac{\sum_j P_j Q_{k|j}}{Q_k} = \sum_j \frac{P_j Q_k e^{s\rho_{j,k}}}{Q_k \sum_l Q_l e^{s\rho_{j,l}}} \Rightarrow \frac{Q_k}{Q_k} = \sum_j \frac{P_j Q_k e^{s\rho_{j,k}}}{Q_k \sum_l Q_l e^{s\rho_{j,l}}}.$$

Step 28. $c_k \equiv \sum_j \frac{P_j e^{s\rho_{j,k}}}{\sum_l Q_l e^{s\rho_{j,l}}} = 1$

Step 29. $d(Q) = \sum_{j,k} P_j Q_{k|j} \rho_{j,k}$

Step 30. $d(Q) = \sum_{j,k} P_j \lambda_j Q_k e^{s\rho_{j,k}} \rho_{j,k} = D$

Step 31. $I(Q) = \sum_{j,k} P_j Q_{k|j} \log\left(\frac{Q_{k|j}}{Q_k} \right)$

Step 32. $I(Q)=\sum_{j,k}P_jQ_{k|j}\log\left(\frac{\lambda_jQ_ke^{s\rho_{j,k}}}{Q_k}\right)$

Step 33. $I(Q)=\sum_{j,k}P_jQ_{k|j}\left[\log\left(\lambda_j\right)+\log\left(e^{s\rho_{j,k}}\right)\right]$

Step 34. $I(Q)=\sum_{j,k}P_jQ_{k|j}\left[s\rho_{j,k}+\log\left(\lambda_j\right)\right]$

Step 35. $I(Q)=\sum_{j,k}P_jQ_{k|j}s\rho_{j,k}+\sum_{j,k}P_jQ_{k|j}\log\left(\lambda_j\right)$

Step 36. $I(Q)=\sum_{j,k}P_j\lambda_jQ_ke^{s\rho_{j,k}}s\rho_{j,k}+\sum_{j,k}P_jQ_{k|j}\log\left(\lambda_j\right)$

Step 37. $I(Q)=sD+\sum_jP_j\sum_kQ_{k|j}\log\left(\lambda_j\right)$, using constraint 2

Step 38. $I(Q)=sD+\sum_jP_j\log\left(\lambda_j\right)=R$

I will now label steps 29 and 37 as points on the rate distortion curve where *s* represents the slope of the curve at that point:

$$D=\sum_{j,k}\lambda_jP_jQ_ke^{s\rho_{j,k}}\rho_{j,k} \tag{3.4.8}$$

and

$$R=sD+\sum_jP_j\log\left(\lambda_j\right). \tag{3.4.9}$$

Equation 3.4.8 and 3.4.9 give the points that lie directly on the *R*(*D*) curve, the desired relation between the complexity of a communication system and the accuracy of its data transmission. The point that lies on this curve is represented as [4]:

$$\left(D_s,R_s\right). \tag{3.4.10}$$

This point was mentioned at the end of section 3.2. If you know the distortion and average mutual information of a given communication system, then you can measure the distance information point, denoted $\left(D,I\left(I;Y\right)\right)$, from a point on the *R*(*D*) curve defined by equation

3.4.10 [1]. If the distance information point lies above the $R(D)$ curve then it has higher mutual information than is necessary to achieve a given fidelity criterion, and if it lies below the curve it has less mutual information than is necessary. An optimal situation would be to choose a point directly on the $R(D)$ curve that shows the exact amount of mutual information necessary for given value of distortion. This is given by (D_s, R_s).

The channel capacity can be found from equation 3.4.9 when D_s=0. This corresponds to no distortion and the uncertainty of the generated output is given simply as the entropy $H(X)$.

To sum up, if you're given a probability distribution with all its statistical properties known and a fidelity criterion, then you can compute the distortion function as demonstrated in section 3.3. The distortion function is then treated as an r.v. in connection with the joint probability distribution. When the joint probability distribution and the distortion measure are summed over all values for j and k, it generates the average distortion $d(Q)$. For a fixed value of distortion D, the rate distortion curve can be generated by taking a minimum of the average mutual information (which can be found by summing the probability distributions with their entropies) when $d(Q) \leq D$ (its maximum allowed value). Using the method of Lagrange multipliers, one can then compute the rate distortion curve once they have the average distortion, it's maximal permissible value, a distortion measure, and a probability distribution. This will produce a set of points that give the minimum amount of mutual information needed to achieve a desired, pre-specified level of fidelity. It provides the bond between the complexity of the communication system and the accuracy of the data it transmits in lossy data compression. Also, it can demonstrate if a preexisting system has too much or too little mutual information to achieve a desired level of fidelity by examining if the information-distortion point is above or below $R(D)$. Finally if we set D=0 on the point of the rate distortion curve, we can obtain the capacity of the channel represented as the maximum limit at which information can be conveyed without any loss in accuracy.

3.5: Blahut-Aritmoto Algorithm

In 1972, Richard Blahut proposed an alternative way of computing channel capacity by defining mutual information as a maximum over an appropriate space, channel capacities as double maxima, and rate distortion functions as double minima [9]. This provides another method for determining the capacity of a channel by portraying a set a probability vectors on itself in way that in converges to a vector that yields the capacity of the channel.

Proof [9]:

Suppose a transition matrix *Q* is $n \times m$, then $J(p,Q,P) = \sum_j \sum_k p_j Q_{k|j} \log\left(\frac{P_{j|k}}{p_j}\right)$ for any $m \times n$ transition matrix P. For a fixed P the function $J(p,Q,P)$ is maximized by $p_j = \frac{e^{\sum_k Q_{k|j} \log(P_{j|k})}}{\sum_j e^{\sum_k Q_{k|j} \log(P_{j|k})}}$.

We'll let $P^*_{j|k} = \frac{p_j Q_{k|j}}{\sum_j p_j Q_{k|j}}$ and $q_k = \sum_j p_j Q_{k|j}$. $I(p,Q)$ is the mutual information between the channel input and the channel output.

Step 1. $I(p,Q) = \max_P \sum_j \sum_k p_j Q_{k|j} \log\left(\frac{P_{j,k}}{p_j Q_k}\right)$

Step 2. $I(p,Q) = \max_P \sum_j \sum_k p_j Q_{k|j} \log\left(\frac{P_{j|k} Q_k}{p_j Q_k}\right)$

Step 3. $I(p,Q) = \max_P \sum_j \sum_k p_j Q_{k|j} \log\left(\frac{P_{j|k}}{p_j}\right)$ also $I(p,Q) = \sum_j \sum_k q_k P^*_{j|k} \log\left(\frac{P^*_{j|k}}{p_j}\right)$

Step 4. $I(p,Q) - \sum_j \sum_k p_j Q_{k|j} \log\left(\frac{P_{j|k}}{p_j}\right) = \sum_j \sum_k q_k P^*_{j|k} \log\left(\frac{P^*_{j|k}}{P_{j|k}}\right)$

If for $P_{j|k}=0$ for some value of *k*, then $I(p,Q)=0$. If you ignore the constraint $p_j \geq 0$, you can set the derivative of p_j equal to 0 to maximize $J(p,Q,P)$ over *p* by using a Lagrange Multiplier [9].

Step 5. $\frac{\partial}{\partial p_j}\left[\sum_j \sum_k p_j Q_{k|j} \log\left(\frac{P_{j|k}}{p_j}\right)+\lambda(p_j-1)\right]$

Step 6. $\frac{\partial}{\partial p_j}\left\{\sum_k\left[p_j Q_{k|j}\log\left(P_{j|k}\right)-p_j Q_{k|j}\log\left(p_j\right)+\lambda p_j-\lambda\right]\right\}=0$

Step 7. $\frac{\partial}{\partial p_j}\left\{\sum_k\left[Q_{k|j}\log\left(P_{j|k}\right)-Q_{k|j}\log\left(p_j\right)-Q_{k|j}+\lambda\right]\right\}=0$ but $\sum_k Q_{k|i}=1$ so

Step 8. $\frac{\partial}{\partial p_j}=\sum_k Q_{k|j}\log\left(P_{j|k}\right)-\log\left(p_j\right)-1+\lambda=0$

Step 9. $\sum_k Q_{k|j}\log\left(P_{j|k}\right)-\log\left(p_j\right)-1+\lambda=0$

Step 10. $\log\left(p_j\right)=\sum_k Q_{k|j}\log\left(P_{j|k}\right)-1+\lambda$

Pick λ such that $\sum_j p_j$=1. Find p_j first then add $\sum_j p_j$ to see what the role of λ is in the sum.

