# Optimal Information Storage: Nonsequential Sources and Neural Channels

by

Lav R. Varshney

B.S., Electrical and Computer Engineering
Cornell University, 2004

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2006



Author .....
Department of Electrical Engineering and Computer Science
May 12, 2006

Certified by .....
Sanjoy K. Mitter
Professor of Electrical Engineering and Engineering Systems
Thesis Supervisor

Certified by .....
Vivek K Goyal
Assistant Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by .....
Arthur C. Smith
Chairman, Department Committee on Graduate Students

# Optimal Information Storage:

# Nonsequential Sources and Neural Channels

by

Lav R. Varshney

Submitted to the Department of Electrical Engineering and Computer Science
on May 12, 2006, in partial fulfillment of the
requirements for the degree of
Master of Science in Electrical Engineering and Computer Science

## Abstract

Information storage and retrieval systems are communication systems from the present to the future and fall naturally into the framework of information theory. The goal of information storage is to preserve as much signal fidelity under resource constraints as possible. The information storage theorem delineates average fidelity and average resource values that are achievable and those that are not. Moreover, observable properties of optimal information storage systems and the robustness of optimal systems to parameter mismatch may be determined. In this thesis, we study the physical properties of a neural information storage channel and also the fundamental bounds on the storage of sources that have nonsequential semantics.

Experimental investigations have revealed that synapses in the mammalian brain possess unexpected properties. Adopting the optimization approach to biology, we cast the brain as an optimal information storage system and propose a theoretical framework that accounts for many of these physical properties. Based on previous experimental and theoretical work, we use volume as a limited resource and utilize the empirical relationship between volume and synaptic weight. Our scientific hypotheses are based on maximizing information storage capacity per unit cost. We use properties of the capacity-cost function, $\epsilon$-capacity cost approximations, and measure matching to develop optimization principles. We find that capacity-achieving input distributions not only explain existing experimental measurements but also make non-trivial predictions about the physical structure of the brain.

Numerous information storage applications have semantics such that the order of source elements is irrelevant, so the source sequence can be treated as a multiset. We formulate fidelity criteria that consider asymptotically large multisets and give conclusive, but trivialized, results in rate distortion theory. For fidelity criteria that consider fixed-size multisets, we give some conclusive results in high-rate quantization theory, low-rate quantization, and rate distortion theory. We also provide bounds on the rate-distortion function for other nonsequential fidelity criteria problems. System resource consumption can be significantly reduced by recognizing the correct invariance properties and semantics of the information storage task at hand.

Thesis Supervisor: Sanjoy K. Mitter
Title: Professor of Electrical Engineering and Engineering Systems

Thesis Supervisor: Vivek K Goyal
Title: Assistant Professor of Electrical Engineering and Computer Science

## Acknowledgments

I have had the great privilege of receiving encouragement, advice, and support from not one but two excellent thesis supervisors, Sanjoy Mitter and Vivek Goyal. The content and presentation of the thesis have been greatly enhanced by their efforts. I thank them both for their mentorship over the past two years.

The look into robustness of optimal information systems was suggested by Sanjoy Mitter; I thank him for his incisive questions and for pushing me in that direction. The problem formulation and central results of Chapter 3 were developed jointly with Mitya Chklovskii of Cold Spring Harbor Laboratory. I thank him for hosting me at CSHL and for facilitating my foray into neuroscience; it has been a true collaboration. I also thank Jesper Sjöström for graciously providing experimental data for comparison with theory, without which the scientific method could not be completed. Investigation into the importance of order in information theory was suggested by Vivek Goyal, and many of the results of Chapter 4 were obtained jointly with him. I thank him for sharing his insights and enthusiasm.

The contents of the thesis have also been influenced in various ways by Aman Chawla, Ronald Kline, Julius Kusuma, Anant Sahai, Ram Srinivasan, and several other colleagues in LIDS and in RLE, as well as numerous people at institutions with which I have been affiliated in the past.

I also thank all members of the institution with which I have had the longest and deepest affiliation, the Varshney family. In particular, I thank my brother Kush for providing encouragement and also, luckily for me, additional technical viewpoints. I thank my parents for providing unwavering and limitless support, and other assistance which cannot be described in words.

- Financial support provided by the National Science Foundation Graduate Research Fellowship Program.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

Information storage and retrieval are fundamental for the preservation and use of cultural, scholarly, and literary heritage. For a typical person, the most local and immediate seat of information storage is the brain. Papers on a desk, books on a shelf, photographic negatives in a box, data files on a computer hard drive, and music on a compact disc are more distant seats of information storage. Still more distant are non-local repositories of these print, film, magnetic, and optical storage media, such as libraries and digital archives. These distant, technological storage media supplement the local, biological storage medium.

Information storage and retrieval systems must meet certain performance objectives, such as maintaining stored signals within a certain fidelity. These systems must also operate with limited resources, such as energy or physical volume. Since information storage and retrieval is a form of communication from the present to the future, the mathematical formalism of information theory is directly applicable to the study of these systems. It is the purpose of this thesis to study the tradeoff between the two fundamental parameters of information storage: the average signal fidelity maintained and the average resources consumed.

In Chapter 2, we review the information theoretic formulation of information storage, and present several characterizations of optimal systems. We also give some examples of optimal information storage, inspired by paper, magnetic, and optical storage media. In Chapter 3, we adopt the optimization approach to biology and discuss physical properties of the mammalian brain that would be required for optimal information storage. These optimal properties are consistent with experimental

results and also lead to non-trivial scientific predictions that would need to be tested in future experimental investigations. In Chapter 4, we discuss a family of fidelity criteria that correspond to a natural invariance property in numerous information storage applications. We deal with non-single-letter fidelity criteria that arise when the order of elements to be stored is irrelevant and develop fundamental bounds on optimal performance from all of the major branches of source coding theory.

The next section gives a more detailed summary of the contributions of the thesis.

## 1.1 Outline and Contributions

#### Chapter 2: General Information Storage Systems

It has long been evident that information storage and retrieval is a special case of point-to-point information transmission,[1] arguably the central problem in information theory. In Section 2.1, we define the two main system parameters that will be used to judge performance—average incurred distortion ($\Delta$) and average incurred cost ($B$)—in terms of fidelity criteria and resource criteria, respectively. The fidelity criterion quantifies the notion of signal reproduction quality, comparing the signal that was stored with the one that was retrieved. The fidelity criterion should incorporate the semantic nature of the information transmission problem; the distortion assigned to a retrieved signal that is almost as useful as the original stored signal for the task should be small. The fidelity criterion will be of central importance in Chapter 4. The resource criterion quantifies the cost of storing and retrieving a particular signal. It should reflect the cost of building and operating the information storage system. The resource criterion will be of central importance in Chapter 3. In Section 2.2, we define an optimal information storage system to be one where $\Delta$ cannot be reduced without increasing $B$ and $B$ cannot be reduced without increasing $\Delta$.

The information storage theorem, Theorem 5 in Section 2.3, divides the set of $(\Delta, B)$ duples into two classes: the possible and the impossible. Optimal information

[1]Multiple retrieval attempts may be thought of as a broadcast system and multiple storage attempts may be thought of as a multiple descriptions system, however this is not the view that we adopt. See Chapter 2 for further discussion and [1] for related discussion.

storage systems lie at the boundary between the two classes. There are numerous ways of characterizing an optimal information storage system, and in the chapter, we review three of them. The first is what we call the *rate matching* interpretation, described in Section 2.4. Here optimality is defined as requiring $R(\Delta) = C(B)$, where $R(\cdot)$ and $C(\cdot)$ are the rate-distortion and capacity-cost functions, which measure information rate. The *factorization* interpretation, given in Section 2.5, emphasizes that marginal probability distributions of the information storage system must be of a particular form. The distribution of letters at the storage channel input must be the capacity-cost achieving input distribution and the joint distribution of the source letters and the retrieved reproduction letters must be the rate-distortion achieving distribution. A third interpretation given in Section 2.6, the *measure matching* interpretation, imposes conditions on the fidelity and resource criteria to make an information storage system optimal. Since these three interpretations are essentially equivalent and address properties of three different parts of an information storage system, they provide complementary views of system optimality. These complementary views also suggest three different architectures of system design that all result in optimality. The different architectures partition incurred distortion in different parts of the system.

The review of the information storage theorem and several characterizations of optimal information storage serve to collect results from information theory and to establish terminology and notation. Although the chapter is mainly a review, we also present some new results. We give two new examples of optimal information storage systems and review a third example of an optimal system. Given in Section 2.4.1, the first new example considers the storage of a source with a context-dependent fidelity criterion in a magnetic or optical medium. Context-dependent fidelity criteria are a first step to incorporate meaningful semantics into the storage problem and so this example is something of a preview of Chapter 4. The duality relationship between storage of a source under the context-dependent fidelity criteria and storage in magnetic or optical media with objectionable intersymbol interference does not seem to have been noticed previously. The second new example, Section 2.6.1, involves using

a spacing storage channel which corresponds to timing channels used in information transmission. Such spacing channels have been used to covertly store information in paper documents. This example uses the measure matching conditions and provides intuition into the measure matching characterization. Not only are the measure matching conditions used, but also the source and channel model are similar to models of excitatory synaptic information storage, so the example serves as somewhat of a preview of Section 3.5. The third example deals with the storage of a Gaussian source in an AWGN channel (Section 2.5.1). Such a system is used to obtain results about the robustness of an information storage system to modeling errors, which forms the final section of Chapter 2.

This final section, Section 2.7, deals with robustness of optimal systems to parameter mismatch. For a system designed using separate source and channel coding, if the rate-distortion value exceeds the capacity-cost value, then the probability of error quickly goes to one, and the average incurred distortion becomes large. Contrariwise, when a Gaussian source is stored in an AWGN channel and the system is designed using single-letter codes, when the capacity degrades below the rate-distortion value, the loss in performance below optimality is not much. In fact, to a first-order approximation, there is no loss between the optimal performance and the performance of the mismatched single-letter scheme. Thus, systems designed according to uncoded or single-letter principles are robust to uncertainty in some noise parameters. Some comments on the robustness of single-letter schemes in the finite alphabet case are also made. These new results show that even though two systems may achieve the same optimal performance for the same source and the same channel in the $(\Delta, B)$ sense, there may be reasons to prefer one system design over another.

#### Chapter 3: Information Storage in Neural Memory Channels

The mammalian brain is a complicated system composed of billions of neurons and trillions of synapses. A common mechanistic view of the brain is as a computer, however besides computing, another important task of the brain is to remember things. Some even argue that "...the brain doesn't 'compute' the answers to problems; it

retrieves the answers from memory.... The entire cortex is a memory system. It isn't a computer at all" [2, p. 68]. Thus the brain falls naturally into the information storage and retrieval framework of this thesis. Although previous philosophical and psychological investigations of the brain as a memory system have used information theoretic principles, particularly in the heyday of information theory euphoria in the 1950s [3, 4], and previous quantitative neuroscience investigations have used the information theoretic framework to study the transmission of information between neurons, particularly in the context of sensory processing [5,6], the work presented in this chapter is, to the best of our knowledge, the first application of the information theory framework to a quantitative neuroscience investigation of memory. Thus the entire chapter, including the problem formulation; the development of the system model; the scientific hypotheses developed from the model through the optimization approach to biology; and some experimental tests of these hypotheses may be considered novel.

We formulate a unified set of optimization principles that explain physical characteristics of synapses that have been observed and also make predictions that would need to be tested in future experiments. In particular we try to explain why synapses are rather noisy (Section 3.3), why synaptic connectivity is sparse (Sections 3.4 and 3.5), and why the distribution of synaptic efficacies contains many strong synapses (Sections 3.4 and 3.5). We also comment on observations of discrete-valued synapses (Section 3.6). Since it is widely believed that long-term memories are stored in synapses through processes of synaptic plasticity modulating synaptic efficacy, the information storage channel is identified with the synapses; each synapse is a discrete-space channel usage. Based on previous theoretical and experimental work, the resource criterion is identified with synaptic volume, since volume is a scarce resource and also measures metabolic energy expenditure.

With the synapse channel model and the volume resource criterion established, we use the factorization interpretation and the measure matching interpretation of optimal information storage, focusing on the channel half. For the neural information storage problem, we would like to not only achieve capacity-cost, but also achieve

optimal capacity per unit cost. This second stage of optimization determines a particular point on the capacity-cost curve. It should be noted that although systems obeying weaker notions of optimality than implied by the factorization interpretation conditions may be able to achieve optimal capacity per unit cost when there are zero-cost symbols [7, 8], the volume resource criterion has no such free symbols, and the factoring interpretation conditions must be met for optimality. The optimality conditions from Chapter 2 lead to optimization principles that describe the distribution of synaptic efficacies and synaptic volumes, which can be measured through electrophysiology experiments and electron microscopy experiments, respectively. Properties required for optimality are garnered from both the factoring characterization and the measure matching characterization. In Section 3.7, we compare the qualitative and quantitative predictions of our theory to the results of electrophysiology experiments. In large part, the experimental results verify the theoretical predictions. Experiments relating synaptic efficacy to synaptic volume as well as experiments determining the distribution of synaptic volume are needed to further test the theory. A final point that is made in the chapter is that many forms of synaptic noise may not have been captured in the experiments, so further experiments may change the particular quantitative predictions, however the method of making predictions and the qualitative predictions themselves would remain the same.

#### Chapter 4: Source Coding for Nonsequential Sources

For a given semantic framework with a given fidelity criterion, Shannon's statement that "semantic aspects of communication are irrelevant to the engineering problem" [9] is certainly true, however establishing the semantic framework is not a trivial task. Desirable properties for a fidelity criterion include such things as tractability, computability, and semantic significance. These desiderata, particularly tractability and semantic significance, are often not attainable simultaneously. In this chapter, we develop a fidelity criterion that is not only mathematically tractable and easily computable, but it also bears significance in numerous information storage applications.

Numerous sources produce sequences of messages that are in arbitrary order. For example, the order of financial transactions carried out by a bank, the order in which results of electrophysiology experiments used in Chapter 3 are collected, or the order in which songs are placed in an online jukebox are rather arbitrary. The records can be reordered without ill effect. Thus these source-destination pairs display order irrelevance in their semantics, and the fidelity criterion should reflect this invariance. There are numerous other invariant representations one can envisage, ones that result in no degradation of utility. Melodies are invariant to changes in key, and can be represented with a differential encoding with no perceivable loss. If photographs are stored in a shoebox and the prints become arbitrarily rotated, it does not degrade quality, as the destination can easily determine the correct orientation. Many algorithms in machine learning and machine vision try to form invariant representations to facilitate detection, recognition or other tasks, with varying degrees of success [10]. There are also numerous suggestions that animals form invariant representations. Although we do not attempt to tie the results of this chapter to a biological substrate, there are certainly results from psychology that are suggestive of the applicability of the semantics that we adopt to human memory. In our formalization, sources with a fixed order are equivalent to sources with order that does not matter. A simple psychology experiment that displays fixed-order storage is to try to recite the alphabet out of order: it is much more difficult and seems to involve buffering the alphabet in the correct order first.

The order irrelevance semantics are particularly nice since the associated invariant representation is mathematically well-characterized. For discrete alphabets, the type (empirical distribution) is a sufficient statistic for sequences where order does not matter. For continuous alphabets, the order statistics are a natural representative of the equivalence class of arbitrarily ordered values. Thus the semantic fidelity criteria operate on the types or the order statistics. If desired, one can also state the fidelity criteria in group theoretic terms using permutation groups and invariance under the group operator. Source coding can be thought of as projecting source outputs to canonical representations for each equivalence class, which results in no

distortion, and then performing further compression if needed. By imposing rules on the arrangement, the source output patterns become ordered and thus the dimension of the possible outcomes is reduced. This dimensionality reduction is a basic tool that we use in developing our new results on information storage systems for nonsequential sources, as now described.

Nonsequential semantics allow interesting statements to be made about joint source and channel coding, using the measure matching conditions. Furthermore, order irrelevance suggests itself as a way of embedding covert information into a source sequence. The main focus, however is on separate source and channel coding in the manner of the rate matching interpretation of optimal information storage. First we show that information about the values of a sequence and the information about the order of a sequence are independent for exchangeable sources. Next we formulate some rate-distortion problems with nonsequential fidelity criteria and show that the rate-distortion function is the $(R = 0, \Delta = 0)$ point. Alternative nonsequential fidelity criteria yield non-trivial results. A first step is to study the design of low-rate and low-dimension quantizers. We obtain conditions on when quantization and projection to canonical representatives can be interchanged without loss of optimality. We also obtain results in high-rate quantization theory, a branch of source coding theory related to rate-distortion theory, that show how much rate can be saved using the nonsequential fidelity criterion as opposed to more traditional nonsemantic fidelity criteria. Bounds on the rate-distortion function for a problem with the alternative nonsequential fidelity criteria are also obtained.

#### Chapter 5: Conclusion

The final chapter recapitulates the results of the thesis, interpreting some results in terms of a hierarchy of difficulty of communication problems. Suggestions for future work are also made. One of the main issues that is not considered in the thesis is the issue of time delay, which is known to be important for the control of dynamical systems and has been considered in the design of computer memory systems.

## Bibliographical Note

The problem formulation and some of the results of Chapter 3 have been presented at some meetings:

- L. R. Varshney and D. B. Chklovskii, "Reliability and Information Storage Capacity of Synapses," presented at *2005 Cold Spring Harbor Laboratory Meeting on Learning & Memory*, Cold Spring Harbor, New York, 20-24 April 2005.

- D. B. Chklovskii and L. R. Varshney, "Noisy Synapses and Information Storage," presented at *Neuroscience 2005: Society for Neuroscience 35th Annual Meeting*, Program No. 965.17, Washington, D. C., 12-16 November 2005.

- L. R. Varshney, P. J. Sjöström, and D. B. Chklovskii, "Optimal Information Storage in Noisy Synapses," presented at *2006 Cold Spring Harbor Laboratory Meeting on Channels, Receptors & Synapses*, Cold Spring Harbor, New York, 18-22 April 2006.

and also appear in the manuscript

- L. R. Varshney, P. J. Sjöström, and D. B. Chklovskii, "Optimal Information Storage in Noisy Synapses under Resource Constraints."

Some of the results of Chapter 4 appear in

- L. R. Varshney and V. K. Goyal, "Ordered and Disordered Source Coding," in *Proceedings of the Information Theory and its Applications Inaugural Workshop*, La Jolla, California, 6-10 February 2006.

- L. R. Varshney and V. K. Goyal, "Toward a Source Coding Theory for Sets," in *Proceedings of the Data Compression Conference (DCC 2006)*, Snowbird, Utah, 28-30 March 2006.

Chapter 2

# General Information Storage Systems

## 2.1 The Problem of Information Storage

The basic problem of information storage and retrieval is the preservation of information generated in the present for use in the future. When stored information is retrieved, a minimum fidelity should be preserved, independent of the way the process of storage and retrieval is constrained by finite resources. Therein lies the difficulty of the storage and retrieval problem: maintaining good fidelity while remaining within resource constraints. The storage and retrieval problem is a physical problem, in that the information must be maintained in some physical medium that can withstand the passage of time. In this thesis, we will cast both fidelity, which is a potentially semantic quality, and resource consumption, which is a physical quality, in mathematical terms. By doing so, we will be able to use mathematical techniques to determine characterizations of the optimal tradeoff between fidelity preservation and resource consumption.

For brevity, in the sequel we refer to the information storage and retrieval problem as the information storage problem. The fundamental nature of the information storage problem is captured in the celebrated Figure 1 of Shannon's "A Mathematical Theory of Communication" [9] (Figure 2-1). Following Figure 2-1, we define an information storage system to have five parts: source, storer, channel, retriever, and destination. We model messages and signals probabilistically and enforce a Markov information pattern consistent with the figure: the retrieved signal is independent of the source message given the stored signal and the destination message is indepen-

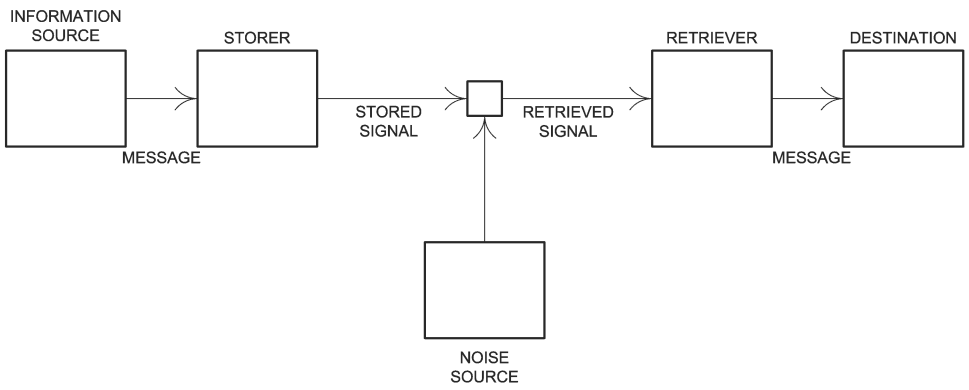


**Figure 2-1.** Schematic diagram of a general information storage system, Figure 1 from Shannon's "A Mathematical Theory of Communication," [9], redrawn and relabeled by author.

dent of the source message and stored signal given the retrieved signal. To illustrate the relationship between the physical/semantic problem of information storage and the mathematical problem of information storage and more fully explicate the mathematical model, we use an example.