Step 11. $p_j=e^{\sum_k Q_{k|j}\log\left(P_{j|k}\right)-1+\lambda}$

Step 12. $\sum_j p_j=\sum_j e^{\sum_k Q_{k|j}\log\left(P_{j|k}\right)-1+\lambda}$

Step 13. $e^{\lambda-1}=\frac{1}{\sum_j e^{\sum_k Q_{k|j}\log\left(P_{j|k}\right)}}$ Now Return to **Step 10.**

Step 14. $\log\left(p_j\right)=\sum_k Q_{k|j}\log\left(P_{j|k}\right)-1+\lambda$

Step 15. $p_j = e^{\sum_k Q_{k|j} \log\left(P_{j|k}\right)} e^{\lambda - 1}$ Now plug **step 13** into **step 15**.

Step 16. $$p_j = \frac{e^{\sum_k Q_{k|j} \log\left(P_{j|k}\right)}}{\sum_j e^{\sum_k Q_{k|j} \log\left(P_{j|k}\right)}}$$

At this point p_j represents the capacity of the channel, because the capacity is defined when the function $J\left(p, Q, P\right)$ is maximized over p (what the proof above demonstrated.) The solution to the problem is [9]:

$$p_j = \frac{e^{\sum_k Q_{k|j} \log\left(P_{j|k}\right)}}{\sum_j e^{\sum_k Q_{k|j} \log\left(P_{j|k}\right)}}. \tag{3.5.1}$$

Using the Blahut algorithm over the methods described in section 3.4 allows for greater flexibility. It expedites the rate of calculating the capacity of the communications system. Capacity can be represented as the entropy $H(X)$ when D=0 on the $R(D)$ curve. However, the capacity is only part of the $R(D)$ curve and usually a communications engineer wishes to pick a point that allows some amount of distortion due to limited amounts of resources to construct it. A similar proof can be shown to generate a sequence of points (I,D) that converges to point on the $R(D)$ curve [9]. This will allow one to create a simulation that can quickly and easily compute the rate distortion curve using statistical software in place of the rigorous calculations discussed in section 3.4. Due to the length of the proof I have omitted it here but the results of it, a variant of the Blahut algorithm, will be demonstrated in chapter 4. It essentially uses the exact same methods and math as the proof demonstrated above. The Blahut-Aritmoto algorithm has two components defined as:

$$P_{k|j} = \frac{p_k e^{-\lambda d \rho_{j,k}}}{\sum_k p_k e^{-\lambda d \rho_{j,k}}} \tag{3.5.2}$$

and $$p_k = \sum_j p_j p_{k|j} \quad . \tag{3.5.3}$$

There are four steps for using the algorithm. First initialize the output probability distribution p_k as a uniform distribution. Second compute the conditional probability distribution $p_{k|j}$ that minimizes the mutual information $I(X;Y)$ subject to the distortion constraint in equation 3.5.2. Given the conditional probability distribution obtained, compute the marginal distribution p_k that minimizes the mutual information subject to the distortion constraint in equation 3.5.3. Finally, repeat steps 2 and 3 until p_k and $p_{k|j}$. The expected distortion $E[d]=D$ is determined by the choice of the Lagrange multiplier λ. By choosing different values of λ you obtain several conditional probabilities $p_{k|j}$, for which the distortion-information point $(D,I(X;Y))$ lies on $R(D)$ curve [1].

Most simulations that use information theory as its method for quantifying cellular decision-making strategies use the Blahut algorithm in programs such as Matlab for the quickest and most efficient computation of the rate distortion curve. The importance of this algorithm will be better illustrated in the next chapter.

Chapter 4

Applications of Rate Distortion Theory in a Biological Context

Everything at the cellular level has a certain degree of randomness, so cellular processes and cellular decision-making strategies can only be defined by probabilistic functions. Cells receive stochastic signals, they identify signals and execute decisions with stochastic biochemistry, and they grow and die in stochastic environments [10]. Cells execute binary (all-or-nothing) decisions based on their potential to evaluate information from their environment. Because the survival, reproduction, and evolutionary stability of these cells rely on correct decision-making, it's pertinent for the cells to employ the correct decision-making strategies [1]. However, the stochastic nature of the cells environment hinders the ability to correctly sense signals and execute the appropriate response.

A way to analyze and quantify how cells make correct assessments of their environment based on noisy, uncertain observations is needed to understand binary decision-making systems. "How close to optimal is a decision-making strategy in a given situation? [1]" What decision-making strategy would be the most favorable for a certain situation? How do cells choose when higher metabolic costs associated with more intricate sensing and signal transduction is advantageous to achieve higher fidelity decision-making? Is it possible to simulate a biological system that executes an ideal decision? [1]

Using the major aspects of rate distortion theory, it's possible to quantify how the cost and performance are correlated by providing a bond between its mutual information (metabolic cost) and expected distortion (performance). Under the framework of rate distortion theory it is often undesirable to have perfect transmission in a communications system due to increased costs. Similarly, it's impossible for cells to always execute "correct decisions," so rate distortion theory provides a method to demonstrate how these decisions can be executed imperfectly in the most efficient way possible [1]. By regarding a cellular decision-making system as a noisy communication systems with the environmental stimulus as the input and the decision as the output, it's possible to demonstrate that the ideal decision-making strategy depends on the cells prior knowledge of its environment, its goals for the decision, and how much energy it is wiling to provide for a correct decision [1].

The rate distortion framework for cellular decision-making supplies a harmonious connection with the three axioms of cellular decision-making as proposed by Perkins and Swain: (1) a cell must infer the state or likely future state of the environment by sensing stochastic stimuli; (2) based on the stimuli sensed it weighs the advantages and disadvantages of each potential decision; and (3) it executes a decision so that it maximizes the fitness of the cellular population [1, 10]. Decision theory can demonstrate how cells execute a decision based on stochastic signals from their environment by evaluating the costs and benefits of each potential response. Evolutionary theory takes into account situations where cellular-decisions are made in conjunction with other cells to increase the variance in fitness of the population over the variance in fitness of any singular cell [10]. These two theories along with the three-step process of cellular decision-making provide complementary perspectives in explaining observable cellular characteristics such as bet-hedging strategies, hysteresis, and irreversibility as being optimal under the framework of rate distortion theory [1].

4.1: Applications of Rate Distortion Theory in a Biological Context

Everything at the cellular level has a certain degree of randomness, so cellular processes and cellular decision-making strategies can only be defined by probabilistic functions. Signal transduction, diffusion, chemotaxing, gene expression, and mating are all stochastic processes [1, 10]. Cells grow and die in stochastic environments, and they make decisions based on stochastic biochemistry from stochastic signals received by their surroundings [10]. Because of the probabilistic nature of cellular decisions and their surroundings, aspects of rate distortion theory may be used to help demonstrate how cells go about deducing their actions based from their stochastic environments [10].

The stochastic interactions between cells and their environment provides a deterrent to correctly sense and interpret the signals that cells receive to make educated decisions based on the environment they're in [10]. However, the implications of correct decision-making are paramount, because the survival of the individual or the population depends on correct decision-making [1]. Using rate distortion theory to quantify cellular decision-making systems poses and additional question analogous to the one mentioned in section 3.1, "How do cells balance the metabolic cost of complex decision-making equipment with the benefit of accurate decisions? [1]." Under this bodywork, error-free transmission in cells is usually disadvantageous because it requires higher mutual information, which corresponds to more intricate biological mechanisms for sensing and signal transduction. This increases the complexity of the biological system and requires the cell to use more energy to achieve higher fidelity decision-making. Rate distortion theory helps to quantify the expense with which higher or lower rates of mutual information can increase or decrease the accuracy of cellular decision-making [1].

Perkins and Swain argue that cellular decision-making occurs in three steps. First, cells must deduce the most likely environmental state and possibly future states based on noisy, uncertain signals they receive. Second, they must evaluate the advantages and disadvantages of each potential response to the most probable state is has inferred. Third, the cell should execute

the appropriate response using a strategy that maximizes the variance in fitness for the cell and for the cellular population so that it can outcompete rivals and endure environmental cataclysms [10].

Furthermore, cells face an intrinsic problem when making decisions based on the environmental cues. Their biochemical decision-making apparatuses are intracellular, but their decisions made are determined from extracellular stimuli such as pheromone concentrations or a predator of the cell [10]. Signals are transmitted to the inside of the cell after being detected on the cell membrane, but these signals are stochastic and can never depict a perfectly accurate notion of the extracellular environment to the intracellular machinery. These signals usually fluctuate to various degrees and are generally accompanied by many other fluctuating signals that often conflict with one another. The rate of diffusion into and out of the cell is also stochastic, so the signals received inside may not provide a perfectly accurate depiction of the outer environment. Additionally, the components of the biological decision-making system are stochastic because they change in concentration and intracellular location [10]. When trying to detect certain stimuli such as sugar, some of the internal organelles of the cell consume sugar for energy. In such examples, the signals detected by the intracellular machinery to exhibit transcriptional factors to metabolize sugar may be subject to even further stochasticity. In response, cells adopt several strategies to explicate and utilize these noisy signals gathered from their extracellular environment such as statistical inference. Cells may reason which future state is most probable by examining a measurable variable that's correlated with an immeasurable variable of interest.

By using statistical inference a cell may infer a likely future environmental state by measuring signals that are associated with that state, because knowing the state is much more important than knowing the parameters that epitomize whichever state it's in [10]. For example, a temperature rise may be indicative of a bacterium entering a host organism or being in the presence of the sun [10]. Each situation would require a different physiological response from

the cell, so only knowing the temperature change isn't sufficient. If *E. coli* enters the digestive tract of a host organism, then it should express the operons necessary to metabolize lactose and maltose. If it expresses these operons in an environment where they're not needed or if it expresses more than necessary, then it has wasted cellular energy that could have been used for alternative means. By using the least amount of energy possible to achieve a certain goal a cell promotes evolutionary stability. Similarly, not expressing operons when they're needed can result in detrimental consequences to the cell. It's better for the cell to recognize what caused the change in environmental parameters then noting just the parameters themselves so it may respond accordingly. The concept of statistical inference is best demonstrated using Baye's rule, where cells deduce the most probable state of the environment by sensing signals that are correlated with that state [10]. Baye's rule states that cells can infer the possibility of an environmental state *E*, based on signals they sense *S* as:

$$P(E|S) = \frac{P(S|E)P(E)}{P(S)}. \tag{4.1.1}$$

The above equation implies that cells can guess the state of their environment from signals that are only correlated with the state [10]. This equation also implies that three forms of "prior knowledge" are available to the cell. First it assumes a certain amount of knowledge about environments *E* and their respective probabilities $P(E)$. Second, it implies a conditional probability of sensing certain signals in certain environmental states $P(S|E)$. The last term $P(S)$ assumes the probability of sensing a signal for all possible states of the environment [10]. The proof for Baye's rule is short and can be found using equations (3.2.8) and (3.2.9):

Step 1. $P(j|k) = \frac{P(j,k)}{P(k)}$

Step 2. $P(j,k) = P(k|j)P(j)$

Step 3. $P(j|k) = \frac{P(k|j)P(j)}{P(k)}$.

To demonstrate how Baye's Rule works suppose the weatherman predicts a 50% chance of a cold front coming through and dropping the temperature by 40 degrees. You've been in the shopping mall all day and are about to leave, but you don't know rather the temperature has dropped or not, because the weather was still warm when you arrived at the shopping mall. Knowing nothing more, you're equally likely to believe that's it's either cold or not cold so, $P(j = Cold) = P(j = NotCold) = 0.5$. Despite not seeing the outside conditions, you can still predict the weather outside by examining the clothes that other people are wearing; denote this observation as *k*. We say *k=Many Clothes* if they have on winter attire such as long pants and heavy coats and *k=Few Clothes* if they are dressed for warmer weather. Thus, you can fairly deduce if it's cold by observing *k*. If it is cold, then some people will have on winter attire so $P(k = ManyClothes|j = Cold) = 1$ and $P(k = FewClothes|j = Cold) = 0$. However, if it's not cold outside, perhaps some people will still wear winter attire in preparation for the weatherman's predictions but the probability is lower. We have $P(k = ManyClothes|j = NotCold) = 0.3$ and $P(k = FewClothes|j = NotCold) = 0.7$. Now imagine if you do see people with winter attire, then Baye's rule will allow you to quantify the probability of your belief that it's cold outside by computing the following equation:

$$P(j = Cold|k = ManyClothes) = \frac{P(k = ManyClothes|j = Cold)P(j = Cold)}{P(k = ManyClothes)}.$$

$P(k = ManyClothes)$ can then be calculated as follows:

$$P(k = ManyClothes) = P(k = ManyClothes, j = Cold) + P(k = ManyClothes, j = NotCold)$$

$$P(k = ManyClothes) = P(k = ManyClothes|j = Cold)P(j = Cold) + P(k = ManyClothes|j = NotCold)P(j = NotCold)$$

$$P(k = ManyClothes) = (1)(0.5) + (0.3)(0.5)$$

$$P(k = ManyClothes) = 0.5 + 0.15$$

$$P(k = ManyClothes) = 0.65$$

Once we know $P(k = ManyClothes)$ we can find the answer to the original question:

$$P(j = Cold | k = ManyClothes) = \frac{(1)(0.5)}{0.65} = \frac{0.5}{0.65} = \frac{10}{13} \approx 77\%.$$

So according to Baye's Rule in our example, the probability that's it's cold outside given the observation that some people are wearing winter attire is about 77%.

A question was posed that asked if it's plausible to construct a model that describes the likelihood of an extracellular environment based on noisy concentrations of signals received inside the cell, in particular the concentration of sugar levels [10, 11]. They posed a bacterium subject to two states: one that's high in extracellular sugar concentration and another that's low in extracellular sugar concentration. Sugar generally enters the cell and interacts with transcriptional factors, which communicates to the cell to express the appropriate genes [10]. Therefore the level of intracellular sugar that the cell has the ability to directly measure is regarded as the signal, and the environment is the extracellular surroundings, which is either high or low in sugar [10, 11]. Using this assumption it's possible to infer the posterior probability of the extracellular state being either high or low in sugar by examining the concentration of intracellular sugar the cell senses. For example, the likelihood of the environment being high in sugar based on the internal signals sensed, $P(high|S)$ can be demonstrated by using an equation similar to 4.1.1 [10]:

$$P(high|S) = \frac{P(S|high)P(high)}{P(S)}. \qquad (4.1.2)$$

The denominator $P(S)$ represents the probability of sensing a signal for all possible states correlated with an environment. In this example there's only the two states of "high" and "low," so it may be expanded as [10, 11]:

$$P(high|S) = \frac{P(S|high)P(high)}{P(S|high)P(high)+P(S|low)P(low)}. \qquad (4.1.3)$$

If any signal is continuous and the environment has only two possible states, then equation 4.1.1 may be used to compute the posterior probability of either one of these two states [10]. The output of any cellular decision-making strategy choosing a decision in response to a stimulus may be proportional to the posterior probability of a cell inferring what environmental state it's in. Thus, the execution of many cellular decisions is contingent upon what environmental state the cell has inferred based on the intracellular signals it has sensed [10].

However, you must keep in mind that these signals are still stochastic because of alterations of sugar being transported across the cell membrane, the consumption of sugar inside the cell to provide energy etc. [10]. The cell can only infer a probability that the extracellular environment is in a certain state given intracellular levels of sugar concentration as demonstrated by Baye's Rule. The cell can never be certain of their surrounding environment. This model merely provides a framework using concepts of rate distortion theory to help quantify how a cell determines the quantity of an unknown variable, by measuring the quantity of a known variable that's associated with the unknown variable of interest.

The cells can also improve their inference overtime by noting the facets of the fluctuating stimuli. To elaborate on the example above, if a cell infers that the environmental state is high in sugar, then the probability of a future state being high in sugar would be more likely, at least for a certain amount of time [10]. Therefore $P(high)$ would increase and $P(low)$ would consequently decrease. In effect, the cell executes cellular decisions by improving its knowledge of the environment through sequential implementation of Baye's rule-the present posterior probability of the environmental state would then become the prior probability of the environmental state in the next step [10]. This will affect how the cell infers the probability of a certain state overtime and consequently affect its decision to express certain genes or to execute decisions that are correlated with whatever state it infers.

Overtime the cells should adopt the resources needed to account for other possible environmental states it could encounter [10]. Using Baye's rule only demonstrates how step 1 is solved in Perkins and Swain's 3-step process for cellular decision-making systems. It only clarifies how cells deduce the most likely current and future state of their environment. Steps 2 and 3 also need to be described-the costs and benefits of each potential decision based on the most likely state and execution of this decision in the presence of other competitive decision-making cells. The use of mutual information and rate distortion functions provide an ideal way for describing how cells accomplish steps 2 and 3 of Perkins and Swains 3-step process in addition to step 1. The rate distortion curve then provides a bond between the performance of cells to properly execute the correct decision and the amount of energy that's related to higher fidelity decision-making. The use of $R(D)$ in a biological context will effectively answer the questions posed in the introduction of this chapter much like it answered the analogous questions with respect to communication engineering systems in chapter 3.

4.2: The use of Mutual Information, Distortion Functions, and Rate Distortion Curves as a Method for Cells to Weigh the Advantages and Disadvantages of Potential Decisions

Once a cell deduces the most probable state of its environment, it must evaluate the advantages and disadvantages of each prospective response by weighing the costs and benefits of each response. The most likely environmental state as well as the most likely future states must be analyzed in order to select both the most advantageous response and the level at which to respond [10]. The costs and benefits of each response may be difficult to pinpoint in cellular systems, but the most common example of a cost is the energy needed to synthesize the RNA and proteins used for gene expression. The benefit will rely on the characteristics of the proteins and the environment of the cell. It's usually the increase in growth rate obtained by metabolizing substances such as sugar, even though it must synthesize enzymes to metabolize the substance(s) of interest [10]. Fitness is then defined by Perkins and Swain to be the expected benefits of a response minus the expected costs [10].