Consider the storage process whose end stage is happening now: the storage of the informational content of the thesis over a paper medium. The paper medium is the channel; the information that has been stored is a sequence of ideas, theorems, principles, and concepts. Owing to the novelty of this sequence, there exists some uncertainty that is mitigated when a representation of the sequence is retrieved. We model the information source in probabilistic terms as producing a sequence of random variables, $U_1^k = (U_1, U_2, \ldots, U_k)$ which are drawn from a common alphabet $\mathcal{U}$ according to a probability distribution $p_{U_1^k}(\vec{u})$, defined for all $k$. In contrast to problems in observation and estimation [11, 12, 13], the source side has a storer by which the source message may be transformed to facilitate storage in the physical medium. Since the ensemble of ideas, theorems, and such is not well-suited for storage in paper, the storer physically acts as a transducer which has a mathematical representation as an operator. The operation of the storer is to express ideas in linguistic and mathematical terms, which are then represented with a symbol alphabet with an associated typography. The storer may also filter the message to reduce redundancy. The typography is the physical manifestation of the signal that is stored in the physical storage medium. Although the stored signal is not the medium and should be understood

independently of the medium, a physical manifestation is necessary nonetheless. The storer converts a block of $k$ source symbols into a block of $n$ channel symbols, thus the system has a fixed code rate $\kappa = k/n$. Mathematically, the storer is specified as a sequence of transition probability assignments

$$K_{ST}^{n} = p_{X_1^n|U_1^k,}(\vec{x}|\vec{u}), \tag{2.1}$$

defined for all $n$. The output of the storer is a sequence of channel input random variables, $X_1^n = (X_1, X_2, \ldots, X_n)$ which are drawn from a common alphabet $\mathcal{X}$. The distribution of the channel input letters, $p_{X_1^n}(\vec{x})$, is determined by the source distribution and the storer.

The channel is the physical medium over which storage is carried out; in our example it is the paper. The channel provides the physical link between the source and destination and may be restricted to allow only a finite number of states or may introduce noise or both. Noise may be manifested through degradation of paper fibers, fading of ink, stains, smudges, or dirt. The mathematical model of the channel is defined as a sequence of transition probability assignment between the channel input space and the channel output space, $p_{Y_1^n|X_1^n}(\vec{y}|\vec{x})$, for all $n$. The resource constraints of the information storage system are specified as constraints on what can be stored over the channel. A word cost function is defined on the channel input alphabet $\mathcal{X}^n$. Formally, we require that the word cost function be a nonnegative function $b_n(\vec{x}) : \mathcal{X}^n \mapsto [0, \infty)$. It is desirable for the cost function to be based on some fundamental system resource such as energy, bandwidth, volume, or money. Sometimes it is difficult to determine the cost constraints that limit system functionality [14]. A sequence of word cost functions,

$$E = \{b_n(x_1^n), n = 1, 2, \ldots\}, \tag{2.2}$$

is called a resource criterion. A resource criterion composed of word cost functions of

the form

$$b_n(\vec{x}) = \frac{1}{n}\sum_{i=1}^{n} b(x_i) \tag{2.3}$$

is called a single-letter resource criterion.

In the specification of the channel, we have implicitly defined the channel output random variables, $Y_1^n = (Y_1, Y_2, \ldots, Y_n)$ which are drawn from the common alphabet $\mathcal{Y}$ according to the marginal distribution $p_{Y_1^n}(\vec{y})$, which is determined by $p_{X_1^n}(\vec{x})$ and the channel. In the thesis, we assume that the channel does not introduce insertion or deletion errors and so the number of channel input and output symbols is identical. In reversing the operations of the storer, the retriever converts the retrieved signal back into the form of the original source message. The retriever converts $n$ channel symbols into $k$ reconstruction symbols, where the same code rate $\kappa = k/n$ is used. Mathematically, the retriever is specified as a sequence of transition probability assignments

$$K_{RT}^n = p_{V_1^k|Y_1^n}(\vec{v}|\vec{y}), \tag{2.4}$$

for all $n$. The destination receives a sequence of reconstruction random variables, $V_1^k = (V_1, V_2, \ldots, V_k)$ which are drawn from the common alphabet $\mathcal{V}$. The probability distribution of $\vec{V}$, $p_{V_1^k}(\vec{v})$, is determined by the distribution of $\vec{U}$, the storer, the channel, and the retriever in the Markov fashion. Note that since the channel does not delete or insert symbols and since the code rate is fixed, there are an equal number of source and reconstruction letters, $k$.

For the source-destination pair, we define a word distortion function that measures the quality of the reproduction $\vec{V}$ for the source $\vec{U}$. Formally, we require that the word distortion function be a nonnegative function $d_k(\vec{u}, \vec{v}) : \mathcal{U}^k \times \mathcal{V}^k \mapsto [0, \infty)$. Desirable properties for word distortion functions include such things as tractability, computability, and subjective significance [15]. These desiderata, particularly mathematical tractability and semantic significance, are often difficult to attain simultaneously [16]. A sequence of word distortion functions,

$$F = \left\{ d_k(u_1^k, v_1^k), k = 1, 2, \ldots \right\}, \tag{2.5}$$

is called a fidelity criterion. A fidelity criterion composed of word distortion functions of the form

$$d_k(\vec{u},\vec{v}) = \frac{1}{k}\sum_{i=1}^{k} d(u_i, v_i) \tag{2.6}$$

is called a single-letter fidelity criterion.

We have specified an information storage system by six sequences of measures that are naturally grouped into three pairs [17, 18]. The source-destination pair is specified by the source distribution and the fidelity criterion, $(p_{\vec{U}}(\vec{u}), F)$. The storer and retriever are specified by their transition probability assignments, $(K_{ST}, K_{RT})$. Finally, the channel is specified by the transition probability assignment and resource criterion $(p_{\vec{Y}|\vec{X}}(\vec{y}|\vec{x}), E)$.

The three functionals that are used to measure the performance of an information storage system are the code rate $\kappa$, the average incurred cost,

$$B_n = \mathop{\mathrm{E}}_{p_{X_1^n}(\vec{x})}\left[b(X_1^n)\right], \tag{2.7}$$

and the average incurred distortion

$$\Delta_k = \mathop{\mathrm{E}}_{p_{U_1^k,V_1^k}(\vec{u},\vec{v})}\left[d(U_1^k, V_1^k)\right]. \tag{2.8}$$

Of course, we would prefer a system with large $\kappa$, small $B$, and small $\Delta$. We will give a definition of an optimal communication system in the next section.

In the framework of information theory, information storage is studied as a problem built upon probability theory. As such we have modeled the mechanisms of source message production and channel signal perturbation with probability laws. The actions of the storer and retriever are also described by probability laws. Although there are alternatives to the probabilistic formulation of information theory [19,20,21], these will not be our concern. In the remainder of this chapter, we will follow [22], considering the basic asymptotic blocklength point-to-point information storage problem with synchronous, discrete-space, stationary, ergodic sources and channels. Asynchronous sources [23], channels with synchronization errors [24, 25, 26], and non-stationary,

non-ergodic systems [27, 28, 29] have been considered elsewhere. The conversion of continuous-space storage systems to discrete-space systems by appealing to the sampling theorem and its generalizations [30, 31, 32] is standard.[1]

One might argue that if the same signal is retrieved from the channel multiple times, then the storage system is not point-to-point, but rather a broadcast system. In fact, assuming that channel degradation is non-decreasing as a function of how long the physical channel medium is maintained, the system is a degraded broadcast system: as each increment of time passes, the signal is passed through an incremental channel cascaded to the previous set of channels. In a similar vein, one might envision storing information more than once, as in a multiple descriptions system. Like the degraded broadcast system, presumably this would reduce to a successive refinement system. For simplicity, we do not adopt either of these views in the thesis. The degraded broadcast view would require the characterization of channel quality as a function of time and would involve uncountably many destinations in which to optimize. There do not appear to be many physical characterizations of such channel degradations as functions of time. The storage procedure suggested by the successive refinement view is suboptimal. Rather than providing a new description to supplement an old description, overwriting the stored signal with a new signal would do better in terms of fidelity and resources, in a sense creating a palimpsestic memory system.

In this section, we have taken the physical information storage problem and cast it into a mathematical information storage problem by adopting the information theoretic formulation of information storage. By analogy with the storage of water or electrical potential, one might wonder what the jar or capacitor for information storage is. While it is true that the stored signal can remain in the physical medium of the channel for an indefinite period of time, it is not quite the same as a jar or capacitor, a priori. The signal, rather than some object which can be measured in units of information, is what is stored. Only when we invoke the separation-based approach

[1] Care may be needed to preserve spacing information, a method of modulating signals, see Section 2.6.1.

to optimal information storage will the quantity that is stored in the channel be measurable in units of information. Otherwise, all that we can measure is the amount of distortion that is introduced by each of the stages of the information storage system. The storer, the channel, and the retriever may all introduce or mitigate distortion. Even when we use the separation-based approach and adopt the rate matching interpretation, as in Section 2.4, dissipating entropy and losing mutual information is not necessarily objectionable if it has no effect on average incurred distortion. The loss of entropy is actually beneficial when it reduces the average incurred cost. Thus if the storer performs a computation that dissipates irrelevant entropy, it is beneficial and should not be seen as degrading the quality of the information storage system, as a leaky jug might. The performance of an information storage system is measured in terms of distortion and cost and so an entropic notion of information supply, storage, and dissipation is not universally applicable. In the next section, we define the notion of an optimal information system in terms of fidelity and resources without reference to amount of information.

## ■ 2.2 Operational Definition of Optimal Information Storage Systems

Information storage system optimality is defined in terms of the performance criteria $\kappa$, $B$, and $\Delta$. To make this definition, assume that the code rate $\kappa$ is a fixed system parameter and not subject to optimization.

**Definition 1 (Definition 5 in [17]).** *For the source-destination pair $(p_{\vec{U}}(\vec{u}), F)$, and the channel $(p_{\vec{Y}|\vec{X}}(\vec{y}|\vec{x}), E)$, a sequence of storers and retrievers $(K^n_{ST}, K^n_{RT})$ of fixed code rate $\kappa$ that achieve average incurred costs $\{B_n\}$ and average incurred distortions $\{\Delta_{\kappa n}\}$ are optimal if the following conditions are satisfied.*

1. *There is no sequence of storers and retrievers, $(\hat{K}^n_{ST}, \hat{K}^n_{RT})$, such that the sequence of achieved average incurred costs $\{\hat{B}_n\}$ is less than $\{B_n\}$, while the achieved average incurred distortion $\{\hat{\Delta}_{\kappa n}\} = \{\Delta_{\kappa n}\}$, and*

2. *there is no sequence of storers and retrievers, $(\hat{K}^n_{ST}, \hat{K}^n_{RT})$, such that the sequence of achieved average incurred distortions $\{\hat{\Delta}_{\kappa n}\}$ is less than $\{\Delta_{\kappa n}\}$, while*

*the achieved average incurred cost* $\{\hat{B}_n\} = \{B_n\}$,

*without any restriction on* $(K_{ST}^n, K_{RT}^n)$ *or* $(\hat{K}_{ST}^n, \hat{K}_{RT}^n)$ *other than the Markov information pattern.*

A cost-distortion curve may be drawn which gives all system performances, $(\Delta, B)$, that are optimal. This curve is given by the set of points that simultaneously satisfy two optimization problems derived from the definition of optimality. In particular, the cost-distortion curve is the intersection of points $(\Delta, B)$ that satisfy

$$\Delta(B) = \liminf_{n\to\infty} \underset{(K_{ST}^n, K_{RT}^n):\mathrm{E}_{p_{X_1^n}}\left[b\left(X_1^n\right)\right]\le B}{\mathrm{E}_{p_{U_1^k, V_1^k}}} \left[d\left(U_1^k, V_1^k\right)\right] \tag{2.9}$$

and

$$B(\Delta) = \liminf_{n\to\infty} \underset{(K_{ST}^n, K_{RT}^n):\mathrm{E}_{p_{U_1^k, V_1^k}}\left[d\left(U_1^k, V_1^k\right)\right]\le \Delta}{\mathrm{E}_{p_{X_1^n}}} \left[b\left(X_1^n\right)\right]. \tag{2.10}$$

The distortion-cost notion of system performance can be used to compare various information storage systems to each other and to the optimality bound. Distortion and cost are also fundamental parameters for information systems other than the basic point-to-point scenario and similar optimal performance tradeoffs between these two parameters may be formed. As an example, for the storage of many correlated sources for a single destination, a natural fidelity criterion would use sum distortion measures on the individual components and a natural resource criterion would use sum costs. This formulation would allow comparison of various styles of storer and retriever designs that run the entire gamut from separate source and channel processing [33,34] to joint source and channel processing [35] to uncoded storage [36].

## ■ 2.3 Informational Definition of Optimal Information Storage Systems

Although we have a definition (Definition 1) and a general formula (intersection of (2.9) and (2.10)) for an optimal information storage system, the general formula is given in terms of an optimization problem over a sequence of storers and re-

trievers, which seems rather intractable. One would like to reformulate the problem from this *operational* definition into some *informational* definition in terms of $(p_{\vec{U}}(\vec{u}), F, p_{\vec{Y}|\vec{X}}(\vec{y}|\vec{x}), E)$. To do so, we will invoke the informational functions, capacity-cost and rate-distortion [9, 37], which are defined in terms of the mutual information functional [38].

**Definition 2.** *For each positive integer $n$ and $B \geq 0$, the $n$th capacity-cost function of the channel $p_{Y_1^n|X_1^n}(y_1^n|x_1^n)$ with respect to the resource criterion $E$ is*

$$C_n(B) = \sup \left\{ I(X_1^n; Y_1^n) : \mathrm{E}\left[b_n(X_1^n)\right] \leq B \right\}. \tag{2.11}$$

*The supremum is taken with respect to any $p_{X_1^n}(x_1^n)$ that satisfies the cost constraint. By convention, the $n$th capacity-cost is 0 when no random vectors in $\mathcal{X}^n$ satisfy the constraint.*

**Definition 3.** *The capacity-cost function with respect to the resource criterion $E$ is*

$$C(B) = \limsup_{n\to\infty} \frac{1}{n} C_n(B). \tag{2.12}$$

**Definition 4.** *For each positive integer $k$ and $\Delta \geq \Delta_{min}$, the $k$th rate-distortion function of the source $p_{U_1^k}(u_1^k)$ with respect to the fidelity criterion $F$ is*

$$R_k(\Delta) = \inf \left\{ I(U_1^k, V_1^k) : \mathrm{E}\left[d_k(U_1^k, V_1^k)\right] \leq \Delta \right\}. \tag{2.13}$$

*The infimum is taken with respect to any $p_{U_1^k, V_1^k}(u_1^k, v_1^k)$ that satisfies the distortion constraint.*

The constant $\Delta_{min}$ is the minimum possible value of average incurred distortion under the fidelity criterion.

**Definition 5.** *The rate-distortion function with respect the fidelity criterion $F$ is*

$$R(\Delta) = \liminf_{k\to\infty} \frac{1}{k} R_k(\Delta). \tag{2.14}$$

Now that we have defined some basic informational functions, we can formulate the information storage theorems that yield an informational formulation of optimal information storage systems.

**Theorem 1 (Converse Information Storage Theorem [22]).** *The performance parameters, $B$, $\Delta$, and $\kappa$, of an information storage system must satisfy $\kappa R(\Delta) \leq C(B)$.*

*Proof.*

$$I(\vec{U};\vec{V}) \overset{(a)}{\leq} I(\vec{X};\vec{Y}) \tag{2.15}$$

$$I(\vec{X};\vec{Y}) \overset{(b)}{\leq} C_n(B) \tag{2.16}$$

$$C_n(B) \overset{(c)}{\leq} nC(B) \tag{2.17}$$

$$I(\vec{U};\vec{V}) \overset{(d)}{\geq} R_k(\Delta) \tag{2.18}$$

$$R_k(\Delta) \overset{(e)}{\geq} kR(\Delta) \tag{2.19}$$

Inequality (a) follows by the Markov relation $\vec{U} \to \vec{X} \to \vec{Y} \to \vec{V}$ implicit in Figure 2-1 and the data processing inequality [38, Theorem 2.42]; inequality (b) is by Definition 2; inequality (c) is by Definition 3; inequality (d) is by Definition 4; and inequality (e) is by Definition 5. Combining these five inequalities

$$kR(\Delta) \leq R_k(\Delta) \leq I(\vec{U};\vec{V}) \leq I(\vec{X};\vec{Y}) \leq C_n(B) \leq nC(B) \tag{2.20}$$

yields the desired result

$$\kappa R(\Delta) \leq C(B). \tag{2.21}$$

□

Note that the converse information storage theorem is simply a consequence of the definitions of the capacity-cost and rate-distortion information functions and the mutual information functional (through the data processing inequality). There has been no consideration of retrievers and storers yet.

To establish an achievability result, we must design the retriever and storer, and the code that they share. Following Shannon [9], we will partition both the storer and the retriever into two parts: a source encoder and channel encoder at the storer, and a channel decoder and source decoder at the retriever. The modularity introduced by this partitioning will simplify the design of the storer and retriever as well as the analysis of the system performance (allowing the use of the channel coding and source coding theorems). Separation, however, is not the only possible storer and retriever design.

Define an $(n, M_f)$ channel code to consist of a channel encoding operation $f_{ST} : \{1, 2, \ldots, M_f\} \to \mathcal{X}^n$ and a channel decoding operation $f_{RT} : \mathcal{Y}^n \to \{1, 2, \ldots, M_f\}$. The sequences $\{f_{ST}(1), \ldots, f_{ST}(M_f)\}$ are called channel codewords and the set of all channel codewords is called the channel codebook. The rate of the code is defined in terms of the size of the codebook, $M_f$, and blocklength $n$, as $r_f = \frac{1}{n} \log M_f$. The code is termed $B$-admissible with respect to the resource criterion $E$ if $b_n(\vec{x}) \leq B$ for all codewords in the codebook. When we use the channel code, we will use the codewords with equal frequency, hence the logarithmic rate measure and the cost constraint imposed on each codeword. Notice, however, that using codewords with equal frequency does not imply that channel letters will be used with equal frequency. For convenience, a channel code performance criterion is introduced, though it is not fundamental to the information storage problem.[2]

**Definition 6.** *The conditional probability of error given that the message $i$ is sent, for $1 \leq i \leq M_f$ is*

$$\lambda_i = \sum_{\vec{y} \in \mathcal{Y}^n : f_{RT}(\vec{y}) \neq i} \Pr[\vec{Y} = \vec{y} | \vec{X} = f_{ST}(i)]. \tag{2.22}$$

**Definition 7.** *The maximum probability of error of the channel code is*

$$\lambda_{max} = \max_i \lambda_i. \tag{2.23}$$

[2]Similarly, other channel code performance criteria such as zero-error reliability [39] or anytime reliability [27, 40] are not fundamental to the information storage problem either.

The channel coding theorem is an achievability statement about maximum probability of error for a sequence of $(n, M_f)$ channel codes for asymptotically large $n$.

**Theorem 2 (Channel Coding Theorem [22]).** *Let a channel with capacity-cost function $C(B)$ be given and fix a value $B_0$. Then for any three values $B > B_0$, $R > C(B_0)$, and $\epsilon > 0$, for all sufficiently large $n$, there exists an $(n, M_f)$ channel code such that the code is $B$-admissible, the size of the codebook is lower bounded as $M_f \geq 2^{\lceil nR \rceil}$, and the maximum probability of error is upper bounded as $\lambda_{max} < \epsilon$.*

*Proof.* The proof is based on a random coding argument. See standard texts, e.g. [22], for details. See [41] for a very general version. □

Given the existence of good channel codes (in the sense of the channel coding theorem), the existence of good source codes must also be established, so that the combination produces good storers and retrievers. Define a $(k, M_g)$ source code to consist of a source encoding operation $g_{ST} : \mathcal{U}^k \to \{1, 2, \ldots, M_g\}$ and a source decoding operation $g_{RT} : \{1, 2, \ldots, M_g\} \to \mathcal{V}^k$. The sequences $\{g_{RT}(1), \ldots, g_{RT}(M_g)\}$ are called source codewords and the set is called the source codebook. The rate of the code is defined in terms of the size of the codebook, $M_g$, and blocklength $k$, as $r_g = \frac{1}{k} \log M_g$. The average incurred distortion of the source code is defined in the same manner as the average incurred distortion of an information storage system. The source coding theorem is an achievability statement about average incurred distortion for a sequence of $(k, M_g)$ source codes for asymptotically large $k$.