Because the physiology of prokaryotes has been determined empirically to optimize the fastest growth rate that's possible, cellular growth rate has been regarded as a suitable measure of fitness for both eukaryotes and prokaryotes [10]. Since cellular growth rate is an experimentally measurable quantity, it has been demonstrated empirically when and at what level a cell should express a set of genes. The effects of cellular growth rate in the bacterium *E. coli* were measured by the expression of the *lac* operon in varying levels of extracellular concentrations of lactose [10, 12]. The *lac* operon is responsible for producing enzymes that metabolize lactose; these enzymes will be represented as Z to quantify their intracellular concentration. The experiment consisted of a control group that did not express the *lac* operon and experimental groups that expressed the *lac* operon to varying extents. All groups were placed in an environment that contained no lactose and the reduction in growth rate of the experimental groups due to the unnecessary gene expression was compared to the control group that didn't express the genes [10, 12]. It was determined that the reduction in growth rate observed in the experimental groups increased super-linearly by the amount of enzymes produced. This can be demonstrated as [10, 12]:

$$g_{low} = g_c - c(Z). \tag{4.2.1}$$

In the above equation g_{low} represents the diminished growth rate of the experimental population, and g_c is the growth rate of the control population not expressing the *lac* operon. The cost of expression of the enzyme $c(Z)$, can be given as [10, 12]:

$$c(Z) = Z_0 Z + Z_0' Z^2, \tag{4.2.2}$$

where Z_o and Z_0' are constants that were determined from observations. This formula shows that the cost of expression for the enzymes necessary to metabolize lactose is a quadratic function of the synthesized enzyme [10, 12]. However, at times it's beneficial for a bacterium to express certain enzymes such as those situations where the extracellular environment does contain lactose. As stated previously it's important for the cell to decide when to express genes as well as

at what level to express them. It has been determined that any additional energy acquired through metabolizing a substance such a lactose may be regarded as its benefit in response to express these genes for the synthesize of enzymes. Even though it costs some amount of energy from RNA and protein synthesize to express these genes, the benefit in growth rate can often exceed the cost in expression [10]. It's important for the cell to know how much gene expression is necessary for an expected environment so that it can maximize its growth rate in all situations.

The way in which a cell decides how much energy to use for gene expression or other vital actions can be modeled under the framework of rate distortion theory. The mutual information between the environmental stimulus and the cells decisions provides a way of quantifying the energy costs associated with cellular decision-making systems. The distortion function provides the expected distortion for each given amount of mutual information when graphed on the $R(D)$ curve. This distortion function can alternatively be considered the cost of performance as defined by Perkins and Swain, because the accuracy of a decision may be used to quantify the cost. If a cell is expected to achieve a certain amount of distortion at a given level of mutual information, then the mutual information represents the energy used to express genes or execute certain actions to achieve this level of distortion [10]. The distortion function then quantifies the quality of the decision by showing how many "false decisions" should be expected for this level of mutual information [1]. Under this mindset, the cost in fitness occurs when the cell makes an incorrect decision, so the lower levels of mutual information would correspond to higher costs in fitness when the correct decision is very important for the cells survival. Another way of thinking about this is by saying it's the cost in fitness for not expressing the genes when they're needed instead of the cost in fitness for expressing the genes. In this way of thinking, if the cost of not expressing the genes is more than the benefit of however many genes of interest are being expressed, then the fitness as defined by Perkins and Swain, will have a negative value. This negative value implies that more genes would need to be expressed in order to obtain a positive desired level of fitness, which is to be maximized.

Andrews, Porter, and Iglesias did a study in which a cellular decision-making system was considered as a noisy communications channel, and the goals defined by the distortion function can demonstrate how cells weigh the costs and benefits of different responses. The output was regarded as a decision based on the stimulus the cell sensed, and this stimulus was considered as the input [1]. In other words, the cell would make a decision Y based on the stimulus X that it has recognized. The decision-making system was modeled as a conditional probability where the cell chooses the decision $Y=y$ when the stimulus is $X=x$. They choose the stimulus to be the level of pheromone concentration in their simulations; however, one can define a stimulus to be any measurable signal of a cells environment when trying to use rate distortion theory to model decision-making strategies.

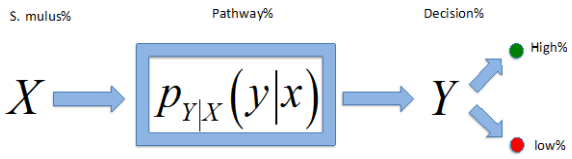

Figure 4.2.1: This depicts the conditional probability that a cell makes the decision $Y=y$ when the stimulus is $X=x$ [1].

The decision to mate or not to mate is a binary decision represented as y=high and y=low respectively. They used a variant of the Hamming distortion function similar to Equation 3.3.9 to quantify the quality of the cellular decisions. The distortion function characterizes the goals of the decision making pathway by quantifying how "distorted" a decision y is in response to a stimulus x; it essentially depicts how disadvantaged a decision is [1]. In their generic model, if the level of pheromone concentration is at or below the threshold level x_{th}, then the cell should not mate, and it should mate if the pheromone concentration is above this threshold value. The hamming distortion function states that the decision to mate has zero distortion when the pheromone concentration is above the threshold value, and it carries one arbitrary unit of

distortion if the cell chooses to mate at pheromone levels at or below x_{th}. The opposite is true for the decision not to mate y=low. The distortion function is represented as [1]:

$$d(x,y)=\begin{cases} & x \le x_{th} & x > x_{th} \\ y=low & 0 & 1 \\ y=high & 1 & 0 \end{cases}. \qquad (4.2.3)$$

The distribution of stimulus $p_X(x)$ was assumed to be exponential with finite support [1]. Stated alternatively, it was assumed that there was more likely to be a lower level of pheromone concentration with a probability of zero to find the concentration above a certain level; this stimulus distribution is referred to as likely low. The stimulus distribution ultimately affects the cell's decision-making strategy.

Once the stimulus distribution $p_X(x)$ and distortion function $d(x,y)$ are both known, it's possible to compute the rate distortion curve $R(D)$. The simplest way to compute the rate distortion curve is by using a Blahut-Aritmoto algorithm in 4 steps using the following two distortion constraints [1]:

$$p_{Y|X}(y|x)=\frac{p_Y(y)e^{-\lambda d(x,y)}}{\sum_y p_Y(y)e^{-\lambda d(x,y)}}. \qquad (4.2.4)$$

$$p_Y(y)=\sum_x p_X(x)p_{Y|X}(y|x). \qquad (4.2.5)$$

First set up the marginal probability distribution $p_Y(y)$ as a uniform distribution. Once this distribution has been specified, use it to calculate the conditional probability distribution $p_{Y|X}(y|x)$ that minimizes the mutual information while satisfying the distortion constraint in equation 4.2.4. Third, use the conditional probability distribution obtained to calculate the marginal probability distribution that minimizes the mutual information subject to the distortion constraint in equation 4.2.5. Finally, repeat the second and third steps until $p_{Y|X}(y|x)$ and

$p_Y(y)$ converge [1]. While these steps are being performed, the limiting mutual information has been determined to be the rate distortion curve $R(D)$. The rate distortion curve produces a slew of points that show how much mutual information between the stimulus and decision is needed to execute a decision with expected distortion $E[d]=D$; this bond can be referred to as the "information rate" that's required for a pre-specified level of D. The expected distortion is established from the choice of the Lagrange multiplier λ. The rate distortion curve forms a bond between the performance (expected distortion) and cost (mutual information) [1].

Once the rate distortion curve has been generated, you can assess the optimality of any cellular decision-making strategy by observing how closely the distortion-information point $\left(D, I\left(X;Y\right)\right)$ is to the rate distortion curve $R(D)$. Once a cellular decision-making strategy has been modeled as the conditional probability distribution $p_{Y|X}\left(y|x\right)$ and the probability distribution of the stimulus $p_X(x)$ has been determined, it's possible to calculate the mutual information $I(X;Y)$ and expected distortion D [1]. Any cellular decision-making strategy can then be assessed with regards to the rate distortion curve.

The mutual information $I(X;Y)$ can be used to quantify the cost of the cellular decision-making strategy, similar to how it was used to quantify the cost of constructing communication engineering systems in chapter 3. The mutual information expresses the reduction of uncertainty in decision Y when the stimulus is X [1]. A noisy decision-making pathway will have lower levels of mutual information, which corresponds to lower levels of fidelity when making decisions. As mutual information increases so does the probability for a cell to execute the correct decision Y given the stimulus X. It can be shown that some situations are more beneficial for the cell to expend the extra energy required for higher fidelity decision-making and others where it's not. The mutual information can be found relatively easy once the stimulus distribution has been assumed and the cellular decision-making strategy has been modeled as the

conditional probability distribution $p_{Y|X}(y|x)$. The entropy $H(Y)$, of any binary decision Y, expresses the uncertainty in the decision [1]:

$$H(Y) = -\sum_{y} p_Y(y) \log_2 p_Y(y). \tag{4.2.6}$$

Equation 4.2.6 essentially represents the maximum amount of mutual information that a cell may achieve; i.e. $I(X;Y) = H(Y) - H(Y|0) = H(Y)$. In reality cells never make completely accurate decisions. However, some cells do make decisions with much less distortion than others depending on how important the decision the cell makes is with regards to the fitness of the cell and its population at large. The mutual information between the decision and the stimulus can then be expressed as:

$$I(X;Y) = H(Y) - H(Y|X) = \sum_{x}\sum_{y} p_{Y|X}(y|x) p_X(x) \log_2 \left(\frac{p_{Y|X}(y|x)}{p_Y(y)} \right). \tag{4.2.7}$$

The reduction of uncertainty in Y given X can be stated as the uncertainty in decision Y minus the uncertainty that still remains in decision Y after the stimulus X has been identified [1]. Equation 4.2.7 shows how this statement is represented mathematically. It characterizes the cellular resources and energy that's required for more accurate decision-making systems.

The expected distortion D may be computed using a formula similar to equation 3.3.4:

$$E[d] = \sum_{x,y} p_X(x) p_{Y|X}(y|x) d(x,y). \tag{4.2.8}$$

Once the distortion-information point $(D, I(X;Y))$ has been determined, you can evaluate the optimality of the decision-making strategy by observing how closely this point approaches the rate distortion curve [1]. If the distortion-information point lies above the $R(D)$ curve, then it has more mutual information than is necessary to achieve the goal of obtaining a minimal amount of D, as defined by the distortion function. If a cell possesses more mutual information than is

necessary, then it's wasting cellular energy to obtain the same amount of distortion possible to achieve with less mutual information. Such strategies would be considered suboptimal, because they will decrease the fitness of the cell. That is, the cost of having more mutual information than is needed for a certain value of D is more than its benefit. This may be analogous of a cell creating more sites for signal transduction at the cell surface to detect spatial heterogeneities than is needed for a certain chemotaxing accuracy [2].

It's possible to use the rate distortion framework to design optimal decision-making strategies as well. To design an optimal strategy you can simply pick any point along the rate distortion curve. These points provide the exact amount of mutual information to achieve an expected level of distortion; i.e. an optimal strategy is one that doesn't use more mutual information than is necessary to achieve a certain level of distortion. Every point of the rate distortion curve may be represented as an ideal conditional probability $p_{Y|X}(y|x)$ of a cell making the decision $Y=y$ when the stimulus is $X=x$. Each point along this curve will have a different level of expected distortion and the exact amount of mutual information that's required to achieve that level of distortion. Ideally you would want decision-making strategies that only fall on the $R(D)$ curve so that the cell doesn't waste energy, regardless of how much distortion the cell can afford in its decision-making. These examples demonstrate how a cells optimal decision-making strategy depend on its goals for the decision and how much metabolic cost its inclined to "pay" for an accurate decision.

The distortion function shown by equation 4.2.3 penalizes decisions near the threshold x_{th} equally, but for real-life situations intuition suggests that it should be more probable to observe more incorrect decision near this threshold. This threshold may not be entirely evident to the cell, it could change in time with a changing stimulus distribution or the cell may be excused for making an incorrect decision close to the threshold [1]. In real life scenarios it may be more practical for cells to use less energy that's required to make an accurate decision when the level

of pheromone concentration that regulates when cells mate or not is closer to the threshold value that compels them to do so. To demonstrate, you would want all cells to mate when the pheromone concentration was far above the threshold. If the majority of cells were observed to not mate at high pheromone concentrations, it could lead to problems in the reproductive success of the cellular population. On the other hand, if the pheromone concentration were very near the threshold value, then observing several cells not mating certainly wouldn't be indicative of a possible breeding debacle. In fact in many cellular populations, one should expect to see such occurrences in cells choosing to mate or not to mate when the pheromone concentration is around the threshold value but not when it's far above or below it.

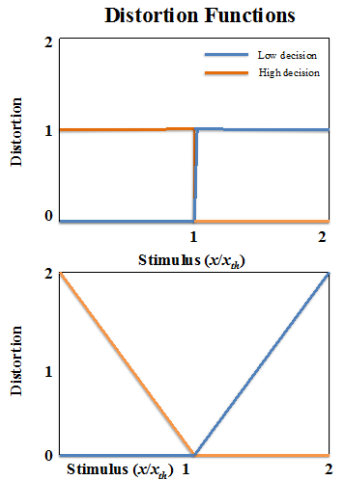

Figure 4.2.2: The top shows the distortion function as defined by equation 4.2.3. The bottom shows the distortion function that penalizes a false decision more as you move further away from the threshold. The same stimulus distribution was used for both distortion functions [1].

To model a more realistic scenario Porter, Andrews, and Iglesias used a graded distortion function shown in figure 4.2.2. This function does not penalize a false low decision more than a false high decision or vice versa. It does penalize either false decision of high or low more as you move further away from the threshold value x_{th}, equally in both directions. They first proposed a graded distortion function and used it with the same probability distribution $p_X(x)$ that represented the stimuli in their first generic model to compute a new $R(D)$. Once this rate distortion curve was calculated with the graded distortion function, it provided an ideal

conditional probability of deciding "high" as a sigmoidal function of the pheromone concentration for each value of distortion on the $R(D)$ curve. Each separate value for D was fitted with a Hill-type coefficient to represent its slope [1]. As expected, their theoretical models depicted that the majority of cells should be expected to mate or not mate at low values of D, when the pheromone concentration was far above or below the threshold respectively. At concentrations that were very close to x_{th}, a bimodal population of cells was expected where many mate but many others don't mate [1].

The process described above where's it's expected to see some cells mate and others not mate is an example of a random strategy referred to as Bet-hedging. Bet-hedging strategies consist of cells with the same genotype expressing different phenotypes [10]. Random strategies such as Bet-hedging are adopted because sometimes the mutual information required for accurate decision-making isn't worth the extra cost in energy needed to obtain more accurate decisions, such as when the pheromone level is near its threshold value [1, 10]. In situations where there's little information available to cells, an isogenic population of cells may express different phenotypes to help increase the variance in fitness for the population. Another example of a bet-hedging strategy is phase-variation in bacteria. The example of differing phenotypes for a cell choosing to mate or not near threshold values demonstrates how cells may benefit from Bet-hedging strategies, because they're conserving energy, corresponding to lower rates of mutual information. However, the example of bacteria expressing different phases isn't as clear. While it's evident that the population of cells as a whole should benefit from an increase in different phenotypes, it seems that the variance in fitness for any cell to express this different phenotype would decrease. This phenomenon is known as a cooperative strategy where the cell may choose a decision-making strategy that decreases the variance in fitness for that single cell, but it increases the variance in fitness for the cell population [1].

A cell's ability to execute advantageous decision-making strategies is contingent upon its prior knowledge acquired about its environment as well, and these decision-making strategies can by adjusted over time through adaptive learning. Moreover, it's important in many situations for cells to anticipate future changes so that they can prepare for any circumstances it may expect [1, 10, 13, 14].

E. coli can also be used to demonstrate the importance of cells acquiring a method of adoptive learning brought forth by changes in the environment and expected future changes. When *E. coli* leaves the soil and enters the human gut it generally makes the decision to express transcriptional factors to compensate for the loss in oxygen [10, 13]. The *E. coli* have learned through evolutionary time scales to associate an increase in temperature to be followed by a loss in oxygen. Experiments have shown that once *E. coli* has been subject to an increase in temperature, it expresses these transcription factors even when there's still some amount of oxygen present [10, 13]. Moreover, once *E. coli* confronts lactose it expresses the genes necessary to metabolize maltose, because lactose appears before maltose when progressing through the human gut. The *E. coli* expects to encounter maltose after confronting lactose. So after weighing the advantages and disadvantages (costs and benefits) of the most probable future state of the environment, it has a high probability of expressing the genes to metabolize the maltose [10, 14]. These decisions are learnt through evolutionary time scales, and microevolution experiments have demonstrated that when *E. coli* was subject to an increase in temperature paired with an untypical amount of oxygen created in an artificial environment, it learned to disassociate the rise in temperature with a decrease in oxygen. It evolved to a point that the probability of expressing the transcription factors for the loss in oxygen was significantly reduced when presented with a rise in temperature [10, 13]. Similarly, it was shown that actuation of the maltose operon was diminished when *E. coli* was subjected to an artificial environment that possessed lactose that wasn't followed by the presence of maltose. The reason for *E. coli*

choosing to omit the superfluous transcriptional factor for gene expression in these examples is to conserve energy so that it's not wasted on nonessential tasks [10, 14].