**Theorem 3 (Source Coding Theorem).** *Let a source destination pair with rate-distortion function $R(\Delta)$ be given and fix a $\Delta_0 > \Delta_{min}$. Then for any two values $\Delta > \Delta_0$ and $R > R(\Delta_0)$, for all sufficiently large $k$, there exists a $(k, M_g)$ source code such that the size of the codebook is upper bounded as $M \leq 2^{\lfloor kR \rfloor}$, and the average incurred distortion $\Delta_g$ is upper bounded as $\Delta_g < \Delta$.*

*Proof.* The proof is based on a random coding argument. See standard texts, e.g. [22], for details. See [41] for a very general version. □

The direct part of the information storage theorem is based on using a source encoder guaranteed to exist by the source coding theorem to produce an index into the set $\{1, 2, \ldots, M\}$ and then applying a channel encoder guaranteed to exist by the channel coding theorem to produce the channel input stream. After the channel perturbations, the corresponding channel decoder and source decoder are used.

**Theorem 4 (Direct Information Storage Theorem [22]).** *Let an information storage system consist of a source, channel, and destination with capacity-cost function $C(B)$ and rate-distortion function $R(\Delta)$ and fix $B_0$, $\Delta_0 > \Delta_{min}$, and $\kappa_0 < C(B_0)/R(\Delta_0)$. Then for sufficiently large $k$, there exists an information storage system that can achieve $(B, \Delta, \kappa)$ for any three values $B \leq B_0$, $\Delta \leq \Delta_0$, and $\kappa \geq \kappa_0$.*

*Proof.* The proof here is based on the proof given for [22, Theorem 5.1.b]. Select numbers $\beta_0$, $\delta_0$, $\delta_1$, $C_0$, and $R_0$ that satisfy the following relations $\beta_0 < B_0$, $\Delta_{min} \leq \delta_0 < \delta_1 < \Delta_0$, $C_0 < C(\beta_0)$, $R_0 > R(\delta_0)$, $\kappa_0 < C_0/R_0$.

The storer is composed of a source encoder followed by a channel encoder. By Theorem 3, for sufficiently large $k_0$, there is a $(k_0, M_g)$ source code where $M_g \leq 2^{\lfloor k_0 R_0 \rfloor}$ and $\Delta_g < \delta_1$. Use this source code. For a certain integer $m$, take $k = k_0 m$.

Define the worst-case distortion of the source code. For each source sequence $\vec{u} \in \mathcal{U}^{k_0}$, let $d_{max}(\vec{u}) = \max\{d_{k_0}(\vec{u}, \vec{v}_i) : \vec{v}_i \in \{g_{RT}(1), \ldots, g_{RT}(M_g)\}\}$. The worst-case distortion for the source code is the average of these:

$$\Delta_g^{max} = \mathop{\mathrm{E}}_{p_{\vec{U}}(\vec{u})}[d_{max}(\vec{U})]. \tag{2.24}$$

Note that the worst-case distortion is bounded, since $d_k(\cdot)$ has range $[0, \infty)$. Now define

$$\epsilon = \frac{\Delta_0 - \delta_1}{\Delta_g^{max}}, \tag{2.25}$$

which will relate the probability of error of the channel code that we design to the amount of end-to-end distortion it may cause.

For each $m = 1, 2, \ldots$, define an integer $n_m = \lceil m k_0 R_0 / C_0 \rceil$. Then by Theorem 2, there is a $(n_m, M_f)$ channel code such that all codewords in the codebook meet the

cost constraint, $M_f \geq 2^{\lceil n_m C_0 \rceil} \geq 2^{mk_0 R_0}$, and $\lambda_{max} < \epsilon$. Use this channel code. Also take $m$ sufficiently large so $k_0 m / n_m \geq \kappa_0$.

Now it follows directly from the choice of $m$ that $\kappa = k/n = k_0 m / n_m \geq \kappa_0$. It also follows directly from the design of the channel code that the channel cost constraint is met. It remains to show that the average distortion incurred $\Delta \leq \Delta_0$. Distortion is incurred through channel decoding failure and through the distortion introduced by the source code. Use the random variable $F_{fail}$ to indicate channel code failure. Then the distortion can be decomposed as

$$\Delta = \mathop{\mathrm{E}}_{p_{\vec{U},\vec{V}}}[d_k(\vec{U},\vec{V})|F_{fail}=0]\Pr[F_{fail}=0] + \mathop{\mathrm{E}}_{p_{\vec{U},\vec{V}}}[d_k(\vec{U},\vec{V})|F_{fail}=1]\Pr[F_{fail}=1]. \tag{2.26}$$

By the design of the source code and the bound $\Pr[F_{fail}=0] \leq 1$, we get

$$\mathop{\mathrm{E}}_{p_{\vec{U},\vec{V}}}[d_k(\vec{U},\vec{V})|F_{fail}=0]\Pr[F_{fail}=0] < \delta_1. \tag{2.27}$$

By the bound on the frequency of error of the channel code, the worst-case performance of the source code, and (2.25),

$$\mathop{\mathrm{E}}_{p_{\vec{U},\vec{V}}}[d_k(\vec{U},\vec{V})|F_{fail}=1]\Pr[F_{fail}=1] \leq \lambda_{max}\Delta_g^{max} \leq \epsilon\Delta_g^{max} = \Delta_0 - \delta_1. \tag{2.28}$$

Adding the bounds on the two terms, we get

$$\Delta < \Delta_0. \tag{2.29}$$

□

Finally, we can present the informational definition of an optimal information storage system, which is equivalent to the operational definition given in Section 2.2.

**Theorem 5 (Information Storage Theorem [22]).** *An information storage system is optimal if and only if*

$$\kappa R(\Delta) = C(B), \tag{2.30}$$

*$\Delta$ cannot be decreased without increasing $R(\Delta)$, and $B$ cannot be decreased without decreasing $C(B)$.*

*Proof.* Follows directly from Theorem 1 and Theorem 4. □

## ■ 2.4 Rate Matching Interpretation of Optimal Information Storage Systems

Since the storer in the achievability proof comprises separate source and channel coding, Theorem 4 is often called the separation theorem. Since the proof is by a particular construction of storer and retriever, it does not preclude other constructions that also yield the achievability result. In particular, methods of joint source channel coding, which we will discuss in later sections, may also be used to show achievability. The separation of source and channel coding yields a *rate matching* characterization of system optimality; it requires the rate-distortion of the source, a mutual information quantity that measures information rate, to be equal to the capacity-cost of the channel, another mutual information quantity that measures information rate.

The separation-based approach concentrates essentially all incurred distortion into the source code. In an information storage system, the end-to-end distortion that is incurred may be divided into three parts:

$$\Delta = \Delta_{Source} + \Delta_{Mixed} + \Delta_{Channel}. \tag{2.31}$$

The $\Delta_{Source}$ term is the amount of distortion introduced if the storer has range space smaller than domain space. In the separation-based approach, this is the distortion caused by the source code, the $\mathrm{E}[d_k(\vec{U}, \vec{V})|F_{fail} = 0]\Pr[F_{fail} = 0]$ term in (2.26) in the proof of Theorem 4. The $\Delta_{Channel}$ term is the amount of distortion introduced by unmitigated channel noise, the $\mathrm{E}[d_k(\vec{U}, \vec{V})|F_{fail} = 1]\Pr[F_{fail} = 1]$ term in (2.26). The $\Delta_{Mixed}$ term is the remaining distortion and arises from the interaction between the several parts of the system; it is zero for optimal systems, including cases where the source code is optimally designed for a noiseless channel [42] or the system is uncoded. As can be noted from the proof, the goal of the separation-based approach is to use channel coding to make the channel essentially deterministic, by driving

$\lambda_{max}$ to be arbitrarily small, so that $\Delta_{Channel}$ is essentially zero, and consequently $\Delta \approx \Delta_{Source}$. Alternative optimal information storage systems have other values of $\gamma$ in decomposing $\Delta = \gamma\Delta_{Source} + (1-\gamma)\Delta_{Channel}$. In uncoded storage, the storer is an identity map, therefore $\Delta_{Source} = 0$ and $\Delta = \Delta_{Channel}$. A side benefit of the separation-based approach to information storage system optimality is the duality between the design of source codes and channel codes that is induced [37, 43].

As a consequence of the rate matching interpretation of optimal information storage, information may be understood in the same manner as fluid flow. Although fluids appear in many of the idioms of electrical engineering theory (the "waterbed effect" in Bode's integral formula, "waterfilling" to achieve capacity in colored noise, or "digital fountains" for multicasting), here the fluid paradigm is often taken to be an ontological description of information. The paradigm of electricity as a water-like fluid has been central to the development of electrical engineering theory, and has allowed connections with other branches of engineering theory such as acoustics, mechanics, thermics, and hydraulics [44, 45, 46]. Since information theory is arguably a branch of electrical engineering theory,[3] it is not surprising that a similar characterization of information would be sought, allowing problems of information storage to yield to previously developed methods of analysis.

In many cases, the fluid framework for electricity led to fruition, but in other cases, this framework led down the wrong path. Very briefly, we can look at the development of electrophysiology as an example of the double-edged nature of the fluid paradigm. This will also serve to introduce the tradition of interaction between neurobiology and electrical engineering theory, of which our contributions in Chapter 3 are a part. It was long held that there was some curious fluid, referred to as "animal spirits," that was involved in muscle contraction and nerve conduction. In the late eighteenth century, Galvani was able to show that this fluid is electrical in nature [47]. The fluid paradigm had also become well-established in another branch

[3]For a discussion of the sociological process that took place in defining the boundaries and proper place of information theory, see [20]; for a discussion of the role played by mathematicians and electrical engineering theorists in the early post-1948 period, see [19].

of electrical science, that dealing with electrical machines and Leyden jars.[4] After Galvani's initial discovery, a great debate developed surrounding the nature of this electrical fluid. Galvani held to the belief that animals possessed intrinsic "animal electricity" (animalis electricitas), whereas Volta eventually came to lead the argument that animals were simply conductors of the electrical fluid and the source of electricity resided in metals [49]. As this debate raged, the fluid paradigm was central to steps forward and to missteps alike. The validity of the fluid paradigm in this electrophysiological situation can be essentially summarized by two facts—when flowing through materials, electricity is very much a fluid; at a distance, water-like fluids cannot act on other water-like fluids, whereas electrical flow can cause distant electrical flow through electrical induction. As Pierce points out [46]:

> ... Maxwell's equations cover the behavior not only of idealized electrical networks but of all physical structures and include the behavior of radio waves, which lies outside of the scope of network theory.

Volta's theory and the voltaic pile batteries that came about from his experiments would come to dominate. To complete the story, Matteucci through experiments such as depicted in Figure 2-2 would eventually confirm that animals have intrinsic Galvanic fluid and this would form the basis of the modern theory of neuroscience, described in Chapter 3.

Returning our attention from electricity, where the fluid paradigm led to Kirchoff's laws and essentially all forms of circuit analysis, to information, one can see evidence of the fluid perspective right from the early period [19]:

> As Jerome Wiesner, director of the Massachusetts Institute of Technology's Research Laboratory of Electronics said in 1953, "Before we had the theory,... we had been dealing with a commodity that we could never see or really define. We were in the situation petroleum engineers would

[4]This can be seen very simply using an etymologically oriented argument: the name "jar" applied to a device for electricity storage implies that this storage was considered to be akin to the storage of fluids in jars. It can be considered as a particular instance of the general phenomenon of interpreting new technologies in terms of old ones [48].

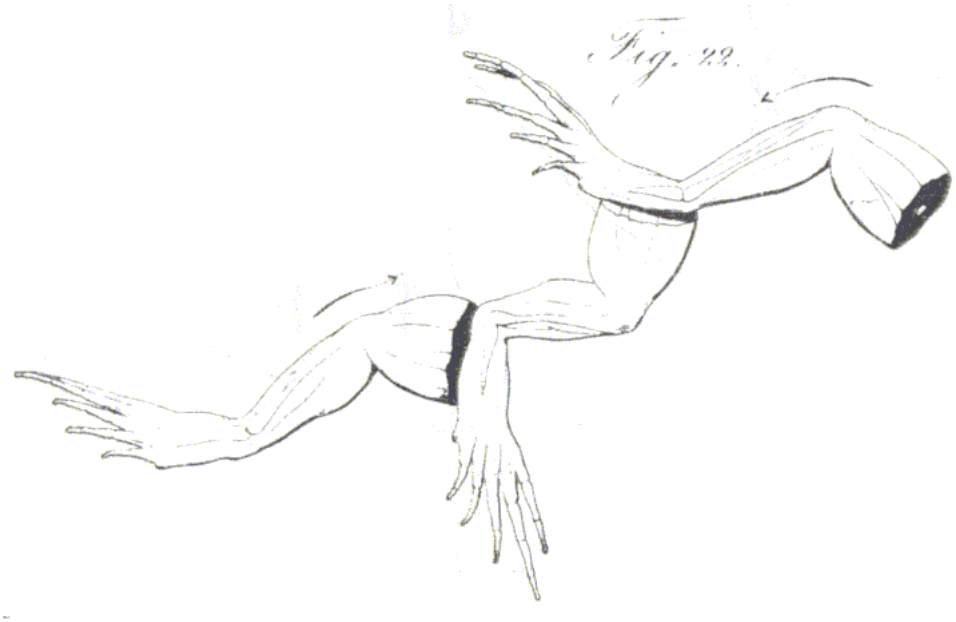


**Figure 2-2.** Experiment to verify the electrical nature of neural activity. A pile of frog legs, Figure 22 from Matteucci's *Traité des Phénomènes Électro-physiologiques des Animaux* [50].

> be in if they didn't have a measuring unit like the gallon. We had intuitive feelings about these matters, but we didn't have a clear understanding."

as well as through to modern views [51], such as [52]:

> "A bit is a bit is a bit" seems to have become the motto of this digital age. Though we often take it for granted, the existence of such a common currency for information is far from obvious. While in practice it has been justified by the rapid advancement of digital electronics and the rise of a communications infrastructure to match, its meaningfulness rests upon the fundamental theorems of information theory—the philosophical foundation that helps us gain insight into problems of communication.

This feeling that the bit is a universal measure of information, like the gallon is for water, has often led to fruition. Many such examples include problems where there is a single destination, with perhaps multiple sources [53, 34, 54]. Furthermore, the rate matching idea, which is based on equating rate to capacity and uses the concept of bit as a fundamental unit of information has been applied to great success in practical information storage systems [55]. However, just as in electricity, where wave

propagation was not explained by the fluid paradigm, there are instances in information theory where the water perspective does not lead to correct theory. Notable examples include the non-ergodic problem [28], the multicast problem [56], and other multiterminal problems [18]. To use the terminology of electricity theory, the *field* problem cannot be reduced to a *circuit* problem [44].

When the fluid paradigm is valid, the optimal information storage problem reduces to a circuit problem of equating the maximum information flow through the circuit to the minimum cut of the network capacity links, by invoking the celebrated max flow min cut theorem [57, 53]. The max flow min cut theorem is in fact another form of more general minimax problems in graph theory. A minimax theorem in graph theory, such as König's Theorem, is an identity of the form

$$\max_p v(G,p) = \min_q w(G,q), \tag{2.32}$$

where $G$ is a graph, $p$ and $q$ are some parameters of the graph, and $v$ and $w$ are functionals of the graph [58]. For the basic information storage theorem that we have been considering, the graph, $G$, is the Markov graph $U \to X \to Y \to V$, the functional $v$ is the mutual information $I(X;Y)$, the functional $w$ is mutual information $I(U;V)$, and the parameters for optimization are the storer and retriever, expressed in single-letter form as $p_{X|U}(x|u)$ and $p_{V|Y}(v|y)$. Information storage problems where rate matching does not quite apply also may have graph theoretic minimax characterizations [59]. In this general minimax characterization, the fact that $I(X;Y)$ is determined by the channel input distribution $p_X(x)$, and that $I(U;V)$ is determined by the transition probability assignment $p_{V|U}(v|u)$ will be used in Section 2.5. Prior to that, we present an example of rate matching.

### ■ 2.4.1 An Example of Rate Matching

This example will introduce a source with a context-dependent fidelity criterion, which is often good for realistic sources. The channel will be closely related to channel models for optical and magnetic recording, and thus is quite appropriate for our

information theoretic investigation of information storage.

First establish the rate-distortion point, $R(0)$, for the following source and fidelity criterion. Let the source alphabet and reconstruction alphabets, $\mathcal{U}$ and $\mathcal{V}$ both be binary, and the source distribution be arbitrary. A local distortion function of span $s$ is any function $\rho_s$: $\mathcal{U}^s \times \mathcal{V}^s \mapsto [0, \infty)$. The fidelity criterion generated from this local distortion function (defined only for $k \geq s$) has a sequence of distortion functions

$$d_k\left(\vec{u}, \vec{v}\right) = \sum_{i=1}^{k-s+1} \rho_s\left(u_i^{i+s-1}, v_i^{i+s-1}\right), \tag{2.33}$$

the sliding sum of $\rho_s$ [37]. Here we consider a local distortion function of span 3 that penalizes for either two consecutive errors or for an error pattern error–no error–error. This is very closely related to a source and fidelity criterion considered in [60]. The matrix form of the local distortion measure is given as

$$[\rho_3\left(\vec{u}, \vec{v}\right)] = \begin{array}{c} \\ \\ \\ \\ 0\ 0\ 0 \\ 0\ 0\ 1 \\ 0\ 1\ 0 \\ 0\ 1\ 1 \\ 1\ 0\ 0 \\ 1\ 0\ 1 \\ 1\ 1\ 0 \\ 1\ 1\ 1 \\ \vec{u} \end{array} \begin{array}{c} \begin{array}{cccccccc} 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{array} \\ \left[\begin{array}{cccccccc} 0 & 0 & 0 & d & 0 & d & d & d \\ 0 & 0 & d & 0 & d & 0 & d & d \\ 0 & d & 0 & 0 & d & d & 0 & d \\ d & 0 & 0 & 0 & d & d & d & 0 \\ 0 & d & d & d & 0 & 0 & 0 & d \\ d & 0 & d & d & 0 & 0 & d & 0 \\ d & d & 0 & d & 0 & d & 0 & 0 \\ d & d & d & 0 & d & 0 & 0 & 0 \end{array}\right] \\ \vec{v} \end{array} \tag{2.34}$$

for some $d > 0$.

As shown in [60], the value for $R(0)$ can be found by enumerating binary strings that avoid the patterns that induce non-zero distortion. In particular, we want to enumerate binary strings that avoid the patterns "11" and "101." By [61] or [62,

Theorem 24], we can find the generating function for this enumeration to be

$$S_{11,101}(x) = \frac{x^2 + x + 1}{-x^3 - x + 1}. \tag{2.35}$$

From the generating function, we find that the number of binary strings that do not have "11" or "101" as substrings satisfies the recursion

$$a_k = a_{k-1} + a_{k-3}, \tag{2.36}$$

with initial conditions $a_0 = 1$, $a_1 = 2$, and $a_2 = 3$. This sequence is very closely related to the second Meru sequence (a shifted version of it) as well as the second Meru constant[5]

$$\text{meru}_2 = \frac{1}{3}\left(\sqrt[3]{\frac{29 + 3\sqrt{93}}{2}} + \sqrt[3]{\frac{29 - 3\sqrt{93}}{2}} + 1\right) = 1.465571\ldots \tag{2.37}$$

Using the equivalent of Binet's formula[6] for the second Meru sequence in the achievability proof, we can get the following:

**Theorem 6.** *For the context-dependent fidelity criterion generated by $\rho_3$ above, the rate-distortion point $R(0) \leq \log(2/meru_2)$ for all binary sources, with equality holding when the probability of all binary $n$-strings is strictly positive for all $n$.*

*Proof.* A line by line recreation of the proofs of Theorems 4 (direct) and 5 (converse) in [60] suffice. The only modification is we let $f_n$ denote the number of strings that have neither "11" nor "101" as substrings. Then $n^{-1} \log f_n \to \log \text{meru}_2$ as $n \to \infty$. □

The achievability proof is based on a random coding argument. Without loss of

[5]Following [63], we use the name $\text{meru}_2$ for this constant in reference to the fact that the constant is associated with Piṅgala's Meru prastāra [64], also known as Pascal's triangle. The relationship between the prastāra and the constants is explicated in [65], but seems to have been known to Nārāyaṅa [64]. The first Meru constant is also known as the golden ratio and appears in the rate-distortion theorem in [60].

[6]Binet's formula generates the first Meru sequence (Fibonacci sequence) and allows easy calculation of the first Meru constant (golden ratio)

generality, to evaluate error frequency, storage of only the all zero string can be used. Then the enumeration given by the second Meru sequence is used.

The converse proof is also based on a standard argument. It is noted that the $R_k(0)$ is the same for any source distribution that satisfies the condition that the probability of all binary $k$-strings is greater than zero. Thus, without loss of generality, the source distribution can be assumed to be equiprobable and memoryless.

As seen, finding the rate-distortion function at zero distortion for context dependent rate-distortion measures that have local distortion measure matrices with many zeros can be reduced to a string enumeration problem. It should be noted that although this example reduced to enumeration of various strings and probabilistic descriptions did not arise, the reduction of the Shannon problem to a Hartley problem [66] does not usually happen in rate matching. The fact that $R(0)$ problems may turn into enumeration problems will also be exploited in Chapter 4.

Second establish the capacity-cost point, $C(0)$, for the following channel and resource criterion. The result will be closely related to previous results on the capacity of discrete noiseless channels [9,67], and in particular to the storage capacity of magnetic and optical disks and tapes [68]. Let the channel input and output alphabets $\mathcal{X}$ and $\mathcal{Y}$ both be binary. Moreover, let the channel be noiseless. A local cost function of span $s$ is any function $\upsilon_s$: $\mathcal{X}^s \mapsto [0, \infty)$. The resource criterion generated from this local cost function (defined only for $n \geq s$) has a sequence of cost functions

$$b_n(\vec{x}) = \sum_{j=1}^{n-s+1} \upsilon_s\left(x_j^{j+s-1}\right), \tag{2.38}$$

the sliding sum of $\upsilon_s$. Here we consider a local cost function of span 3 that charges for either two consecutive zeros or for a pattern of zero–one–zero, whereas other patterns

incur no cost

$$\upsilon_3(\vec{x}) = \begin{cases} b, & \vec{x} = 000 \\ b, & \vec{x} = 001 \\ b, & \vec{x} = 010 \\ 0, & \vec{x} = 011 \\ b, & \vec{x} = 100 \\ 0, & \vec{x} = 101 \\ 0, & \vec{x} = 110 \\ 0, & \vec{x} = 111 \end{cases} \tag{2.39}$$

for some $b > 0$. By a form of duality, the enumeration used for the rate-distortion problem can be used for the capacity-cost problem as well. To meet the cost constraint for $C(0)$, only channel inputs that use zero-cost channel input sequences will be used. Moreover, to maximize the rate, these will be used equiprobably. Thus the problem will reduce to an enumeration of zero-cost strings. The enumeration of the strings with non-zero cost is given by the second Meru sequence. This is also the number of 2-limited binary sequences for runlength-limited recording channels [69, Table I: d = 2].

**Theorem 7.** *For the context-dependent resource criterion generated by $\upsilon_3$ above, the capacity-cost point $C(0) = \log(2/meru_2)$, with each $(B = 0)$-admissible channel input codeword used equiprobably.*

*Proof.* Since we require each codeword in the codebook to be $B$-admissible, here for $B = 0$, our optimization is so restricted. Since the channel is noiseless, the optimal input distribution should use each admissible codeword equiprobably to maximize rate. Then the desired result follows directly from the previous enumeration. □

Finally, we can see that the information storage system specified in this example is optimal for $\kappa = 1$ since $R(\Delta) = C(B) = \log(2/\mathrm{meru}_2)$ for some $\Delta$ and $B$, and neither $\Delta$ nor $B$ can be lowered.

## 2.5 Factoring Interpretation of Optimal Information Storage Systems

Theorem 5 shows the equivalence of the operational and the informational definitions of optimal information storage. The theorem is stated in a form that is conducive to the rate matching interpretation given in the previous section. There are, however, alternate interpretations that are based on the probability measures that parameterize the mutual information functionals rather than the values of these functionals themselves. Note that although the pair of codes that were used in the proof of Theorem 4 are fixed, when viewed with the single-letter characterization, they appear to be governed by a probabilistic description.

Our original operational definition of an optimal information storage system was based on two main system performance parameters, $B$ and $\Delta$. For a given resource criterion, $E$, it can be noted that $B$ only depends on the marginal channel input distribution $p_X(x)$. Similarly, for a given fidelity criterion, $F$, $\Delta$ only depends on the marginal source-destination distribution $p_{U,V}(u,v)$ and in particular on $p_{V|U}(v|u)$ since the source distribution $p_U(u)$ is given. This leads to the conclusion that information storage system optimality is simply a matter of having the correct marginal distributions. That is to say, we require two conditions for an information system to be optimal [17, 70, 71].

**Condition 1.** *The channel input distribution should be the capacity-achieving input distribution for the channel $p_{Y|X}(y|x)$:*

$$p_X(x) = \arg\max_{q_X(x)} I(X;Y) = \arg\max_{q_X(x)} I(q_X(x), p_{Y|X}(y|x)). \tag{2.40}$$

**Condition 2.** *The source-destination conditional distribution should be the rate-distortion achieving forward test channel for the source $p_U(u)$:*

$$p_{V|U}(v|u) = \arg\min_{q_{V|U}(v|u)} I(U;V) = \arg\min_{q_{V|U}(v|u)} I(p_U(u), q_{V|U}(v|u)). \tag{2.41}$$

With these two conditions and the given source distribution and channel transition probability assignment, we can formulate the design of an optimal information storage

system as a problem of factoring probabilistic kernels. This is most easily formulated in the memoryless case. We want to choose the storer, $p_{X|U}(x|u)$, and retriever, $p_{V|Y}(v|y)$ so that

$$p_X(x) = \int_{\mathcal{U}} p_U(u) p_{X|U}(x|u) du \tag{2.42}$$

and

$$p_{U,V}(u,v) = \int_{\mathcal{X},\mathcal{Y}} p_U(u) p_{X|U}(x|u) p_{Y|X}(y|x) p_{V|Y}(v|y) dx dy, \tag{2.43}$$

where the integrals are understood to be sums when appropriate for the alphabets. In the separation-based approach, as given in the proof of Theorem 4, asymptotically long codewords are used to achieve the correct matchings. Moreover, intermediate message index spaces, denoted $\mathcal{M}$ and $\hat{\mathcal{M}}$, are introduced between source coding and channel coding. Thus, the source-destination distribution is given by

$$p_{U,V}(u,v) = \int_{\mathcal{M},\mathcal{X},\mathcal{Y},\hat{\mathcal{M}}} p_U(u) p_{M|U}(m|u) p_{X|M}(x|m) p_{Y|X}(y|x) p_{\hat{M}|Y}(\hat{m}|y) p_{V|\hat{M}}(v|\hat{m}) dm dx dy d\hat{m}. \tag{2.44}$$

The channel input distribution is given by

$$p_X(x) = \int_{\mathcal{U},\mathcal{M}} p_U(u) p_{M|U}(m|u) p_{X|M}(x|m) du dm. \tag{2.45}$$

The design of the information storage system, i.e. the determination of the unknown factors in the two product expressions, is often easier in this separated way by using the asymptotics. To reduce the complexity of long codes, one might try restriction to linear codes, but this is hopeless (in the unlimited resource, lossless information storage problem, however, this restriction does not preclude achievability [72]). The use of long, separate codes, however, is not required, and joint source channel codes can be designed directly. In fact, if $p_U(u) = p_X(x)$ and $p_{V|U}(v|u) = p_{Y|X}(y|x)$, then uncoded storage is optimal. If the source distribution is already the capacity-cost achieving channel input distribution and the channel distribution is already the rate-distortion achieving forward test channel, optimality is achieved by directly sending

source outputs through the channel and using the channel outputs as the reconstruction [73, 18, 17, 14].

This factoring problem, like the factorization of integers, seems difficult; even more so, since the optimal $p_X(x)$ and $p_{U,V}(u,v)$ are usually difficult to find. Due to the difficulty of the problem, the optimality of uncoded information storage will be investigated in Section 2.6. Prior to that, we present an example of optimality through factorization.

### ■ 2.5.1 An Example of Factoring

Consider an information storage system that has a memoryless Gaussian source and an additive white Gaussian noise (AWGN) channel. The resource criterion restricts average signal power to $B$ and the fidelity criterion requires average mean square error of at most $\Delta$. AWGN channels often arise as simple models of magnetic recording channels [74, 68, 75], and are also a rough model of synaptic storage channels (Section 3.3). The storage of a Gaussian source over an AWGN channel with quadratic cost and distortion has been discussed numerous times in the past, see e.g. [76, 17, 73, 22], [77, pp. 100–101], [78, Section II.E].

It can be noted that the set of jointly Gaussian distributions is closed under linear operations such as addition, scalar multiplication, and integration, and our factoring will result in probabilistic transformations that stay within this set.

It is well known that the rate-distortion function for a Gaussian source with mean zero and variance $P$ under squared error distortion is

$$R(\Delta) = \begin{cases} \frac{1}{2}\log_2\left(\frac{P}{\Delta}\right), & \Delta < P \\ 0, & else \end{cases} \tag{2.46}$$

and the capacity-cost function for an AWGN channel with noise variance $\sigma^2$ under quadratic cost is

$$C(B) = \frac{1}{2}\log_2\left(1 + \frac{B}{\sigma^2}\right). \tag{2.47}$$

Taking $\kappa = 1$, equating these two expressions yields the fact that the smallest achiev-

able distortion at cost $B = P$ is

$$\Delta = \frac{P\sigma^2}{P + \sigma^2}, \tag{2.48}$$

by Theorem 5.

To invoke factoring, two other well known facts are used. First that the capacity-achieving input distribution for an AWGN channel under power constraint $B$ is a Gaussian distribution with variance $B$. Second that the rate-distortion achieving reverse test channel for a Gaussian source under distortion constraint $\Delta$ is an AWGN channel with noise variance $\Delta$. Due to the closure property of jointly Gaussian distributions under conditioning and Bayes' rule, since the reverse test channel is AWGN, the rate-distortion achieving forward test channel is also AWGN. So we want $p_X(x)$ to be Gaussian with the correct power. The source is already Gaussian with the correct power, so if we were to use it uncoded, one of the factoring conditions would be met. If the channel input and the channel noise are both independent Gaussian, then the channel input and output would be jointly Gaussian, as desired. The distortion incurred would be that desired for optimality with a retriever that scales the channel output, so almost uncoded storage seems optimal.

Consider the information storage system shown in Figure 2-3. The source/channel input is distributed as $U = X \sim \mathcal{N}(0, P)$; the independent additive noise is distributed as $N \sim \mathcal{N}(0, \sigma^2)$; the channel input cost function is quadratic; and distortion is the squared difference. Computing the performance of this scheme, we find that

$$B = P, \tag{2.49}$$

and

$$\Delta = \frac{P\sigma^2}{P + \sigma^2}, \tag{2.50}$$

which is the best possible, so we have achieved an optimal information storage system by factoring with uncoded storage, without using the separation-based achievability method in the proof of Theorem 4.

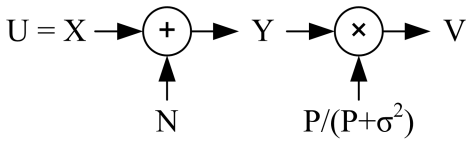


**Figure 2-3.** Optimal quadratic Gaussian information storage system

## 2.6 Measure Matching Interpretation of Optimal Information Storage Systems

In Section 2.5.1, almost uncoded storage resulted in an optimal information storage system with a simple achievability scheme.[7] The reason that this happened was that the redundancy in the source was exactly matched to the redundancy required to protect against the channel. As noted by Shannon [9], "any redundancy in the source will usually help if it is utilized at the receiving point. In particular, if the source already has a certain redundancy and no attempt is made to eliminate it in matching to the channel, this redundancy will help combat noise."

In Section 2.5, the problem of optimal information storage was characterized as a problem of having the correct marginal channel input and end-to-end distributions, rather than having the rate equal the capacity. The two distributions that have to be used are determined from the capacity-cost and rate-distortion optimization problems, which are difficult to solve. Rather than determining the optimal solution over a fixed feasible set, an alternative method of achieving an optimal information storage system is to fix a solution and change the feasible set to make the solution optimal. The feasible set of the capacity-cost optimization problem is determined by the resource criterion, whereas the feasible set of the rate-distortion optimization problem is determined by the fidelity criterion. Choosing the resource and fidelity criteria to make an information storage system optimal is the measure matching characterization [18, 17]. The measure matching paradigm yields, in closed form, the single-letter $E$ and $F$ that make a memoryless, single-letter system with given

[7] According to Massey [72], the separation-based approach "provides jobs for those who take out redundancy and jobs for those who put redundancy back in." Such a scheme not only eliminates the middlemen ($\mathcal{M}$ and $\hat{\mathcal{M}}$) like complicated joint source-channel coding does, but actually eliminates all jobs.

$(p_U, p_{Y|X}, K_{ST}, K_{RT})$ optimal, and can be generalized for systems with memory. In the usual end-to-end information storage problem formulation, however, the goal is to design the $(K_{ST}, K_{RT})$ for a given $(p_U, F, p_{Y|X}, E)$ to make the system optimal. This problem inversion is often a reasonable approach, as good fidelity and resource criteria are difficult to determine [14]. Not only does the measure matching characterization yield a closed form solution, but also like Section 2.5.1, a system designed according to this principle is uncoded and so incurs no delay and is easy to implement. Such systems harken back to systems like AM radio. In fact uncoded operation may be universally optimal for a large set of channels, as encountered in radio broadcast; see Section 2.7.[8]

**Theorem 8 (Theorem 2.6 in [18]).** *An information system with a discrete-time memoryless source $p_U(u)$ and a discrete-space memoryless channel $p_{Y|X}(y|x)$ for which there exist values of $B$ and $\Delta$ such that $R(\Delta) = C(B)$ and has a single-letter storer and retriever $(K_{ST}, K_{RT})$ with $0 < I(U;V) = I(X;Y) < C(\infty)$ is optimal for single-letter fidelity and resource criteria generated by*

$$b(x) \begin{cases} = \nu D(p_{Y|X}(\cdot|x)\|p_Y(\cdot)) + \nu_0, & p_X(x) > 0 \\ \geq \nu D(p_{Y|X}(\cdot|x)\|p_Y(\cdot)) + \nu_0, & else \end{cases} \tag{2.51}$$

*for some $\nu > 0$ and $\nu_0$, and*

$$d(u,v) = -\mu \log p_{U|V}(u|v) + d_0(u), \tag{2.52}$$

*for $\mu > 0$ and arbitrary $d_0(\cdot)$.*

*Proof.* See [18]. □

For discrete alphabets, the theorem holds in "necessary" form in addition to this "sufficient" form [18, 17]. The cost function is an affine scaling of the derivative of mutual information with respect to the input distribution, which is sometimes

[8] In non-probabilistic treatments of AM with envelope detection, the modulation index parameterizes the operational tradeoff between cost and distortion, however our optimal tradeoff between cost and distortion is of the type defined at the beginning of the chapter.

called the Bayesian surprise [79]. The distortion function is closely related to the derivative of mutual information with respect to the test channel, which one might call the conditional entropy density, following [28]. Seeing that the optimizing cost and distortion are the derivatives of the functionals being optimized, one can see that Theorem 8 is simply a consequence of solving the Karush-Kuhn-Tucker (KKT) conditions from optimization theory.

One can interpret these optimizing resource and fidelity criteria in terms of the mechanics interpretation of the KKT conditions [80, Section 5.5.4]. In the finite alphabet setting, for each letter there are blocks with associated mass. These blocks are connected by springs with associated stiffness. The goal is to minimize the tension. In the factoring interpretation, stiffness is fixed and the mass of the blocks are optimized. Here, the stiffness function is changed so that the fixed masses of blocks minimize the tension. Thus, if the cost is proportional to the information storage utility (Bayesian surprise), things balance. Similarly, if the distortion is proportional to the uncertainty induced by the channel (conditional entropy density), things balance. The spring-mass balance view will allow us to make statements about robustness. Before moving to an analysis of robustness of uncoded storage, we present an example of measure matching. Although the example is not for a discrete alphabet, we will show uniqueness of the solution.

### ■ 2.6.1 An Example of Measure Matching

Timing information [81] is used in practice for overt [82] and covert [83] information transmission. Timing channels are also of interest in neuroscience where the timing of post-synaptic potentials are thought to perhaps provide information [5]. A degenerate form of a covert timing channel can be developed by the techniques of Chapter 4, where order is similar to timing. Spacing, the information storage equivalent to timing, may also be used for covert information storage. Techniques of embedding information in the spacing between rows in printed text and other such schemes are discussed in [84, 85].

Consider a rough model of information transmission through $M/M/\infty$ timing: a

system with a memoryless exponential source and an additive white exponential noise (AWEN) channel. The AWEN channel was previously considered in [86, Section 5] with magnitude cost; here we have not yet determined the cost. Note that this system is closely related to, but different from the $M/M/1$ single-server queue system considered in [81, Section II], [18, Section 3.6.1] and the telephone signaling system considered in [81, Section III]. Also, unlike traditional telephony, where unanswered telephone calls are toll-free [81], here there will be a resource criterion, e.g. as in modern cellular telephony pricing where unanswered calls are also tolled.

Consider a fixed channel input distribution that is given by the exponential distribution

$$p_X(x) = \lambda_0 e^{-\lambda_0 x} \mathtt{1}(x), \tag{2.53}$$

where $\mathtt{1}(\cdot)$ is the unit step function. Also consider a channel transition probability assignment that corresponds to independent additive exponential noise

$$p_{Y|X}(y|x) = \lambda e^{-\lambda(y-x)} \mathtt{1}(y-x). \tag{2.54}$$

This corresponds to an independent noise variable $Z$ with distribution

$$p_Z(z) = \lambda e^{-\lambda z} \mathtt{1}(z). \tag{2.55}$$

Consider the case when the signal to noise ratio is unity, i.e. $\lambda = \lambda_0$. Then the channel output distribution is given by the convolution theorem in a simple form according to the gamma distribution

$$p_Y(y) = \lambda^2 y e^{-\lambda y} \mathtt{1}(y). \tag{2.56}$$

For an arbitrary additive channel, the relative entropy term in (2.51) can be

simplified as follows.

$$\begin{aligned} D(p_{Y|X}(y|x)\|p_Y(y)) &= \int p_{Y|X}(y'|x)\log\frac{p_{Y|X}(y'|x)}{p_Y(y')}dy' \\ &= \int p_{Y|X}(y'|x)\log p_{Y|X}(y'|x)dy' - \int p_{Y|X}(y'|x)\log p_Y(y')dy' \\ &= \int p_Z(z)\log p_Z(z)dz - \int p_Z(z)\log p_Y(x+z)dz \\ &= -\int p_Z(z)\log p_Y(x+z)dz + h(Z). \end{aligned} \tag{2.57}$$

For the exponential-exponential example, setting $\lambda = 1$ for simplicity, the cost function that results in optimal performance is

$$\begin{aligned} b(x) &= -\nu\left\{\int_0^\infty e^{-z}\log\left[(x+z)e^{-(x+z)}\right]dz + h(Z)\right\} + \nu_0 \\ &= \nu\left\{2 - e^x E_1(x)) - \log(xe^{-x})\right\} + \nu_0, \end{aligned} \tag{2.58}$$

where

$$E_1(x) = \int_x^\infty \frac{e^{-t}}{t}dt \tag{2.59}$$

is the exponential integral function. The cost function is shown in Figure 2-4. Evidently, it is very similar to the magnitude function, especially if the constants are chosen properly for correct shifting and scaling. In fact,

$$\lim_{x\to\infty}\frac{b(x)}{x} = 1 \tag{2.60}$$

for $\nu = 1$. This result is quite similar to the result that was demonstrated in [18] for the Laplacian source and additive white Laplacian noise channel. There the cost function also resembled the magnitude function.