Cells must note the facets of a stimulus and responses associated with it after encountering them to increase their variance in fitness. They should terminate actions associated with the stimulus once it ceases or decreases below the threshold value for prolonged periods of time. Through adaptive learning, the threshold of any stimulus to initiate a response may increase or decrease for future responses. This concurs with idea of the 3-step decision-making plan for cells as it weighs the costs and benefits related with the responses of their environment, which is always in flux and change with time. By measuring quantities through adaptive physiology, the cells allow for the greatest rate of reproduction and growth that's possible. Eliminating all unneeded gene expression it can allows for more energy to be used for purposes that directly advance evolution, such as finding food sources, applying it to gene expressions that are pertinent for survival, or for reproduction. The decision-making strategy where cells learn to respond accordingly, to a new probability distribution of a stimulus, is referred to as hysteresis. This phenomenon, which has been observed experimentally as just described, can also be demonstrated using models under the rate distortion framework, where the theoretical models agree with empirical studies.

The concept of hysteresis was also demonstrated by Porter, Andrews, and Iglesias when the rate distortion curve was computed for two classes of source distributions. They computed an $R(D)$ curve for both a "likely low" and a "likely high" source distribution for three different sets of source distributions. The "likely low" distribution is similar to the one described in the two examples above where it has been assumed that the stimulus is more likely to be below the threshold x_{th} then above it. The "likely high" source distribution assumes that the stimulus is more likely to be above the threshold x_{th} then below it. They used the same graded distortion function shown in figure 4.2.2 for all three simulations [1].

It was found that the optimal conditional probability of deciding "high" when the distribution is "likely low" is a sigmoidal function of stimulus level where the decision to choose "high" occurs at the threshold x_{th}, much like the example with the graded distortion function described above. However, the probability of deciding "high" when the distribution was "likely high" also produced a optimal conditional probability of deciding "high" as a function of stimulus level, but the decision to choose "high" occurred at a stimulus level below the threshold x_{th} [1]. This shows that when cells have been exposed to a high level of stimulus for an extended amount of time, they choose to respond to lower levels of the stimulus than the same cells experiencing the same stimulus with a "likely low" distribution. This is similar to the cell described in section 4.1 that tried to infer the state of extracellular sugar concentration. When this concentration was inferred to be "high," the probability of experiencing a future "high" stimulus increased, at least for some amount of time. The advantages and disadvantages of each response will change with an intrinsic change in the stimulus distribution. This was represented when *E. coli* eventually learned to not express the maltose operon in artificial environments to conserve energy, because what was once an advantage to express the genes that would metabolize maltose became a disadvantage from wasting cellular energy on genes no longer needed [10]. For the three simulations with both "likely low" and "likely high" there was a space between the two sigmoidal curves. This space characterized an expectation to observe both cells responding to the stimulus are not responding to the stimulus depending on if they were conditioned to the "likely high" distribution or "likely low" distribution respectively. The area between the curves representing both decisions depending on what the cells where accustomed to grew larger as the distributions became more disjointed [1].

To further demonstrate hysteresis in real-world scenarios, it has been demonstrated experimentally that *E. coli* cells grown in 1 mM TMG to prompt the *lac* operon gene needed levels below 3 μM of TMG to completely turn off this gene. It required a change in stimulus level on the order of magnitude of 10^3 for the *E. coli* to become conditioned to a new probability

distribution of stimulus. Additionally, these cells grown without TMG needed treatment with levels above 30 μM to completely turn on the gene [1]. The theoretical models produced that described hysteresis thus agrees with experimental observations of cells responding to new stimuli distributions. Furthermore, the response to these distributions agrees with the 3-step cellular decision-making process as proposed by Perkins and Swain by using decision theory [10].

As stated previously in the third step of the 3-step decision-making process, cells often have to make decisions in the presence of other decision-making cells what action is the best. Within a population of cells, for any given cell, interactions between other cells and between these other cells and their extracellular environment may alter the fitness of decision-making strategies for the given cell over time, and therefore for the population of cells at large. Competition could encourage cells to use strategies that seem undesirable for the success of the individual but are of particular importance for the survival of the population [10]. Decision theory supplies a way to determine the best response by weighing the costs and benefits of each response when presented with uncertain information. The decision-making strategies at the level of cell populations can no longer by analyzed with just decision theory. Cells executing decisions to help improve the fitness of the population can be evaluated under evolutionary theory, which states that decisions are not made in isolation by individual cells. Instead decision-making strategies are made along side other competing decision-makers [10]

Hamilton states that natural selection may be considered under the context of inclusive fitness-the reproductive success of any organism depends upon the reproduction of other organisms of its species, because they share the genes of that organism and contribute to the total gene pool of the species [10, 15]. In a cellular context, indirect fitness consists of the direct fitness of the cell, the offspring produced by the cell, and the indirect fitness from the progeny of other cells in the population. Thus, a decision-making strategy that appears to reduce the cells direct fitness should be considered to increase the indirect fitness of the cell (the fitness of the cell

population) and is therefore desirable at the level of cell populations [10, 15]. Hamilton proposed a rule referred to appropriately as Hamilton's rule. It can be represented as:

$$rb > c. \tag{4.2.9}$$

In this equation r is a measure of genetic relatedness of a cell and the population to which it belongs, b is the benefit of fitness to the population, and c is the cost in fitness for the cell making the decision [10, 15]. A decision-making strategy that benefits the fitness of the cellular population but possibly provides detrimental consequences to the cell may be described with Hamilton's rule. A cooperative strategy such as bet-hedging or apoptosis may be opted for if $r>0$, the benefit b is high, and the cost c is low.

There's a higher probability of a cooperative strategy benefiting a population evolutionarily if a cell demonstrating the action is surrounded by other similar cells, because following Hamilton's rule, r increases with cooperative strategies, which is a measure of genetic relatedness between the recipient and cooperator [2, 15]. Porter, Andrews, and Iglesias expanded on their generic model again by claiming that asymmetric distortion functions lead to irreversibility. The asymmetric distortion function is different from the graded distortion function, because it penalizes a false "low" decision more heavily than a false "high" decision for a given threshold. The graded distortion function simply allows more false decisions of either "high" or "low" as the stimulus distribution moves away from the threshold; it doesn't penalize one decision more than the other [1].

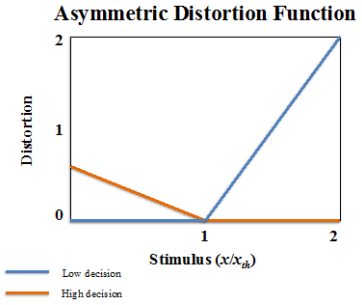

Figure 4.2.3: The asymmetric distortion function shown penalizes false a low decision more then false "high" decisions, because it represents real-world situations where a false "low decision is more disadvantageous to the cell then false "high" decisions. Asymmetric distortion functions demonstrate cells that show irreversible behavior [1].

Irreversibility can be demonstrated by situations where a false "low" decision is more detrimental to the cell, or more appropriately the cell population, than a false "high" decision. They considered a scenario where an incorrect "low" decision was penalized more than an incorrect "high decision;" the asymmetric distortion function is shown if figure 4.2.3. When the stimulus distribution was likely to by high the decision was always y=high, despite the actual value of the stimulus x [1]. However, when the stimulus distribution was likely low the decision made was comparable to the decision of their generic model. A false "high" decision has a lower level of distortion, which moves the $R(D)$ curve to the left for a likely high distribution. A prespecified value of $E[d]$ can be acquired with less mutual information between the stimulus and the decision, so the complexity of the cell is decreased with respect to decision making. In one theoretical scenario where the model desired a small value for D, the model with the asymmetric distortion function required no mutual information for a likely "high" stimulus distribution to achieve the same amount of distortion [1].

Cells that exhibit behavior where a false "low" decision is penalized more than a false "high decision" display irreversible behavior. Cells continually make the decision y=low when subjected to a "likely low" stimulus distribution, and convert to the "high" decision state when they encounter a substantially large stimulus. Once the cell becomes conditioned to the "likely high" distribution, it will not go back to the "low" decision state no matter how far the stimulus drops [1]. These conversions occur because it's much more harmful for a cell to execute an incorrect "low" decision than it is to execute an incorrect "high" decision.

Irreversible behavior may be demonstrated with apoptosis where cells decide to systematically destruct when exposed to a certain level of caspase 8 for the benefit of the population [1]. Once the cell makes the decision to undertake this programmed cell death by transitioning to the "high" decision state, no levels of caspase 8 can undo the decision for the cell to self-destruct, no matter how low it drops. This action aids in limiting disorderly cell-proliferation, which could possibly lead to ailments such as cancer [10]. The example of

apoptosis ideally demonstrates how Hamilton's rule works in real-life scenarios. The benefit gained from the cellular population may be significant if the cell is infected, mutated, or not functioning properly. The fitness of the cell wouldn't increase when it decides to destroy itself, but it may help the survival and reproduction of the cellular population at large.