By Theorem 8, this cost function is a sufficient condition for optimality. We would also like to show necessity, by establishing that the capacity-cost problem here is a convex optimization problem. First we show that the subset of input distributions over the nonnegative reals that also meet the average $b$ constraint is convex. Let

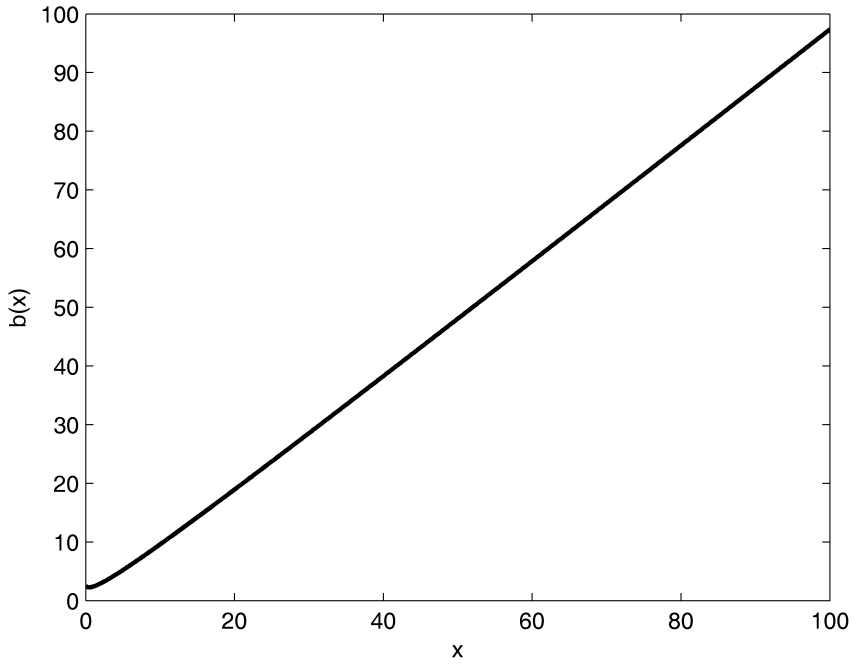


**Figure 2-4.** Optimal cost function for exponential input over additive white exponential channel; arbitrary constants $\nu$ and $\nu_0$ in (2.58) taken to be 1 and 0 respectively. The limiting value for $x = 0$ is $2 + \gamma$, where $\gamma \approx 0.577215665$ is the Euler-Mascheroni constant.

$\mathcal{F}$ denote the set of all distribution functions, and let $\Omega \subset \mathcal{F}$ be the distribution functions of nonnegative random variables with an average $b$ constraint. The set $\Omega$ is the set of $F_X$ that satisfy

$$F_X(0^-) = 0 \tag{2.61}$$

and

$$\int_0^\infty b(x) dF_X(x) \leq B. \tag{2.62}$$

The set $\Omega$ is convex because convex combinations of distribution functions are distribution functions; for two distribution functions with $F_X^{(1)}(0^-) = 0$ and $F_X^{(2)}(0^-) = 0$, the convex combination $F_X$ will also have $F_X(0^-) = 0$; and the $b$-moment of $F_X$ will be the convex combination of the $b$-moments of $F_X^{(1)}$ and $F_X^{(2)}$, thus meeting the constraint.

The second part of establishing that the optimization problem is convex is to show that the objective function is strictly concave over the feasible set. We now prove that

the mutual information $I(X;Y) = I(F_X)$ is a strictly concave function over not only $\Omega$, but all of $\mathcal{F}$ for the AWEN channel. For additive channels, the mutual information can be decomposed as

$$I(F_X) = h(Y; F_X) - h(Y|X; F_X) = h(Y; F_X) - h(Z; F_X), \tag{2.63}$$

where $h(Z; F_X)$ is a constant in $F_X$, so we want to show that $h(Y; F_X)$ is strictly concave. Following the method of [87], we can show that the channel, which gives an output $p_Y(y)$ for an input $F_X$, is one-to-one. The function $h(Y)$ is clearly a strictly concave function of $p_Y(y)$. Since the channel is a convolution with noise, it is a linear, one-to-one operator, so $p_Y(y)$ is a linear, one-to-one function of $F_X$. Thus $h(Y)$ is a strictly concave function of $F_X$, and so $I(F_X)$ is a strictly concave function of $F_X$.

Since the feasible set is convex and the objective is strictly concave, the capacity-cost problem is a convex optimization problem where the global optimum is unique.

**Theorem 9.** *The additive white exponential noise channel with noise variance 1 operates at capacity-cost with channel input distribution that is exponential with noise variance 1 if and only if the single-letter resource criterion $E$ is generated by cost function (2.58) for some $\nu > 0$ and $\nu_0$.*

*Proof.* Follows from Theorem 8 and the fact that the capacity-cost problem is a convex optimization problem. □

Now our attention turns to the fidelity criterion. Fix the source to be exponential with unit variance. Fix the forward test channel to be the AWEN channel with unit variance noise. This implies that the conditional distribution $p_{U|V}(u|v)$ is uniform over the interval $[0, v]$, taking value $1/v$. Due to the uniform distribution, the channel output is rather uninformative about the channel input. The conditional entropy density form of the optimal distortion function turns out to be $d(u, v) = \log v$ for $0 \leq u \leq v$. The distortion function does not even depend on $u$, which is rather strange indeed, and points out the danger of the measure matching approach. The fidelity and resource criteria that are generated may be quite unnatural.

Before closing this example, we give an intuitive explanation for why the capacity-achieving input distribution for the AWEN channel under our cost function $b(x)$ is different from the capacity-achieving input distribution under Verdú's magnitude cost function [86, Section 5]. Under the magnitude cost function, the capacity-achieving input distribution is

$$p(x) = \tfrac{1}{2}\delta(x) + \tfrac{1}{4}e^{-\frac{1}{2}x}\mathbb{1}(x), \tag{2.64}$$

whereas it is (2.53) under $b(x)$ from (2.58). As can be seen from Figure 2-4, the cost function starts at the Euler-Mascheroni value and then decreases for input letters less than $\sim 0.434818$, before starting to increase like the magnitude function. The dip in $b(x)$ makes low-valued input letters costly, and thus it is advantageous to use them less frequently than other letters. This acts to eliminate the delta function that appears in (2.64); under the magnitude cost function the zero-valued letter is free.

## ■ 2.7 On Robustness of Optimal Information Storage Systems

In the previous three sections, we have given three different characterizations of optimal information storage systems: rate matching, factoring, and measure matching. System design based on rate matching involves separate source and channel codes that are usually long and complicated. Design based on the factorization principle results in joint source channel codes, which may or may not be long or complicated. Systems designed according to measure matching are simple uncoded single-letter systems. Blocklength is often used as a proxy for complexity, which leads to the lossy joint source channel coding error exponent [88,89]. Since practical systems cannot use asymptotically long codes, the error exponent gives a characterization of robustness to complexity limitation. We will not discuss robustness to complexity constraints in greater detail. Rather, we will focus on robustness to system modeling errors.

To design an optimal information storage system using the rate matching, separation-based approach, the rate-distortion is taken at the capacity-cost. As was seen in the achievability proof, the separate system hinged on channel codes providing essentially a noise-free "bit pipe." But what happens if the source code rate increases beyond

the channel capacity or the channel code rate dips below the rate-distortion? That is to say, what happens when things are mismatched, perhaps due to slight errors in modeling the source distribution and the channel transition probability assignment? One approach is to study the sensitivity of the codes to mismatch or to mismatched retrievers, however a coarse look shows that the channel code error frequency must go to 1 exponentially fast when the rate exceeds the capacity [90], so the end-to-end distortion must exceed the constraints that are set. The main difficulty is that a new code cannot be designed to compensate for the modeling error.

Information storage systems are much more robust when designed according to the uncoded transmission principle inherent in the measure matching characterization. A prime example is provided by AM radio. To quantify the robustness of uncoded transmission to small modeling errors, we focus on the effects of mismatch on the Gaussian source-AWGN channel system that was explicated in Section 2.5.1. Inspired by ultrawideband communication for the information transmission context, and short-term plasticity interacting with long-term memory in the neural information storage context (see Chapter 3), we will investigate the effect of weak additive interference on the uncoded Gaussian system. These weak interferences, at least in the UWB case, remain unmodeled and so systems should be robust to them.

### ■ 2.7.1 Quadratic Gaussian Mismatch due to Interference

Before proceeding with this analysis, we review some results of Pinsker, Prelov, and Verdú on the sensitivity of channel capacity to weak additive interference [91].

#### Sensitivity of Channel Capacity

Consider the channel shown in Figure 2-5, with average power constraint $B$ on the input $X$; nominal Gaussian noise $N \sim \mathcal{N}(0, \sigma^2)$; and contaminating noise normalized so that $E[Z] = 0$ and $E[Z^2] = 1$. The scalar $\theta$ determines the contaminating noise power, $\theta^2 = \zeta$. We denote the capacity of this channel as $C_B(\theta)$, and for any $\theta > 0$, this capacity is usually not known in closed form; for $\theta = 0$, the channel reduces to the usual AWGN channel which has known solution. Since we cannot find $C_B(\theta)$, we

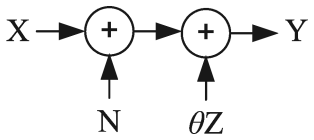


**Figure 2-5.** Quadratic AWGN channel with additive interference

would like to find an approximation to it for small $\theta$, based on the known value for $C_B(0)$. One can think of this as the technique from perturbation theory which uses the result of a known solvable problem to find an approximation to the solution of a difficult problem. Here, we aim to find the linearization of $C_B(\theta)$ around $\theta = 0$. We define the sensitivity of channel capacity with respect to contaminating noise power as

$$S_B = \lim_{\theta \to 0} -\frac{C_B(\theta) - C_B(0)}{\theta^2}. \tag{2.65}$$

By the limit definition of the derivative, this is also

$$S_B = -\frac{\partial C_B}{\partial \zeta}\bigg|_{\zeta=0}. \tag{2.66}$$

Then the linearization of the channel capacity is

$$C_B(\theta) \approx C_B(0) - S_B\theta^2. \tag{2.67}$$

When the contaminating noise is Gaussian, we can find the sensitivity directly from the definition:

$$S_B = \frac{B}{2\sigma^2 (B + \sigma^2) \log_e 2}. \tag{2.68}$$

This is clearly an upper bound on arbitrary white noise, since Gaussian noise results in worst case capacity degradation. The surprising result in [91], however is that the sensitivity is actually equal to the upper bound for any contamination noise $Z$ that is drawn i.i.d.[9] Thus we can approximate the channel capacity for a wide class of

[9]The theorems in [91] are much more general, however the result as stated is sufficiently general for our purposes.

interference noise distributions as

$$C_B(\theta) \approx \frac{\log_2\left(1+\frac{B}{\sigma^2}\right)}{2} - \frac{B\theta^2}{2\sigma^2\left(B+\sigma^2\right)\log_e 2}. \tag{2.69}$$

#### Sensitivity of Distortion to Interference

The previous subsection established the sensitivity of channel capacity to interference. Now we want to find the sensitivity of end-to-end distortion to interference. For a $\mathcal{N}(0,B)$ source, the distortion-rate function, the inverse of the rate-distortion function, is

$$\Delta(R) = B2^{-2R}. \tag{2.70}$$

Using a separation-style method, we want to use the distortion-rate function operating at capacity to find the sensitivity of distortion to the contamination noise power. Using the chain rule of differentiation, and assuming that (2.69) is exact, we find that

$$\begin{aligned} -\left.\frac{\partial\Delta}{\partial\zeta}\right|_{\zeta=0} &= -\left.\left(\frac{\partial\Delta}{\partial C_B}\frac{\partial C_B}{\partial\zeta}\right)\right|_{\zeta=0} \\ &= -\left.\left(\frac{\partial\Delta}{\partial C_B}\right)\right|_{C_B=C_B(0)}\left.\left(\frac{\partial C_B}{\partial\zeta}\right)\right|_{\zeta=0} \\ &= \left.\left(\frac{\partial}{\partial C_B}B2^{-2C_B}\right)\right|_{C_B=C_B(0)} S_B \\ &= \left(-\frac{2B\log_e 2}{1+\frac{B}{\sigma^2}}\right)\left(\frac{B}{2\sigma^2\left(B+\sigma^2\right)\log_e 2}\right) \\ &= -\frac{B^2}{\left(B+\sigma^2\right)^2} \end{aligned} \tag{2.71}$$

The Gaussian rate-distortion function is differentiable, so this computation makes sense. We have found the sensitivity of distortion to changes in contamination noise power. By the limit definition of the derivative, this is also

$$S_{\Delta,B}^{\text{optimal}} \triangleq \lim_{\theta\to 0}\frac{\Delta_B(0)-\Delta_B(\theta)}{\theta^2} = -\left.\frac{\partial\Delta}{\partial\zeta}\right|_{\zeta=0} = -\frac{B^2}{\left(B+\sigma^2\right)^2}, \tag{2.72}$$

where we have defined the new quantity, $S_{\Delta,B}$, the distortion sensitivity to contamination noise power for a fixed cost $B$. Thus the end-to-end distortion, for fixed cost $B$, as a function of contamination noise power may be approximated by

$$\begin{aligned} \Delta_B(\theta) &\approx \Delta_B(0) - \theta^2 S_{\Delta,B} \\ &\approx \frac{B\sigma^2}{B+\sigma^2} + \frac{B^2\theta^2}{(B+\sigma^2)^2}. \end{aligned} \tag{2.73}$$

This approximation involves two stages of linearization, first the capacity is linearized using the results of [91], and then the distortion is linearized using the chain rule of differentiation. Alternatively, we can get a better approximation of the distortion function if we use only one stage of linearization. This would involve substituting (2.69) into (2.70) to get

$$\Delta(C_B(\theta)) = \Delta_B(\theta) \approx \frac{B\sigma^2}{B+\sigma^2} 2^{S_B\theta^2}. \tag{2.74}$$

#### Performance of Single Letter Codes in the Presence of Interference

Now that we know an approximation of the best that we can do, let us determine how well our original single letter scheme does. The new system is created by substituting the channel in Figure 2-5 into the system in Figure 2-3. Since everything is mean zero, the second moments and the variances are equal. Working through the calculation, making use of various independence relationships, we get that

$$\Delta_B(\theta) = \frac{B\sigma^4 + B^2\theta^2 + B^2\sigma^2}{(B+\sigma^2)^2}. \tag{2.75}$$

Taking the derivative of this with respect to $\theta^2$ and evaluating the negative at $\theta = 0$ yields

$$S_{\Delta,B}^{\text{uncoded}} = -\frac{B^2}{(B+\sigma^2)^2}. \tag{2.76}$$

Now consider a decoder that has knowledge of the interference statistics and takes that into account. The decoder is now a scalar multiplication by $B/(B+\sigma^2+\theta^2)$.

Computing the distortion, we get

$$\Delta_B(\theta) = \frac{B\left(\sigma^2 + \theta^2\right)}{B + \sigma^2 + \theta^2}. \tag{2.77}$$

The associated sensitivity is

$$S_{\Delta,B}^{\text{uncoded,PI}} = -\frac{B^2}{\left(B + \sigma^2\right)^2}. \tag{2.78}$$

Evidently, partial knowledge of the interference noise statistics does not reduce the sensitivity to the interference noise power.

**Theorem 10.** *To a first-order approximation, the end-to-end distortion performance in the single-letter mismatched case, in the single-letter partially-matched case, and in the optimal case are the same.*

*Proof.* Compare (2.76), (2.78), and (2.71). □

#### Writing on Dirty Paper

Since we are looking at channels with interference, for completeness, we should also comment on dirty paper [92] (or more precisely generalized writing on dirty paper [93]) results. In the previous subsection where we considered a partially informed single-letter decoder, the decoder had access to the variance of the contamination noise (dirt) distribution. In the dirty paper scenario, the encoder has access to the actual realization of the dirt, not just a statistical parameter. It has been shown that for any i.i.d. interference noise, the capacity is not reduced, so the distortion is not increased, and so the sensitivity of distortion to interference, $S_{\Delta,B}^{\text{dirtypaper}}$, is zero.

#### Comments on Robustness

First, we have shown that distortion as a function of interference noise power is continuous for arbitrary zero-mean, memoryless interference. This follows from the continuity of the channel capacity as a function of interference noise power and the continuity of the distortion-rate function for a Gaussian source with respect to squared

error. One can regard the optimal cost-distortion point as a saddlepoint, with perturbations causing smooth changes.

Second, we have found an approximation of the optimal sensitivity of distortion to interference noise power through a first-order approximation of the channel capacity as a function of interference noise power.

Third, we have found the distortion as a function of interference noise power for two single-letter coding schemes: one uses the decoder designed for no interference, and the other uses the decoder that is designed with interference noise power taken into account. These distortion-interference noise power functions give associated sensitivity values. We find that the sensitivity in the optimal case, the partially informed decoder case, and the uninformed decoder case are all equal. Thus to a first-order approximation, the distortion-interference noise power functions are all equal for weak additive interference. That is to say, in an approximate sense, single-letter codes perform as well as the optimality bound, despite not being designed with the interference in mind.

The robustness of the single-letter code is in stark contrast to a separation-based approach for which additive interference causes the probability of error to go to one. To continue operating at optimal levels, new source and channel codes would be required. The single-letter coding scheme, on the other hand, remains fixed, yet has performance degradation commensurate with the optimal performance degradation.

One can characterize this robustness to uncertain weak channel interference as a type of weak universality of uncoded transmission, to complement other true universality properties. A true universality property for the partially informed scenario considered here would hold if the interference was restricted to be white Gaussian. Then the resultant channel would be AWGN, and the uncoded system is optimal for the Gaussian source with squared error fidelity and any member of the entire class of AWGN channels with power constraint. There would be some degradation in the fully uncoded system, since the decoder would not apply the correct scaling.

### 2.7.2 Finite Alphabets

Since the space of probability distributions is not compact and information is not a continuous functional in general, a priori, it is not clear that channel input distributions close to the capacity-achieving distribution will perform almost as well. Similarly, it is not clear that channels close to the rate-distortion achieving forward test channel will perform almost as well. When the alphabet is finite, however, nice analytical properties of the space of probability distributions emerge, and we can make statements about robustness to parameter mismatch.

The cost constraints that we impose on the channel input distributions are of the form $\mathrm{E}[b(X)] \leq B$, which include the boundary, and so the spaces of input distributions are closed sets. Moreover, they are bounded subsets of $\mathbb{R}^{|\mathcal{X}|}$ and are thus compact. The operation of the channel on the channel input distribution to produce the channel output distribution is a linear function and hence a continuous function, so a neighborhood in the channel input distribution space is mapped to a neighborhood in the channel output distribution space. Therefore, the set of possible channel output probability distributions, $\Pi$, is closed. As shown, e.g. in [94], if $\Pi$ is a closed subset of $\mathbb{R}^{|\mathcal{Y}|}$, then the set of all probability measures on $\Pi$ is compact in the topology of weak convergence. Since mutual information is continuous in that topology, there exists a capacity-achieving input distribution. Uniqueness of the capacity-achieving input distribution can also be shown. Due to these analytical properties, Theorem 8 provides both necessary and sufficient conditions for optimality.

These analytical properties also simplify statements about robustness of uncoded information storage systems in the finite alphabet case. Consider a storage system where the channel input distribution (source distribution) is slightly different than what has been designed for. Due to continuity of the channel mapping, the channel output distribution (reconstruction distribution) $p_Y(y)$, is also only slightly different than what is expected. By the continuity of the relative entropy function in (2.51), this implies that the cost function for which the system is now optimal is only slightly different than the cost function for which the system was designed. Similarly, by the fact that both $p_U(u) = p_X(x)$ and $p_V(v) = p_Y(y)$ are only slightly perturbed,

by Bayes' rule, and by continuity of the logarithm function in (2.52), the distortion function for which the system is now optimal is only slightly different than the distortion function for which the system was designed.

Supposing the cost and distortion functions are fixed, the perturbation will result in suboptimality, however the suboptimality will be small for small perturbations. In addition to our previous characterizations of capacity-cost, there is also an output-centered characterization, based on the KKT conditions.

**Theorem 11 (Corollary 2.3.4 in [95]).** *For $B > B_{min}$*

$$C(B) = \min_{p_Y(y)} \min_{\gamma \geq 0} \max_{x \in \mathcal{X}} \left[ D(p_{Y|X}(y|x) \| p_Y(y)) + \gamma(B - b(x)) \right]. \tag{2.79}$$

*Proof.* See [95]. □

Like the mechanics interpretation of the KKT conditions (Section 2.6), this output-centered minimax characterization of capacity-cost has a nice geometric interpretation. Since relative entropy is a measurement of distance between distributions, $C(B)$ is the so-called I-radius [94] of the smallest I-sphere that contains the set of distributions $p_{Y|X}(\cdot|x), x \in \mathcal{X}$. The capacity-cost achieving *output* distribution, $p_Y(y)$, is the I-centroid of the I-sphere. Perturbing $p_Y(y)$ slightly from the I-centroid position increases the maximal I-radius only slightly due to compactness, and so the system is only slightly suboptimal. A similar result can be shown for the rate-distortion using Corollary 2.3.7 in [95]. Equivalent statements for the cost function, the distortion function, and suboptimality for capacity-cost and rate-distortion can be made when the perturbation is to the channel transition probability assignment rather than to the channel input distribution. Overall with finite alphabets, small modeling errors do not appreciably degrade performance of information storage systems designed according to uncoded principles, due to continuity and compactness.

Chapter 3

# Information Storage in Neural Memory Channels

We have promulgated several interpretations and characterizations of optimal information storage systems in Chapter 2. Here we apply these characterizations to study scientific questions that arise in neuroscience.[1] It is widely believed that memories are stored in the brains of mammals and that the physical mechanism for information storage in the brain involves changes in synapses, termed synaptic plasticity. As such, our theory provides explanations for and makes predictions about the physical properties of synapses.

Our approach is rooted in the optimization approach to biology [97, 98, 99, 100] in which biological systems are thought to be best solutions to some optimization problem and thus seem to be "designed" or "engineered" under some optimality criterion. Since evolution by random mutation and natural selection favors genotypes of high fitness, qualities that affect fitness tend to improve through evolution. Thus, mathematical analysis of fitness optimization seems reasonable as an approach to understand why animals are the way that they are. Since establishing a quantitative relationship between fitness and observable features is difficult, the mathematical problem posed is to optimize beneficial features under cost constraints.

The scientific method for the optimization approach that we will follow uses a few basic steps [98].

[1] It is assumed that the reader has basic knowledge of neuroscience, though important details of the nervous system will be reviewed in the text. A leisurely introduction to the subject is provided by [96].

1. Ask an explicit scientific question, informed by existing experimental findings.

2. Define a feasible set of strategies or structures in the model, relating to the question.

3. Define the objective fitness function to be optimized in the model. This performance criterion should be mathematical in nature.

4. Given the feasible set and the objective function, determine the optimal strategy or structure by appropriate mathematical analysis. This yields optimization principles as hypotheses, which in turn give explanations and predictions.

5. Test the hypotheses with experimental measurements. Either confirm the theory or falsify it.

This epistemic framework is different than the framework usually adopted in engineering theory or in mathematics, since in science, one is always concerned about the physical world that is to be described by the theory. The connection of science to the physical world necessitates the comparison to experiment, either confirming or falsifying the theory. Comparison to experiment is in some sense a more stringent requirement because not only must the solution be mathematically optimal for the model (Step 4 above), the solution must also exist in the physical world.

Before we go through these steps for our problem on information storage in the brain, we should state some caveats about the optimization approach. First and foremost, the goal of such investigations is not to prove that a particular animal is optimal; rather it is to generate general optimization principles that can unify and explain experimental observations. It must also be kept in mind that animals may have been "designed" for niches different from the ones that they currently occupy, features may have been inherited unchanged from ancestors, and there is always the possibility of pure randomness in evolutionary progression. Notwithstanding these caveats, it is hoped that the idea of optimality leads to a harmonious theory of the physical architecture of the brain.

## 3.1 Scientific Questions Posed

As noted in Section 2.4, early work on establishing the electrical nature of the nervous system was carried out in the eighteenth century. In the interceding centuries, many facts about the physical nature of the nervous system have been collected. With the late nineteenth century discovery of synapses and the recognition of their importance in the two principal tasks of the brain, information processing and information storage [101], their physical properties have been the subject of extensive experimentation. In this section, we review the results of some experiments and use them to formulate our explicit scientific questions.

#### Average Noisiness of Synapses

Numerous studies have found that typical synapses in the central nervous system, including the brain, are noisy devices in the following sense. Arriving pre-synaptic action potentials occasionally fail to evoke an excitatory post-synaptic potential (EPSP) due to probabilistic release of neurotransmitters at the synaptic terminal; furthermore, when an EPSP is evoked, the amplitude varies from trial to trial [102, 103, 104, 105, 106, 107, 108]. This has led other investigators to ask whether unreliable synapses are a feature or a bug [109, Section 13.5.5]. Taking an optimistic viewpoint, we ask:

**Question 1.** *Why are typical central synapses noisy?*

The noisiness of typical central synapses has seemed puzzling because synapses act as conduits of information between neurons [109], and in general, synaptic unreliability is detrimental for information transmission. Several previous theoretical studies have considered the impact of synaptic noise on information transmission through a synapse, generally in the context of sensory processing [110, 111, 112, 113]. It has been shown that, under some suitable constraint models, synaptic noisiness facilitates the efficiency of information transmission. Moreover, Laughlin et al. [114] have noted that splitting information and transmitting it over several less reliable but metabolically cheaper channels reduces energy requirements. However, adding information channels invokes costs associated with building and maintaining those

channels [115, 116], which must also be taken into account [117]. Our approach to addressing this question will focus on information storage rather than on information transmission.

#### Distribution of Synaptic Efficacies

A second observation that has been made regarding synapses is the broad distribution of synaptic efficacy (synaptic weight) across different synapses in the brain. Although the majority of synaptic weights are relatively weak (mean EPSP $< 1$ mV), there is a notable tail of stronger connections [118, 104, 119, 105, 108, 120], with EPSP values exceeding 20 mV having been observed.

**Question 2.** *Why does the distribution of synaptic efficacies display a notable tail of strong connections?*

The optimal design of digital filters under a minimax criterion yields solutions with equal ripples [121], and maximizing volume of a rectangular prism for fixed surface area yields a cube with equal sides. Optimization problems in information theory also have solutions with symmetry properties, but in terms of equal Bayesian surprise, rather than equiprobable channel outputs (Section 2.7.2). Given the pervasiveness of equality and symmetry in the solutions to optimization problems, one might think at first that if there is some optimal value of synapse strength, then this value should be used for all synapses, hence the puzzling nature of this observation. Brunel et al. [122] have explained the distribution of synaptic weights in the cerebellum by maximizing the storage capacity of a perceptron network. When we attack this question from the information storage perspective, we will use an information theoretic approach. The implementation of storage through either neural networks or coding will be abstracted. Memory storage will be considered from a physical perspective, looking at the information storage density of neural tissue, rather than previous work on the memory storage capacity of specific neural network models [122,123,124,125,126,127], which is quite different.

#### Sparsity of Synaptic Connectivity

A third observation that has been made regarding synapses is their absence. One can experimentally create an approximate adjacency matrix of neuron-neuron connections. It has been found that synaptic connectivity, i.e. this adjacency matrix, is sparse in the brain. One might explain this away by inferring that the absent connections are between neurons that are physically distant, however the synaptic connectivity is sparse in local circuits also. That is to say, the chance of finding a synaptic connection between a randomly chosen pair of neurons, even nearby ones, is much less than one [118, 104, 128, 119, 105, 129, 130, 131].

**Question 3.** *Why is synaptic connectivity sparse?*

In local circuits, neuronal axons and dendrites are close enough so that the potential connectivity is nearly all-to-all, so the cost of adding connections is not overwhelming. It seems that increased connectivity would enhance system functionality, so the sparse connectivity is puzzling. Sparseness has previously been explained by maximizing the information storage capacity of a neural network [122].

#### Discrete-valued Synaptic Efficacies

There have been some controversial reports that synaptic efficacies vary in discrete steps, rather than in a graded manner [132,133,134], which leads to our final question.

**Question 4.** *Why do synaptic efficacies vary in discrete steps?*

A caveat with this question is that due to the difficulty in differentiating continuous and discrete objects in experimental science, it is unclear whether any hypothesis can be verified experimentally. In our discussion of this question from the information storage perspective, we will be able to make some predictions, but not nearly as well as for the first three questions.

Having formulated the questions to be addressed, we move to the second and third steps of our scientific program: the development of a physicomathematical model of synapses as information storage elements.

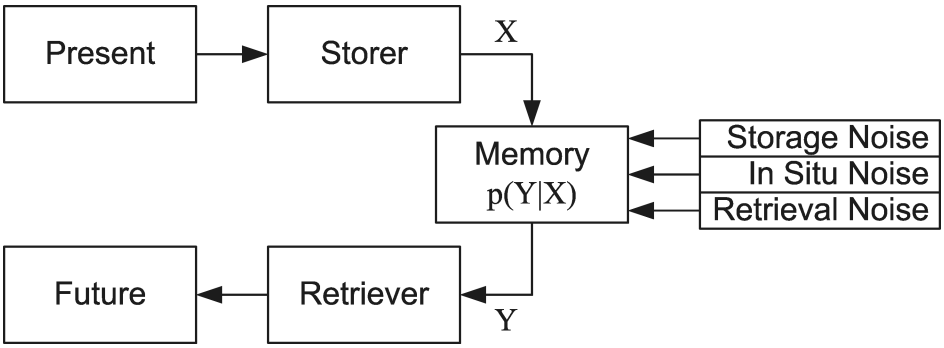


**Figure 3-1.** Schematic diagram of a neural memory system cast as a general information storage system. The various sources of noise in the system have been explicitly notated. Storage noise refers to noise that arises in the storage process; in situ noise refers to noise that perturbs the information while it is stored; and retrieval noise refers to noise in the retrieval process.

## ■ 3.2 Model of Synaptic Memory

We develop a model of synapses, based on their role in long-term information storage, which will lead to a unified theoretical framework answering Questions 1-4. It is widely believed that long-term memories are recorded in neuronal circuits through alteration in the efficacy of existing synapses [135, 136], through long-term potentiation (LTP) and long-term depression (LTD) [137, 138]. Synaptogenesis and synapse elimination are alternative ways that neuronal circuits may be altered. Memories are retrieved from storage by chemical and electrical activity of neurons, which generates synaptic potentials determined by the pattern of synaptic connectivity between them. Thus a synaptic memory system is naturally cast into the model of a general information storage system (Figure 2-1) with suitable identification as shown in Figure 3-1. We should note that although information storage is well recognized as a case of a general communication system [95, 74, 68] and information theory has been successfully applied in neuroscience [5], the application of information theory to the analysis of synapses as memory elements has received little attention previously.

To more fully specify the model in terms of the definitions of Chapter 2, the source is determined by the things that an animal is to remember; there is also some associated fidelity criterion, perhaps based on the utility of degraded memories. Since the source and the fidelity criterion seem particularly hard to describe, we will not consider them further and will focus strictly on the channel. We will return

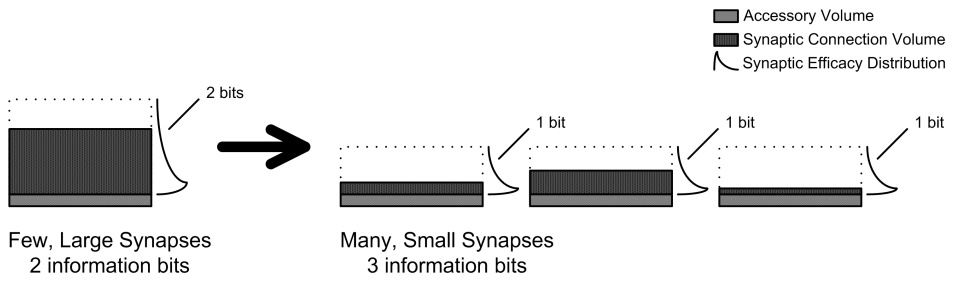


**Figure 3-2.** Illustration of two possible volume-constrained signaling strategies for synaptic information storage. Replacement of a large synapse with several smaller ones (connecting different neurons) may increase information storage capacity (and does, as shown in Section 3.3). Notice that the total volume of the smaller synapses is the same as that of the larger synapse. Since synapse volume is a monotonic function of synapse efficacy, the realization of synapse volume also represents the efficacy realized by that synapse.

to sources and fidelity criteria in Chapter 4. The storer is most likely composed of some sort of coding, followed by a process of synaptic weight change. Synaptogenesis creates a synapse where there was none previously; long-term potentiation increases the weight of an existing synapse; long-term depression decreases the weight of an existing synapse; and synapse elimination removes an existing synapse. Each synapse represents a channel usage. The channel input alphabet is the synaptic efficacy and the distribution of synaptic efficacies across synapses is the channel input distribution. Figure 3-2 illustrates the basic form that the signaling scheme can take. For example, a large synapse can be replaced by a number of smaller synapses that take the same volume.

There is noise associated with the LTP, LTD, synaptogenesis, or synapse elimination processes, which we term the storage noise. While information is stored in the synapse, computational and short-term memory processes may act to perturb the synaptic efficacy; this is termed in situ noise. The retriever consists of activity to retrieve synaptic efficacies, as well as some sort of decoding. There is noise associated with this EPSP-generating chemical and electrical activity, which we term retrieval noise.

The channel alphabet $\mathcal{X}$ is the alphabet of synaptic weights, which we consider to be the real line. Excitatory synapses are represented by positive values, and inhibitory synapses would be represented by negative values, though we do not consider such

synapses in this investigation. A true characterization of the synaptic weight alphabet would require identification and description of the so-called engram, the physical embodiment of memory. Since this is unavailable, we will identify the input letters with the mean EPSP amplitude generated over numerous firings of the pre-synaptic neuron and will measure it in units of electrical potential. Numerous experimental studies have shown synaptic weight to be positively correlated with synaptic volume [139, 140, 141, 142, 143, 144]. Synaptic volume increases with LTP and decreases with LTD [145, 146]. Synapse volume also increases with synaptogenesis and decreases with synapse elimination. Other anatomical and physiological observations show that the volume of individual synaptic contacts is correlated with many microscopic characteristics, such as the number and area of active zones, number of vesicles, area of the post-synaptic density, and the number of receptors [141,142,147,143,148,144,149]. As stated in Ramón y Cajal's laws of economy of space, time, and matter [101], and numerous subsequent theoretical investigations, volume in the brain is a costly resource [150, 151, 152, 153]. Volume not only measures the amount of space that is taken, but also measures the amount of material and metabolic energy required to maintain a structure. The fact that volume sums linearly and can be evenly exchanged among various structures makes it a convenient measure. Given these properties of synapse volume, we take our resource criterion to be generated by the single-letter synapse volume cost function.

Although synaptically connected pairs of cortical neurons usually share multiple synaptic contacts [154, 155, 119, 156], here we refer to these contacts collectively as a synapse. Such a definition is motivated by electrophysiological measurements, which record synaptic weight of all the contacts together. The total synaptic volume is correlated with the total synaptic weight because contributions of individual contacts to synaptic weight may also add up linearly [157] and a neuron can be viewed as a single computational unit [158]. There is evidence that multiple synaptic contacts within a connected pair of neurons have correlated release probability [155] and that the total synaptic connection weight correlates with the number of synaptic contacts [154]. Thus this definition seems well justified. In an alternative scenario where the

integrative compartment is smaller, such as a single dendritic branch [159, 160], and individual synaptic contacts can vary their weights independently, our definition of channel usage would need to be redefined at this finer level.

In addition to the volume cost of the synapse itself, some accessory volume is needed to support a synapse. Accessory volume includes the volume of axons, dendrites, cell bodies, glia, and perhaps extracellular space. Since each potential synapse represents a channel usage, this accessory volume may be interpreted as the cost of providing discrete-space bandwidth. Unlike in other information storage systems, bandwidth and channel alphabet cost (power in other systems) consume the same resource and thus an interesting tradeoff between allocating resources for bandwidth or for power is presented. An alternative view is that there are no free symbols, and that the minimal achievable cost $B_{min}$ is this accessory volume. We will denote it by $V_0$.[2]

As stated before, in our model, the channel input letters are the average EPSP amplitudes, $x \in \mathcal{X}$, where $x \triangleq \mathrm{E}_{p_{Y|X}}[Y]$. A second-order characterization of the noise is given by the standard deviation of the EPSP readout trials. The standard deviation of the EPSP amplitude from trial to trial in a given synapse is denoted as the noise amplitude, $A_N$. If the noise were from an additive channel, with the mean-zero noise variable $Z$ distributed according to $p_Z(z)$, then $A_N = \sqrt{\mathrm{E}_{p_Z(z)}[Z^2]}$. Synaptic release may be modeled by a Poisson process [161, 162], which leads to the conclusion that noise amplitude increases sublinearly with the synaptic weight. In fact, the relationship between synapse amplitude and noise amplitude is well-approximated by a power law with exponent about 0.38 [119, 129], see Figure 3-3.

Since the noise amplitude, $A_N$, is related by a power law to the mean EPSP amplitude, $x$, which is itself strongly correlated with the synapse volume, $V$, we can formulate the following relationship:

$$\frac{V}{V_\nu} = \left(\frac{x}{A_N}\right)^\alpha, \tag{3.1}$$

[2]In this chapter, $V$, $V_0$, $V_\nu$, and related expressions will refer to the volume cost, rather than to the source reconstruction as in previous and subsequent chapters.

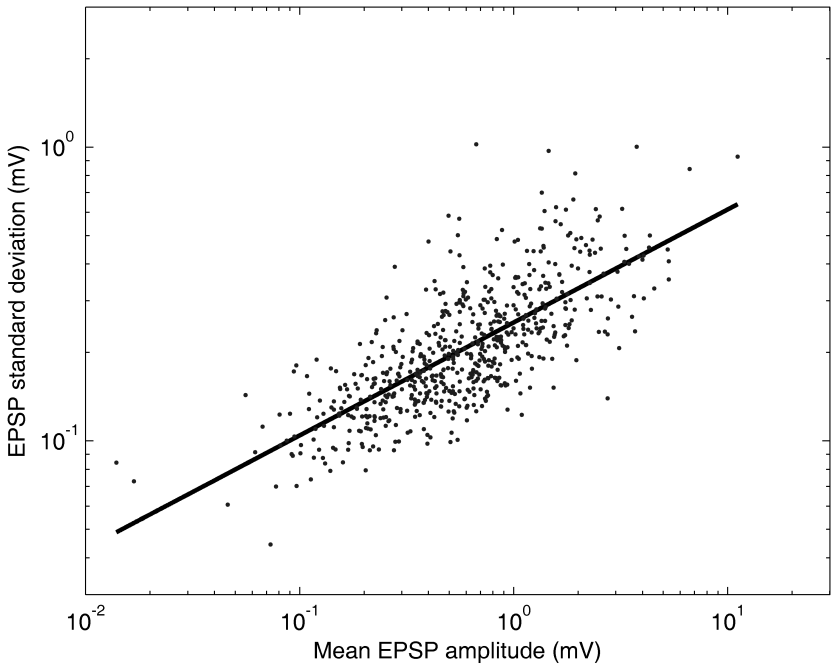


**Figure 3-3.** Standard deviation of the EPSP amplitude as a function of the mean EPSP amplitude. Data presented corresponds to 637 synapses that have been measured experimentally (see [120, 163, 129] for details on experimental procedure). Solid line is least squares fit by a power law with exponent 0.38.

where $V_\nu$ is the volume of a synapse with a signal to noise ratio of 1. This normalization constant should not be confused with other volume quantities. Although existing experimental measurements [139, 140, 141, 143, 129, 144] support (3.1), they are not sufficient to establish the value of the exponent, $\alpha$. Determination of $\alpha$ would specify part of the cost, $V(x)$, of the input letter $x$.

This establishes the physicomathematical model of synapses in terms of channel input alphabet $\mathcal{X}$, the resource criterion $E$ generated by the single-letter cost function $V(x) + V_0$, and a description of the channel noise by its moment $A_N$. The overall objective function we would like to meet is to have optimal information storage in the sense of Chapter 2. To do so, we will try to meet Condition 1 from the factorization interpretation of optimal storage and use (2.51) from the measure matching interpretation of optimal storage.

For the Condition 1 problem, the decision variable in the optimization problem is the signaling scheme which is described by the distribution of synaptic efficacy across the synapses, $p_X(x)$. For the (2.51) problem, the decision variable will be the single-letter resource criterion $E$, specified in terms of $b(x) = V(x) + V_0$. Note that constraints imposed by the evolutionary process and by some biological mechanisms have been implicitly incorporated. That is to say, the basic structure of the system has been assumed, and the details are to be determined by the optimization theory.

## ■ 3.3 Noisy Synapses Maximize Information Storage Capacity

Question 1 asked why synapses are noisy on average. By the relationship between noisiness and small volume established in (3.1), the question can be recast as the question of why synapses are small on average. That is to say, why is the average incurred cost $B$ small. Our mathematical formalization based on the capacity-cost function $C(B)$ is inadequate to address this question as it is a curve parameterized by $B$, and we are interested in determining a particular point on the curve. Rather than trying to just achieve some point on the capacity-cost curve, it seems that one should want to achieve the capacity per unit cost. At this point, the rate achieved

is the least costly on a per unit basis; the most "bang for the buck," as it were. As shown by Verdú [7], the capacity per unit cost is given by

$$\mathfrak{C} = \sup_{B>0} \frac{C(B)}{B} = \sup_{p_X(x)} \frac{I(X;Y)}{\mathrm{E}_{p_X(x)}[V(X)+V_0]}, \tag{3.2}$$

where $V(X)+V_0$ is the volume cost for the channel input letters. So in this section, we will deduce optimal average synaptic weight and volume by maximizing information storage capacity per unit volume. We invoke the synaptic weight/volume relationship formulated in the previous section (3.1) with $\alpha = 2$, other cases to be considered in the following sections. For $\alpha = 2$, the problem of maximizing information storage capacity in a given volume reduces to the well-studied problem of maximizing channel capacity for a given input power. When the channel contributes additive white Gaussian noise (AWGN), the solution is well known. For concreteness and ease of exposition, we assume here that the noise is Gaussian with a given variance; this model is also reasonable for the observed noise. The qualitative result and the general optimization principle do not depend on this assumption; at the end of this section, we will argue that the conclusions hold for other noise models.