Although growth rate (reproduction) is what primarily defines fitness, it's possible to analyze the efficiency of other biochemical networks that do not directly affect growth to define fitness [10]. Many cellular systems depend on spatial variations in chemoattractant concentration to direct cell migration [2]. Because the goal of this pathway is distinctively defined, it serves an admirable and alternative example of fitness [10].

Tumbling occurs when a cell is trying to move closer to or further away from the source of a chemical during chemotaxing. After traveling a certain distance, a chemotaxing cell stops to sense the level of the chemical and then compares it to a signal that determined the chemical concentration at an earlier time. Cells seek to travel in the direction of the highest or lowest concentration gradient of whichever stimulus it's searching for [10]. If signals reach the cell informing it that it's not aligned with the most optimal direction of a concentration for a certain stimulus, then it will stop and reorient itself in a process called tumbling before proceeding in the new direction it has determined to have the steepest or lowest gradient of the chemical it desires to reach. If the cell receives signals informing it that it's traveling in the direction of the desired gradient, then tumbling is suppressed [10].

In models created by Andrews and Iglesias they modeled the true angle of the chemical gradient as $\theta_{s,}$ which a chemotaxing cell wishes to move towards. The actual angle that the cell is moving in due to stochastic factors in the chemotaxing process is θ_r [10]. The accuracy of the decision to tumble and change direction in response to the inferred likely environment can be used to quantify the cost in fitness of the response [10]. Perkins and Swain state that if the cell fails to chemotax towards the source then it receives a cost in fitness defined as [2, 10]:

$$c(\theta_s, \theta_r) = \frac{1}{2}\left[1 - \cos(\theta_s - \theta_r)\right]. \qquad (4.2.10)$$

The equation obtains the minimal value of 0 when $\theta_s = \theta_r$, and obtains the maximum value of 1 when the two are misaligned by 180° [2, 10]. Perkins and Swain stated that the expected cost may be computed using a Bayesian approach as follows [2, 10]:

$$\bar{c} = \int d\theta_s d\theta_r P(\theta_r | \theta_s) P(\theta_s) c(\theta_s, \theta_r). \qquad (4.2.11)$$

However, Andrews and Iglesias who constructed the mathematical models for chemotaxing cells regarded the formula shown above as cost in equation 4.2.10 as the distortion function for chemotaxing cells. This can be alternatively shown as [2]:

$$d(\theta_s, \theta_r) = \frac{1}{2}\left[1 - \cos(\theta_r, \theta_s)\right]. \qquad (4.2.12)$$

Much like the distortion function for cells exposed to a given source of pheromone concentration, this distortion function defines the goals of the decision-making pathway of the biological system. It quantifies how "disadvantaged" or "distorted" a decision is with respect to a stimulus and effectively depicts the quality of a decision [1]. If the cell does not expend enough energy to achieve a level of distortion needed to adequately flourish, then it will receive a cost in fitness as defined by Perkins and Swain represented by an integration of the distortion function with the joint probability distribution. The expected cost as defined by Perkins and Swain is effectively the expected distortion, which can be shown by equation 4.2.8. The mutual information will provide the bond for how accurately the cell can relate θ_s to θ_r that will achieve the maximum value of D [10]. This information-theoretic calculation was performed by Andrews and Iglesias.

In the mathematical models constructed with the distortion function shown in equation 4.2.12, they defined two classes of source distributions $P_{\Theta_s}(\theta_s)$. They then modeled the signal transduction network of the chemotaxing cell choosing θ_r based on θ_s as the conditional probability $p_{\Theta_r|\Theta_s}(\theta_r | \theta_s)$ with downstream binding components shown if figure 4.2.4 [2]. The

input Θ_s is a random variable that represents the angle of the chemoattractant field; this is the true direction that the cell desires to move in. The cell makes a decision Θ_r based on the observed gradient that represents either the location of intracellular markers used to detect the chemoattractant gradient or the biased direction in which pseudopods are generated. The first class of source distribution considered was a uniform source distribution in which the cell assumes Θ_s is uniformly distributed. This scenario represents naive cells that have no a priori bias regarding the direction of the chemoattractant gradient. It may represent real-world scenarios where the cell has been newly introduced to a gradient or cells that are constantly experiencing changes in the chemoattractant source [2]. The second class is the normal source distribution with mean μ_s and variance σ_s. The mean is a reflection of the bias direction the cell has towards a chemoattractant gradient and the variance is a reflection of how certain the cell is about this biased direction. Normal source distributions may be indicative of cells that have been chemotaxing in a given direction for a prolonged period of time [2]. Many cells such as *D. discoideum* develop distinctive polarized leading and trailing edges after being exposed to a chemoattractant gradient over an extended time period. The anterior edges of such cells become more sensitive and cause the cell to change direction when they encounter alterations in the chemoattractant field. These polarized trailing edges are what provide the biased direction of chemotaxing cells represented by the normal stimulus distribution described above. Under the framework of rate distortion theory, it's possible to demonstrate that the degree to which these biased directions influence the cells can be overcome by steeper gradients or more strict distortion requirements [2]

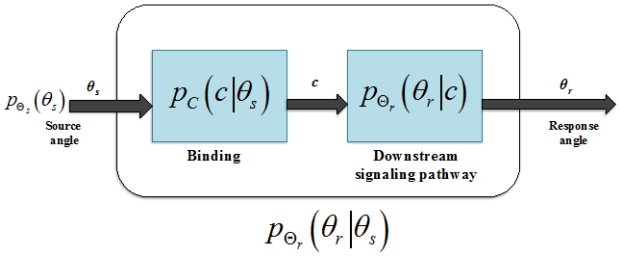

Figure 4.2.4: The response angle is determined by the conditional probability shown above, which has ligand-bound receptor complexes *C* and downstream binding components [2].

Once the probability distribution $P_{\Theta_s}(\theta_s)$ and distortion function $d(\theta_s, \theta_r)$ are both specified, it's possible to compute the rate distortion curve *R*(*D*) [1]. The rate distortion curve was solved computationally much in the same way as it was for the example with pheromone concentration using the same four steps. First the marginal output distribution $p_{\Theta_r}(\theta_r)$ is set up as a uniform distribution. Then the conditional distribution $p_{\Theta_r|\Theta_s}(\theta_r|\theta_s)$ is computed to minimize the mutual information subject to the distortion constraint in equation 4.2.13, and then the marginal distribution $p_{\Theta_r}(\theta_r)$ is computed to minimize the mutual information subject to the distortion constraint in equation 4.2.14. The later two steps are repeated until $p_{\Theta_r}(\theta_r)$ and $p_{\Theta_r|\Theta_s}(\theta_r|\theta_s)$ converge. The points produced form the rate distortion curve, where the amount of distortion is determined by the choice of Lagrange multiplier λ [1, 2]:

$$p_{\Theta_r|\Theta_s}(\theta_r|\theta_s) = \frac{p_{\Theta_r}(\theta_r)e^{-\lambda d(\theta_s,\theta_r)}}{\sum_r p_{\Theta_r}(\theta_r)e^{-\lambda d(\theta_s,\theta_r)}} \tag{4.2.13}$$

$$p_{\Theta_r}(\theta_r) = \sum_s p_{\Theta_s}(\theta_s)p_{\Theta_r|\Theta_s}(\theta_r|\theta_s). \tag{4.2.14}$$

The rate distortion curve was derived for both the normal and uniform source distributions to show how a directional bias can affect the optimal decision-making strategy of chemotaxing cells. Simulations showed that greater mutual information was required for more

accurate chemotaxing. Also the cells with an a priori bias chemotax more efficiently than cells with no a priori bias when the direction of bias was aligned or close to the actual chemoattractant field [2]. Because the rate distortion curve for the normal distribution lies below that for the uniform distribution for all values of *D*, it requires more mutual information to achieve a desired level of accuracy in naïve cells. When the biased angle and the chemoattractant gradient were not aligned, larger values of distortion caused the cell to follow the direction of bias and the conditional distribution $p_{\Theta_r|\Theta_s}(\theta_r|\theta_s)$ was independent of the chemoattractant field. As the distortion was decreased, the optimal decision became progressively more aligned with the chemoattractant gradient [2].

Andrews and Iglesias also simulated how increasing the hill coefficients, which correspond to receptor sites used in signal transduction, affected the distortion for both types of distributions. It was shown that the expected distortion decreased for both distributions when the Hill coefficients were increased [2]. An increase in Hill coefficients is analogous to having more sites for signal transduction for detecting spatial heterogeneities in the chemoattractant field. This is represented by the binding sites *c* in figure 4.2.4. This can also be represented by what is increased when the mutual information is increased to achieve higher fidelity decisions. So under the framework of rate distortion theory it's possible to demonstrate that cells can achieve more accurate chemotaxis by expressing more receptor sites for signal transduction to detect spatial heterogeneities in the chemoattractant field. This corresponds to using more energy to achieve more accurate and/or faster propagation towards the desired chemical source. The cost in fitness of this response can be the gene expression of binding sites, and the benefit in fitness is reaching the desired chemical of the cell [2, 10].