Information storage capacity per synapse (measured in nats) for the AWGN channel is given by [9]

$$C_{synapse} = \frac{1}{2} \ln \left( 1 + \frac{\langle X^2 \rangle}{A_N^2} \right), \tag{3.3}$$

where $\langle X^2 \rangle$ is used to denote $\mathrm{E}_{p_X(x)}[X^2]$. The quantity $\frac{\langle X^2 \rangle}{A_N^2}$ is the average SNR among synapses with respect to the channel input and to the noise distributions. SNR for each synapse is defined as the square of the mean EPSP amplitude divided by the trial-to-trial variance of EPSP amplitude. Using (3.1), we can rewrite information storage capacity in terms of volume

$$C_{synapse} = \frac{1}{2} \ln \left( 1 + \frac{\langle V \rangle}{V_\nu} \right), \tag{3.4}$$

where $\langle V \rangle$ is the average synapse volume, $\mathrm{E}_{p_X(x)}[V(X)]$, excluding the accessory volume.

As volume is a scarce resource, information storage capacity is likely to be optimized on a per volume basis. For example, placing two or more smaller synapses (connecting different pairs of neurons) in the place of one larger synapse may increase memory capacity, see Figure 3-2. Maximum storage capacity is achieved when synaptic weights are independent, since the capacities of channels with memory are bounded by the capacities of channels without memory. Dividing by the average cost of a synapse, the total storage capacity of a unit volume of neural tissue is

$$\mathfrak{C}_{volume} = \frac{C_{synapse}}{\langle V \rangle + V_0} = \frac{1}{2\left(\langle V \rangle + V_0\right)} \ln\left(1 + \frac{\langle V \rangle}{V_\nu}\right). \tag{3.5}$$

Information storage capacity per unit volume, $\mathfrak{C}_{volume}$, as a function of the average size of the synapse, the relationship in (3.5), is plotted in Figure 3-4 for different values of $V_0$. The location of the maximum of such a curve gives the normalized average volume, $\langle V \rangle / V_\nu$, that provides the optimal storage capacity in the capacity per unit cost sense.

Next, we can analytically calculate the optimum average synapse volume $\langle V \rangle$ that maximizes information storage capacity per volume $\mathfrak{C}_{volume}$ for given accessory volume $V_0$ and normalization $V_\nu$. This problem is mathematically identical to maximizing information transmission along parallel pathways [117]. We take the derivative of (3.5) and set it to zero to obtain

$$2\frac{\partial \mathfrak{C}_{volume}}{\partial \langle V \rangle} = \frac{-1}{\left(\langle V \rangle + V_0\right)^2} \ln\left(1 + \frac{\langle V \rangle}{V_\nu}\right) + \frac{1}{\langle V \rangle + V_0} \frac{1}{\langle V \rangle + V_\nu} = 0. \tag{3.6}$$

This implies that the optimal $\langle V \rangle$ can be found by solving the equation

$$\frac{\langle V \rangle + V_0}{\langle V \rangle + V_\nu} = \ln\left(1 + \frac{\langle V \rangle}{V_\nu}\right). \tag{3.7}$$

In the limiting case, when $V_0 \ll V_\nu$, the optimizing average volume $\langle V \rangle$ and the maximum storage capacity achieved are [117]:

$$\langle V \rangle = \sqrt{V_0 V_\nu}, \tag{3.8}$$

and

$$\max \mathfrak{C}_{volume} = \frac{1}{2V_\nu}. \tag{3.9}$$

In the opposite limiting case when $V_0 \gg V_\nu$,

$$\langle V \rangle \ln \left( \langle V \rangle / V_\nu \right) = V_0, \tag{3.10}$$

and

$$\max \mathfrak{C}_{volume} = \frac{1}{2\langle V \rangle}. \tag{3.11}$$

If one looks at the dependence of information storage capacity $\mathfrak{C}_{volume}$ (peak height in Figure 3-4) and optimal synaptic volume (horizontal coordinate of the peak in Figure 3-4) on the accessory volume $V_0$, as would be expected, maximum information storage capacity per unit volume is achieved when the accessory volume $V_0$ is the smallest possible. In this regime, average synapse volume $\langle V \rangle$ is much less than $V_\nu$ and according to (3.1), synapses should therefore be noisy.

In reality, the accessory volume $V_0$ may not be infinitesimal as this could affect system functionality adversely. Although these requirements are abstracted in the information theory, they are present in the physical system. As an example, there is a hard limit on how thin axons can be [164]. Also, reducing wiring volume may increase conduction time delays and signal attenuation [165]; delay is objectionable for control applications [52, 40]. In fact, delay and attenuation are optimized when the wiring volume is of the same order as the volume of synapses [166]. Since the increase in capacity achieved by reducing accessory volume below $V_\nu$ is not too large, it is reasonable to think that the channel input distribution is such that the average synapse volume $\langle V \rangle$ is either less than or of the order of $V_\nu$. In either case, we arrive at the conclusion that typical synapses should be noisy, in agreement with experimental observations.

The advantage of having greater numbers of smaller synapses is valid not only for the AWGN model that was considered above, but also for many reasonable noise and cost models. The best way to see this advantage is through the non-decreasing and

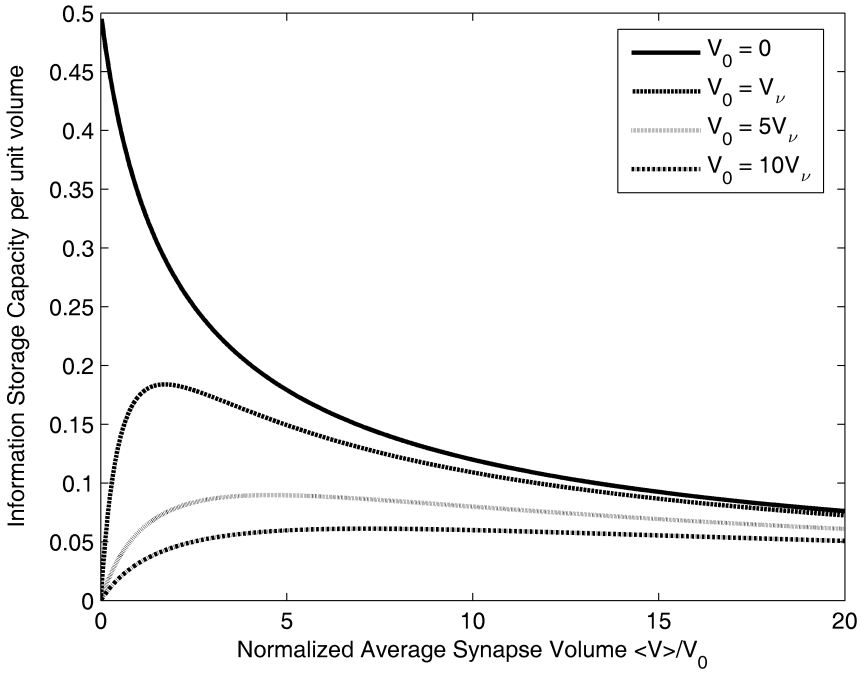


**Figure 3-4.** Information storage capacity per unit volume of neural tissue as a function of normalized average synapse volume for AWGN synapse approximation. The relationship between signal to noise ratio and volume for this plot uses accessory volume, $V_0$, values of 0, 1, 5, and 10, normalized with respect to $V_\nu$. When $V_0 = 0$, the maximum storage capacity per unit volume occurs when average synapse volume is infinitesimal. When $V_0 > 0$,, the finite maximum storage capacity per unit volume occurs at some non-zero normalized synapse volume.

concave downward nature of the capacity-cost function [22,37]. Due to concavity, the slope (capacity/cost) increases at lower costs. If there are no zero-cost symbols, then the capacity per unit cost is maximized at the average cost value where a line through the origin and tangent to the capacity-cost function has its point of tangency, the location of the peaks in Figure 3-4. If there is a zero-cost symbol, then the optimum is for zero average cost ($V_0 = 0$ curve in Figure 3-4). It is usually difficult to find the optimum capacity per unit cost analytically [7], however there is a numerical algorithm that can be used for such a computation [167]. Similar mathematical arguments have been used in the context of information transmission to show that having parallel less reliable channels, such as synapses [114] and ion channels [116], reduces metabolic costs and to show that scheduling of packets for wireless radio networks should go as slow as possible [168]. Along these lines, Berger notes that "information handling always should be conducted at as leisurely a pace as the application will tolerate" [73].

In this section, we have developed the following general optimization principle, provided that the accessory volume needed to support a synapse is small.

**Optimization Principle 1.** *To optimize information storage capacity per unit volume of neural tissue, synapses should be small and noisy on average.*

## ■ 3.4 Optimal Distribution of Synaptic Efficacies in the Discrete States Model

Having established the fact that synapses should be small and noisy on average, we are interested in how volume and EPSP amplitude should be distributed among synapses. The distribution of synaptic weight addresses both Question 2 and Question 3 because absent synapses are synapses with zero weight. In the AWGN model used in Section 3.3, the capacity-achieving input distribution is also Gaussian [9], as shown in Section 2.5.1. By (3.1) with $\alpha = 2$, a capacity-achieving Gaussian input distribution implies a distribution of synaptic volume that follows an exponential distribution. If the noise amplitude $A_N$ is constant, synaptic weight has a Gaussian distribution, as suggested previously by different methods [122]. If $A_N$ scales as a power of $X$ (Figure 3-3), the synaptic weight distribution is a stretched (or compressed) exponential. The

cumulative distribution function

$$F_X(x) = \exp\left[-\left(\frac{x}{x_0}\right)^c\right], \tag{3.12}$$

with suitable normalization $x_0$, is a stretched exponential for $c < 1$, an exponential for $c = 1$, and a compressed exponential for $c > 1$. However, it is not clear whether the prediction for the synapse weight and volume distributions from the AWGN model can be taken at face value for the following reasons. First, the AWGN channel model allows both negative and positive signals whereas synaptic weight is positive for excitatory synapses; remember we have not been considering inhibitory synapses. Second, the Gaussian noise assumption is unlikely to hold especially if synaptic weight must be non-negative. Third, synaptic volume may not scale as the synaptic weight SNR squared, i.e. the parameter $\alpha$ in (3.1) may not be 2.

We now consider an alternative noise model, where the cost function can be chosen arbitrarily and the synaptic weight is non-negative, but still yields an exactly solvable optimization problem. The reason an exact solution can be found is that the noise is treated approximately. Rather than considering a continuous distribution of synaptic weights, we consider a set of discrete synaptic states, with each state representing the range of weights in the continuous distribution that could be confused on retrieval due to noise. Then the difference in synaptic weight between adjacent states is given by twice the noise amplitude $2A_N$. We are engaging in hard sphere packing rather than soft sphere packing, hence the factor of 2. Each of these finite states is viewed as an input symbol with associated volume cost.

The conversion of the noisy continuous-valued input channel into a zero-error, discrete-valued input channel is a convenient approximation [169], especially when a probabilistic description of the noise is not known. This approximation allows us to calculate the $\epsilon$-capacity-cost, a generalization of the Kolmogorov $\epsilon$-capacity [170], rather than the Shannon capacity. Since the converted channel is noiseless, the mutual information reduces to the self-information of the channel input distribution (or, equivalently, channel output distribution). By resorting to this $\epsilon$-capacity-cost

approximation, we do not wish to imply that synaptic weights in the brain vary in discrete steps. In Section 3.7 we will validate this approach by comparing its predictions to the predictions from a continuous channel model (Section 3.5).

Since the self-information is identical to entropy in the discrete alphabet case, the maximization of information storage capacity per volume can be solved using the maximum entropy method, a standard statistical physics technique [171]. In neuroscience, such a mathematical problem has been solved in the context of neuronal communication by the spike code [172, 173]. Consider a set of synaptic states, $i$, characterized by the EPSP amplitudes, $x_i$, and volume cost, $b_i$, which includes both the synapse volume and the accessory volume $V_0$ . We search for the probability distribution over synaptic states, $p_i$, that maximizes information storage capacity:

$$C_{synapse} = -\sum_i p_i \ln p_i \tag{3.13}$$

per average volume of a synapse

$$B = \underset{p_i}{\mathrm{E}}[b(x)] = \sum_i p_i b_i. \tag{3.14}$$

To reiterate, the average synaptic volume cost $B$ includes the accessory volume, $V_0$, which was excluded from the definition of $\langle V \rangle$ used in the previous section.

It is well-known that the maximum entropy distribution takes an exponential form [171]. Therefore the probability distribution over synaptic states, $p_i$ that maximizes information capacity is given by

$$p_i = \frac{1}{Z} \exp(-\lambda b_i), \tag{3.15}$$

where $Z = \sum_i \exp(-\lambda b_i)$ is a normalization constant [171]. The Lagrangian quantity $\lambda$ is implicitly specified by the condition

$$B = \frac{1}{Z} \sum_i b_i \exp(-\lambda b_i). \tag{3.16}$$

Recall that we are interested in capacity per unit cost and not just capacity-cost. Such an optimization problem can also be easily solved, as pointed out by Balasubramanian et al. [172], by choosing $\lambda$ such that $Z = 1$, i.e.

$$\sum_i \exp(-\lambda b_i) = 1. \tag{3.17}$$

In this case, the probability expression (3.15) simplifies to

$$p_i = \exp(-\lambda b_i) \tag{3.18}$$

where $\lambda$ is defined by the condition $\sum_i p_i = 1$. This input distribution gives

$$C_{synapse} = \lambda \sum_i b_i \exp(-\lambda b_i), \tag{3.19}$$

which in turn implies that

$$\mathfrak{C}_{volume} = \frac{C_{synapse}}{B} = \frac{\lambda \sum_i b_i \exp(-\lambda b_i)}{\sum_i b_i \exp(-\lambda b_i)} = \lambda. \tag{3.20}$$

To make further progress, we assume that synaptic state volume is distributed equidistantly, i.e. the volume cost of the $i$th synaptic state is given by

$$b_i = V_0 + i2V_\nu. \tag{3.21}$$

This is exactly what is obtained when the volume-SNR exponent $\alpha$ from (3.1) is taken to be 1. With this assumption, we can rewrite the normalization condition as

$$\begin{aligned} 1 &= \sum_{i=0}^{\infty} \exp(-\lambda b_i) \\ &= \exp(-\lambda V_0) + \exp(-\lambda V_0) \sum_{i=1}^{\infty} \exp(-\lambda i 2V_\nu) \\ &= \exp(-\lambda V_0) + \frac{\exp(-\lambda(V_0 + 2V_\nu))}{1 - \exp(-\lambda 2V_\nu)} \end{aligned} \tag{3.22}$$

where we used the geometric series sum. Multiplying both sides of this expression by the denominator we find

$$\exp(-\lambda V_0) + \exp(-2\lambda V_\nu) = 1. \tag{3.23}$$

The average volume per potential synapse (including accessory volume, $V_0$) given by the total volume divided by the number of potential synapses (including actual ones) is:

$$\begin{aligned} B &= \sum_{i=0}^{\infty} b_i \exp\left(-\lambda b_i\right) \\ &= -\frac{\partial}{\partial \lambda} \sum_{i=0}^{\infty} \exp\left(-\lambda b_i\right) \\ &= -\frac{\partial}{\partial \lambda} \left( \frac{\exp\left(-\lambda V_0\right)}{1 - \exp\left(-2\lambda V_\nu\right)} \right) \\ &= \frac{2V_\nu \exp\left(-2\lambda V_\nu\right) + V_0 \exp\left(-\lambda V_0\right)}{\exp\left(-\lambda V_0\right)} \\ &= V_0 + 2V_\nu \exp\left(\lambda(V_0 - 2V_\nu)\right) \end{aligned} \tag{3.24}$$

In the limiting case when $V_0 \ll V_\nu$, these expressions reduce to:

$$B = 2\lambda V_\nu V_0, \tag{3.25}$$

and

$$\lambda V_0 = \exp(-2\lambda V_\nu). \tag{3.26}$$

In the opposite limiting case when $V_0 \gg V_\nu$,

$$B = V_0, \tag{3.27}$$

and

$$2\lambda V_\nu = \exp(-\lambda V_0). \tag{3.28}$$

To allow comparison with empirical measurements (Section 3.7), we also derive an

expression for the average volume of actual synapses, i.e. states with $i > 0$, and excluding accessory volume, $V_0$.

$$\langle V \rangle_{i>0} = \underset{q_i}{\mathrm{E}} \left[ b - V_0 \right], \tag{3.29}$$

where we have defined a distribution on states $i > 0$, $q_i$, which is the conditional probability of being in state $i$ given not in state 0. This simplifies as follows:

$$\begin{aligned}
\langle V \rangle_{i>0} &= \sum_{i=1}^{\infty} (b_i - V_0) q_i \qquad (3.30) \\
&= \sum_{i=1}^{\infty} 2iV_\nu \frac{\exp(-\lambda b_i)}{\exp(-2V_\nu)} \\
&= \exp(2\lambda V_\nu) \sum_{i=1}^{\infty} 2iV_\nu \exp(-\lambda b_i) \\
&= \exp(2\lambda V_\nu) \sum_{i=1}^{\infty} 2iV_\nu \exp(-(2iV_\nu + V_0)) \\
&= \exp(2\lambda V_\nu) \exp(-\lambda V_0) \sum_{i=1}^{\infty} 2iV_\nu \exp(-2\lambda V_\nu i) \\
&= -\exp(2\lambda V_\nu) \exp(-\lambda V_0) \frac{\partial}{\partial \lambda} \left\{ \sum_{i=1}^{\infty} \exp(-2\lambda V_\nu i) \right\} \\
&= -\exp(2\lambda V_\nu) \exp(-\lambda V_0) \frac{\partial}{\partial \lambda} \left\{ \frac{\exp(-2\lambda V_\nu)}{1 - \exp(-2\lambda V_\nu)} \right\} \\
&= \exp(2\lambda V_\nu) \exp(-\lambda V_0) \left\{ \frac{2V_\nu \exp(-2\lambda V_\nu)}{[\exp(-2\lambda V_\nu) - 1]^2} \right\} \\
&= \exp(2\lambda V_\nu) \exp(-\lambda V_0) \left\{ \frac{2V_\nu \exp(-2\lambda V_\nu)}{[\exp(-\lambda V_0)]^2} \right\} \\
&= 2V_\nu \exp(\lambda V_0),
\end{aligned}$$

where the steps follow by algebraic manipulations or by (3.23). The optimal average volume of actual synapses increases with the accessory volume. This result has an intuitive explanation: once the big investment in wiring ($V_0$) has already been made, it is advantageous to use bigger synapses that have higher SNR.

The ratio between the number of actual synapses and the number of potential

synapses (including actual) is called the filling fraction, $f$. In our model the filling fraction is just the fraction of synapses in states $i > 0$ and is given by

$$f = \exp(-2\lambda V_\nu) = 1 - \exp(-\lambda V_0). \tag{3.31}$$

Information storage capacity per volume can be calculated using (3.20) and (3.23). Just as in the AWGN model, information storage capacity increases monotonically with decreasing accessory volume. Unlike the AWGN model where storage capacity per unit cost is capped at 0.5, information storage capacity per unit cost can grow without bound. Since information storage capacity per unit cost diverges with decreasing accessory volume, $V_0$, optimal information storage is achieved when $V_0$ is as small as possible. As seen from (3.31), in the small $V_0$ limit, the filling fraction, $f$, is much less than one. This prediction is qualitatively consistent with empirical observations of sparse connectivity, but see the next paragraph and Section 3.7. In addition, most actual synapses have volume $2V_\nu$, and thus have SNR of order one, (3.1). This prediction is in agreement with the empirically established noisiness of typical synapses.

Although local cortical circuits are sparse and typical synapses are noisy, the filling fraction is not infinitesimal. As discussed in the previous section, system functionality would be adversely affected by very small $V_0$. The condition that accessory wire volume is of the order of synapse volume [166], along with (3.23), implies that $V_\nu \sim 1/\lambda$, meaning that information storage capacity of volume $V_\nu$ is on the order of one nat. This sets the filling fraction at a value below 1/2 but not infinitesimal.