To sum up, it has been shown using rate distortion theory that a cells decision-making strategy depends on three things: (1) it's prior knowledge about its environment represented by a change in stimulus distribution $p_X(x)$, (2) it's goals for the decision as specified by the distortion

function $d(x,y)$, and (3) how much metabolic cost it's willing to "pay" for an accurate decision represented by the mutual information $I(X;Y)$ [1]. These three factors of decision-making strategies are what ultimately make-up the components for weighing the costs and benefits as described by step 2 in Perkins and Swains 3-step decision-making process. This 3-step process provides an intimate relation with the rate distortion-theoretic approach for describing binary cellular decision-making by providing quantitative calculations for how cells deduce and make decision based on stochastic stimuli. The theoretical models computed using rate distortion theory agree with empirical observations for how cells react to their environment.

Chapter 5

Conclusions

Under the framework of rate distortion theory it's possible to explain the strategies incorporated by cellular decision-making systems that detect, process, and respond to environmental states and how they can possibly change with time. This theory analyzes these strategies in terms of the functions of information processing that biochemistry performs and not how the signaling network senses and evaluates this information with respect to the characteristics of the biochemistry [10]. Employing this method of information processing as described by rate distortion theory enables us to conceive evolutionarily conserved principles such as the choice of a cell expressing genes or not expressing them for the conservation of energy, or the cell choosing it's decision to undergo apoptosis or not to. In other words, this theory addresses questions with regards to the goals of the decision-making pathway with the advantage that these goals are mechanism-independent [1]. Much of the focus on biological studies addresses questions towards the mechanisms behind how the biochemistry functions: what mechanism is responsible for a certain response or what decision is produced by a specific mechanism? [1]. However, rate distortion theory addresses question regarding the goals of the decision-making pathways without accounting for the mechanisms of biochemistry responsible for how they function. For example, what are the goals of the biological system or what decision-making strategy will best attain a specific goal [1]? The goals of the decision-making pathway

are quantified by the distortion function, which demonstrates how "disadvantaged" or "distorted" a decision is with respect to a stimulus. Paired with the mutual information that quantifies the cost in energy required to achieve a desired decision, the rate distortion curve may be contrived. This curve demonstrates the intrinsic limit on how effectively a cell can achieve its goals at a given cost in mutual information. The rate distortion curve provides a slew of optimal decision-making strategies that achieve this inherent limit for different values of expected distortion [1].

By utilizing rate distortion theory, it's possible to describe the design and evaluation of decision-making systems with regards to their goals [1]. The former can be demonstrated by determining what decision-making strategy is the most ideal in achieving a specific fidelity requirement, when responding to a stochastic stimulus. As an evaluation tool it can analyze the optimality of a pathway if the goals are known or it can describe the goals of a pathway that's believed to be ideal [1]. For example, the rate distortion function given by equation 4.2.12 can consider the pathway with downstream signaling components of a chemotaxing cell as a system that has been optimized to attain the goals as described by the distortion function [1]. Perkins and Swain also considered this same distortion function in chemotaxing cells as the cost in fitness when it's integrated with the joint distribution, because the accuracy of a decision as described by the expected distortion can also be used to quantify the cost in fitness of a response [10]. This differs from the cost represented by mutual information, which essentially represents the metabolic cost of achieving a prespecified fidelity in decision-making. Instead it refers to the cost in fitness of a response defined as the expected benefits minus the expected costs of the response (usually as the expected benefit in growth rate minus the expected cost) [10]. Thus, the lower quality of decision fidelity as described by certain distortion functions is analogous to the cost in growth rate of the population brought forth by more incorrect decisions that these functions quantify. Higher levels of distortion leads to increased rates of incorrect decisions, and this is what ultimately determines the evolutionary stability of cells as defined by the fitness of a response. Therefore, the distortion functions can be considered to quantify the goals of a

decision-making pathway independent of the mechanisms responsible for attaining these goals, and it may also be regarded as the cost in fitness of the response [1, 10].

Quantifying the goals of decision-making pathways under the framework of rate distortion theory provides an alternative way of analyzing evolution. One of the major driving forces behind evolution is the need for organisms to execute correct decisions while using less energy than the other organisms in their environment [10]. This increase in efficiency to outcompete other organisms corresponds to the distortion-information point approaching the rate distortion curve, or using less mutual information to achieve the same amount of distortion. This may occur due to a fundamental change in the goals of the decision-making pathway as described by the distortion function or from a different stimulus distribution [1]. A demand for increased performance is an additional incentive for evolution, because higher fidelity decision-making results in more advantage for the organism when its environment or neighbors augment in complexity. This can be represented by the distortion-information point shifting from right to left on the rate distortion curve, because an increase in performance is worth the extra expended energy that's necessary in such circumstances [1].

Finally Porter, Andrews, and Iglesias argue that the rate distortion framework gives a supportive model to Perkins and Swains 3-step process: (1) a cell must infer the state or likely future state of the environment by sensing stochastic stimuli; (2) based on the stimuli sensed it weighs the advantages and disadvantages of each potential decision; and (3) it executes a decision so that it maximizes the fitness of the cellular population [1, 10]. Step 1 can be represented by how well the distortion function quantifies correct decisions and penalizes incorrect decisions, and how the expected distortion $E[d]$ depicts how accurate sensing must be. Step 2 can be demonstrated by how much the distortion function quantifies the disadvantages of alternative decisions and by how the expected distortion shows how much distortion the cell can spare in making a decision. Step 3 is carried out in proportion to how much information is accessible to the cell [1, 10]. Several observable cellular characteristics such as bet-hedging strategies,

hysteresis, and irreversibility can then be shown to be optimal under the framework of rate distortion theory and the 3-step process proposed by Perkins and Swain. When there's a scarcity of information available to the cell it will integrate randomness into its decision-making strategies to improve the variance in fitness of the population. This can be represented in models by Bet-hedging strategies and has been observed in empirical studies. More information is available to the cell when the fidelity of certain decisions is deemed more important to the cell (or population), such as in apoptosis. In these circumstances randomness is renounced and the cell merely executes the correct decision. The rate distortion framework enables design and analysis of a cellular decision-making process with a foundational optimality criterion by combining many aspects of these decision-making systems [1]. Establishing and understanding the strategies of decision-making pathways may possibly yield a method that associates systems and evolutionary biology to help understand biological design [10].

Works Cited

1. Porter, J.R., B.W. Andrews, and P.A. Iglesias, *A framework for designing and analyzing binary decision-making strategies in cellular systems.* Integr Biol (Camb), 2012. **4**(3): p. 310-7.
2. Andrews, B.W. and P.A. Iglesias, *An information-theoretic characterization of the optimal gradient sensing response of cells.* PLoS Comput Biol, 2007. **3**(8): p. e153.
3. Cover, T.M. and J.A. Thomas, *Elements of information theory*. Wiley series in telecommunications. 1991, New York: Wiley. xxii, 542 p.
4. Berger, T., *Rate distortion theory; a mathematical basis for data compression*. Prentice-Hall series in information and system sciences. 1971, Englewood Cliffs, N.J.,: Prentice-Hall. xiii, 311 p.
5. Aftab, C., Kim, Thakkar, and Yeddanapudi, *Information Theory and the Digital Age*. Massachusetts Institute of Technology: Cambridge, MA.
6. Kline, R.R., *What is Information Theory a Theory Of? Boundary Work Among Information Theorists and Information Scientists in the United States and Great Britain During the Cold War.* Interdiscipliary Perspectives by The American Society for Information Science and Technology, 2004: p. 15-28.
7. *Lagrange Multipliers*. University of California San Diego. p. 1-6.
8. Sullivan, M., *Fundamentals of statistics : informed decisions using data*. 2nd ed. 2008, Upper Saddle River, N.J.: Pearson Prentice Hall.
9. Blahut, R.E., *Computation of Channel Capacity and Rate-Distortion Functions.* IEEE Transactions on Information Theory, 1972. **IT-18**(4): p. 460-473.
10. Perkins, T.J. and P.S. Swain, *Strategies for cellular decision-making.* Mol Syst Biol, 2009. **5**: p. 326.
11. Libby, E., T.J. Perkins, and P.S. Swain, *Noisy information processing through transcriptional regulation.* Proc Natl Acad Sci U S A, 2007. **104**(17): p. 7151-6.
12. Dekel, E. and U. Alon, *Optimality and evolutionary tuning of the expression level of a protein.* Nature, 2005. **436**(7050): p. 588-92.
13. Tagkopoulos, I., Y.C. Liu, and S. Tavazoie, *Predictive behavior within microbial genetic networks.* Science, 2008. **320**(5881): p. 1313-7.
14. Mitchell, A., et al., *Adaptive prediction of environmental changes by microorganisms.* Nature, 2009. **460**(7252): p. 220-4.
15. Hamilton, W.D., *The genetical evolution of social behaviour. I.* J Theor Biol, 1964. **7**(1): p. 1-16.