By using (3.18) and (3.1), we can find the probability of synaptic states in terms of the EPSP amplitude,

$$p_i = \exp\left(-\lambda V_\nu \left(x_i/A_N\right)^\alpha\right). \tag{3.32}$$

These are samples from a stretched (or compressed) exponential distribution. In the continuum limit, when the probability changes smoothly between states, we can convert (3.32) to the probability density. Considering that there should be one synaptic

state per two noise amplitudes, $2A_N$, the probability density of the EPSP amplitudes is given by

$$p_X(x) = \frac{1}{A_N} \exp\left(-\lambda V_\nu \left(x/A_N\right)^\alpha\right), \tag{3.33}$$

which we compare with experiment in Section 3.7.

Interestingly, the explicit consideration of noise does not alter the result, following from (3.18), that for $V_0/V_\nu \to 0$ optimum information storage is achieved by using mostly the $i = 0$ state, with $i = 1$ used with exponentially low frequency. If $V_0 = 0$, this type of problem can be solved exactly [7,8] and the information storage capacity is maximized when, in addition to the zero cost symbol, only one other symbol is chosen. The additional symbol is chosen to maximize the relative entropy between conditional probabilities of that symbol and of the zero cost symbol divided by the cost of the additional symbol. If $V_0 > 0$, however, the problem of optimizing information storage capacity cannot be solved analytically, prompting us to pursue a reverse approach discussed in the next section.

To summarize the results of this section, we state two further general optimization principles.

**Optimization Principle 2.** *To optimize information storage capacity per unit volume, the filling fraction should be small, an exponential with exponent determined by the SNR exponent $\alpha$ and accessory volume $V_0$. Small filling fraction is equivalent to sparse synaptic connectivity.*

**Optimization Principle 3.** *To optimize information storage capacity per unit volume, although there will be many absent synapses and numerous small synapses, there will also be some large synapses, with synaptic efficacy distribution like a stretched exponential.*

## 3.5 Calculation of the Synaptic Cost Function from the Distribution of Synaptic Efficacies

In Section 3.4, we used an $\epsilon$-capacity-cost approximation to obtain a capacity-achieving input distribution. This was motivated by Condition 1 of the factorization interpretation of optimal information storage. The problem of directly and analytically finding the capacity-achieving input distribution and the channel capacity for a specified cost function is often rather difficult and is only known in closed form for certain special cases. In most cases, the channel capacity and capacity-achieving input distribution are found using numerical algorithms [174, 175]. In neuroscience, this algorithm was used by [172, 173] in the context of optimal information transmission by the spike code.

To move away from the $\epsilon$-capacity-cost approximation, we will apply conditions from the measure matching interpretation of optimal information storage. The channel conditional probability distribution and the channel input distribution can be measured experimentally and used in (2.51) to determine the single-letter cost function for which the system is operating at capacity-cost [95, 17]. This methodology does not seem to have been used for neuroscience investigations, other than a brief look at sensory processing [18]. As noted in Chapter 2, although this method inverts the problem specification, it seems reasonable if we are not sure of what the channel input cost function is (e.g. here we do not know what $\alpha$ is). The result of the computation is a cost function, which may then be examined for relevance to the problem at hand. Note that due to the arbitrary constants in (2.51), the cost function will be optimizing for any accessory volume $V_0$ and normalization constant $V_\nu$.

We use the dataset from [120, 163], also analyzed in [129], where EPSPs were recorded in several consecutive trials for each of 637 synapses from the cortex of several young rats. To carry out this calculation, we rely on the assumption that information stored at a synapse, $x$, can be identified with the mean EPSP amplitude. Then, the conditional density, $p_{Y|X}(y|X = x)$, is estimated for each synapse as the distribution of EPSP amplitude across trials, Figure 3-5A. The marginal density, $p_Y(y)$, is the

distribution of EPSP amplitude over all trials and synapses. By substituting these distributions into (2.51) we find estimates of the cost function, $V(x)$, for each synapse (Figure 3-5B). A power law with exponent 0.48 provides a satisfactory fit. Error bars are obtained from a bootstrapping procedure. Statistical procedures are given in the next paragraphs.

In order to compute the optimal cost function according to (2.51), we require the channel output distribution as well as the channel output distribution conditioned on the input. To estimate the channel output cumulative distribution function, $F_Y(y)$, we simply use the empirical cumulative distribution function, $F_{emp}(y)$. To account for the variable number of EPSPs acquired from each synaptic connection, the step size in the empirical cumulative distribution function contributed by each data point is inversely proportional to the number of EPSPs obtained from the synapse in question. To model the effect of absent synapses (which were not measured), we included in the empirical distribution function a narrow Gaussian distribution function with mean at zero EPSP amplitude and standard deviation of 0.1 mV (typical noise amplitude). The area under this Gaussian distribution function is given by one minus the filling fraction of 11.6% [129]. To estimate the channel conditional density, $p_{Y|X}(y|x)$, we assume that all EPSPs from a given synapse correspond to the same input letter; furthermore, we make a correspondence between the mean EPSP amplitude and this input letter. For each synapse, we use a histogram with 10 uniformly spaced bins to estimate the conditional density, $p_{emp}(y|x)$. Then the relative entropy is approximated by the following:

$$D(p_{emp}(y|x)\|p_{emp}(y)) = \sum_{i=1}^{10} p_{emp}(Y = y_i|X = x) \log \frac{p_{emp}(y \in y_i|X = x)}{F_{emp}(y = r_i) - F_{emp}(y = l_i)}, \tag{3.34}$$

where $y \in y_i$ implies presence in the $i$th histogram bin and the right and left bin edges are denoted $r_i$ and $l_i$ respectively. This relative entropy is computed for each synapse, and the result is the optimal cost function (Figure 3-5B).

We use the bootstrap method to determine confidence intervals. The confidence interval represented by the horizontal error bars is a small sample estimate of the

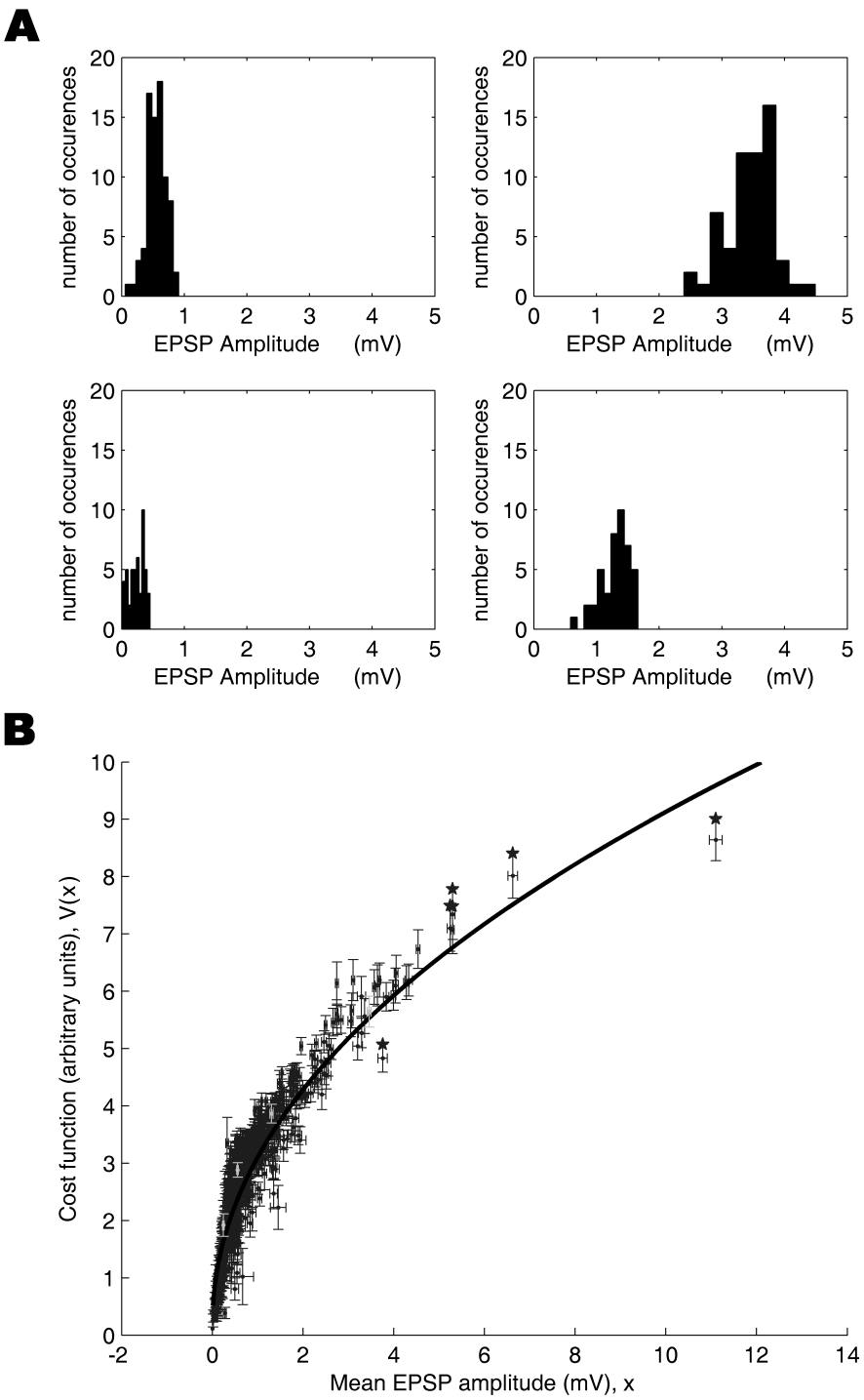


**Figure 3-5.** Optimizing synaptic cost function calculated from EPSP measurements. A. Typical distributions of EPSP amplitude among trials for synapses characterized by different mean EPSP amplitudes. B. Synaptic cost function as a function of mean EPSP amplitude calculated from (2.51) under assumption of optimal information storage. Each data point represents a different synapse, with those appearing in (A) highlighted in light gray. Horizontal error bars represent the standard error for the mean EPSP amplitude; vertical error bars represent the standard error for the relative entropy quantity in (2.51). The standard error was estimated by a bootstrap procedure. The points shown with starred upper vertical error had infinite vertical error, as estimated by the bootstrap procedure. The line shows a least squares fit by a power law with exponent 0.48.

standard error of the mean EPSP amplitude. The estimate is the sample standard deviation of the mean EPSP amplitudes generated by sampling with replacement from the empirical distribution that were measured for each synapse. The result is based on 50 bootstrap trials. The confidence interval represented by the vertical error bars is a small sample estimate of the standard error of the relative entropy quantity in (3.34). In order to do the sampling with replacement for the bootstrap procedure, randomness is introduced in two ways. First, 637 synapses are selected uniformly with replacement from the set of 637 measured synapses. These resampled synapses are used to generate the empirical cumulative distribution function, $F_{emp}(y)$. Second, for each of the measured synapses, EPSP values are selected uniformly with replacement to produce the empirical conditional probability mass functions $p_{emp}(y|X = x)$ for each of the 637 values of $x$. Finally, (3.34) is used to calculate a bootstrap version of the data points in Figure 3-5B. This doubly stochastic resampling procedure is repeated for 50 bootstrap trials, and the estimate is the sample standard deviation of the relative entropy quantities for each of the synapses. When we resample the synapses to generate the $F_{emp}(y)$ for each of the bootstrap trials, there is a possibility that no large EPSP amplitude synapse is selected. In such a case, the relative entropy quantity will come out to be infinite since the output distribution will be zero where the conditional distribution will be non-zero, so the two distributions will not be absolutely continuous with respect to each other. In actuality, such lack of absolute continuity should not occur because of the relationship $p_Y(y) = \int p_{Y|X}(y|x)p_X(x)dx$. This quirky phenomenon of the distributions not being absolutely continuous with respect to each other results in an infinite discontinuity in the relative entropy functional and thus causes the infinite upper standard deviation errors represented by stars in Figure 3-5B. For these points, the lower standard deviation is estimated by excluding points greater than the mean. For other points, the bootstrap distributions are approximately symmetric, so the vertical error bars are symmetrically replicated. The unweighted least squares fit in the MATLAB curve fitting toolbox is used to generate the fit to a function of the form $\nu x^{\eta}$ with $\eta = 0.48$.

In this section, we did not formulate new optimization principles but provided a

reverse characterization for Optimization Principles 2 and 3. In fact we essentially assumed the results of Optimization Principles 2 and 3, and saw that the cost function for which these principles hold falls neatly into the power law cost function that we had initially formalized as (3.1). In Section 3.7 we will make further quantitative comparisons regarding our predictions.

## ■ 3.6 Discrete Synapses May Provide Optimal Information Storage

The models of synaptic information storage presented in Sections 3.3 and 3.5 might have given the impression that the optimal distribution of synaptic strength must be continuous. Indeed, Section 3.3 modeled information storage in synapses by the AWGN channel with average power constraint, for which the optimal (capacity-achieving) distribution is the continuous Gaussian distribution [9]. In addition, using the methods of Section 3.5, one can construct innumerable average cost constrained channels with capacity-achieving distributions that are continuous.

There is experimental evidence, albeit controversial, that synaptic weights change in discrete steps [132, 133, 134]. Implicit in Question 4, can these experimental observations be consistent with the framework of optimal information storage? Here we make two observations that suggest that discrete synaptic states may achieve optimal, or close to optimal, information storage. First, we point out that, surprisingly, not all continuous input channels have optimal input distributions that are continuous. In particular, imposing a constraint on the maximum weight (or volume) of a single synapse may change the optimal, capacity-achieving distribution of synaptic weights from continuous form to a set of discrete values. Such a maximum amplitude constraint is quite natural from the biological point of view, because neither volume nor EPSP can be infinitely large. Note that, unlike in Section 3.4, where discreteness was an assumption used to simplify mathematical analysis, here the discrete solution emerges as a result of optimization.

For concreteness, we return to the AWGN channel model considered in Section 3.3, but now we impose a maximum volume constraint in addition to the average volume

constraint that was originally imposed. The problem then reduces to the well-studied problem of maximizing channel capacity for a given average input power and peak input power. For the AWGN channel, the unique optimal input distribution consists of a finite set of points. A proof of this fact, based on methods of convex optimization and mathematical analysis, is given in [176]. Note that the Blahut-Arimoto algorithm for continuous channels is based on sampling the input space [175], and cannot be used to determine whether the optimal input distribution is continuous or discrete; an analytical proof is necessary.

Since it is known that the optimal input distribution consists of a finite number of points, one can numerically search over this sequence of finite-dimensional spaces to find the locations and probabilities of these points for particular average power and peak power values. Moreover, there is a test procedure, based on the necessity of satisfying the Karush-Kuhn-Tucker conditions, to determine whether the obtained numerical solution is in fact optimal. So one can apply the numerical procedure to generate a possible solution and unmistakably recognize whether this solution is optimal [176]. Applying Smith's optimization procedure, including both the search and the test for optimality, yields the following result for the AWGN channel. For noise power 1, symmetric peak amplitude constraint $[-1.5, 1.5]$, and input power constraint 1.125 (an SNR close to 1), the optimal input distribution consists of the zero point with probability 1/2, and the $-1.5$ and 1.5 points with equal probabilities of 1/4 [176].

The conclusion that the distribution of synaptic weights should be discrete-valued holds not only for the AWGN channel with hard limits imposed on synapse size and weight, but also for other noise models. In particular, the discreteness result holds for a wide class of additive noise channels under maximum amplitude constraint [177]. Some fading channels that have both additive and multiplicative noise and are similarly constrained or even less stringently constrained also have this discrete input property [178,87]. Furthermore, channels other than AWGN with constraints on both average power and maximum amplitude have optimal input distributions that consist of a finite number of mass points [179].

A second observation is that although some channels have optimal input distributions that have discrete, even channels that have continuous optimal input distributions can be used with discrete approximations of the optimal input distribution and perform nearly at capacity. In the average power constrained AWGN example, it is well known that in the limit of small SNR, using an alphabet with just two symbols ($\pm\langle X^2\rangle^{1/2}$) does not significantly reduce information storage capacity (see e.g. [180]). In addition, Huang and Meyn [179] demonstrate numerically that discrete input distributions, in some cases generated by sampling the optimal continuous distribution, are only slightly suboptimal. Although it is not clear a priori that a distribution close to the capacity-achieving one will achieve mutual information close to capacity, since the space of probability distributions is not compact without imposition of constraints, this is almost always the case in practice. Existence or uniqueness of the capacity-achieving input distribution is not guaranteed either, but must be shown in the manner of Section 2.6.1.

Although we cannot formulate this as a general optimization principle because a sufficiently fine characterization of noise is not available, extant analysis of optimal information storage under hard volume constraints suggests that synaptic weights should take a set of discrete values. Experimental observations of discrete changes in synaptic weights are therefore consistent with optimization of information storage. In addition, even if the capacity-achieving input distribution is continuous, the use of discrete rather than continuous levels is not likely to reduce information storage capacity appreciably.

## ■ 3.7 Theoretical Predictions and Comparison to Experiment

In this section, we compare the theoretical predictions made in the previous sections with known experimental facts and also suggest further experimental tests of the theory. This is the final step in the scientific method described at the beginning of the chapter, though of course the steps should be repeated iteratively, if experimental findings do not confirm the theory.

In Section 3.3, by considering an AWGN channel we find that information storage is optimized by synapses with average volume given by the geometric mean of $V_\nu$ and $V_0$ (in the relevant regime $V_0 < V_\nu$). Although storage capacity increases in the limit $V_0 \to 0$, the conduction time delay grows without bound. Although delay is not explicitly considered in our performance metrics, it is certainly an important issue. Since the minimum conduction time delay is achieved when synaptic volume is of the order of the accessory volume [165, 166], the competition between these requirements would yield the average synaptic volume less or equal to $V_\nu$, though an explicit characterization of the tradeoff between storage capacity and propagation delay has not been made. Therefore, typical synapses should be noisy, in agreement with experimental data where the signal to noise ratio is found to one or less. This result is corroborated in Section 3.4, where optimal synaptic volume was found to be $2V_\nu$. Although the noise is represented only through the hard sphere packing, the optimal synapse volume is the minimum possible with the discrete states model.

In Section 3.4, we argue that optimal information storage requires sparseness of synaptic connectivity and, assuming that synaptic states are equidistant in volume space, predict a relationship between the filling fraction and the relative volume occupied by synapses and accessory volume. To make a quantitative comparison with empirical observations, we use experimental data from a mouse cortical column. Potential synaptic connectivity in a cortical column is all-to-all, meaning that axons and dendrites of any two neurons pass sufficiently close to each other that they can be connected through local synaptogenesis [158, 154]. According to [181], the fraction of potential synapses converted into actual synapses in a mouse cortex is $\sim 0.3$. We take this fraction to be our filling fraction, $f = 0.3$. Then by (3.31), we find $2\lambda V_\nu = -\ln 0.3 = 1.2$ and by (3.23), we find that $\lambda V_0 = -\ln 0.7 = 0.36$. Then, the average volume of actual synapse is of the same order as the accessory volume per actual synapse $V_0/f$, in agreement with experiment. A more detailed calculation using (3.30) on the ratio between the actual synapse volume $\langle V \rangle_{i>0}$ and the accessory

volume per actual synapse $V_0/f$,

$$\begin{aligned}\frac{1}{K} &= \frac{V_0}{f}\frac{1}{\langle V\rangle_{i>0}} = \frac{V_0}{2V_\nu f\exp(\lambda V_0)} \\ &= \frac{V_0}{V_\nu}\frac{(1-f)}{2f} \\ &= \frac{\log(1-f)}{\log f}\frac{1-f}{f}\end{aligned} \tag{3.35}$$

shows that actual synapse volume should be about 40% greater than accessory volume for an actual synapse. In reality, accessory volume is greater than synapse volume. This may be a consequence of minimizing conduction delays as discussed in Section 3.4. Hopefully, an optimization framework combining conduction delays and information storage capacity will emerge in the future. The anytime information theory framework seems poised in that direction [52, 40].

One might wonder whether the sparseness theory applies to the entire brain network beyond just the cortical column. In principle, the sparseness of the global connectivity seems consistent with the high cost of wiring. A detailed quantitative analysis seems difficult, however, because the network does not possess all-to-all potential connectivity, and $V_0$ would change. The cost and benefit of growing axonal arbors would need to be quantified, however this is not straightforward since a single axon can implement synaptic connections with several neurons at a time; furthermore, possible correlations in synaptic connectivity patterns would need to be taken into account.

In Section 3.4, we predicted that synaptic volume should follow an exponential distribution with parameter $\lambda$, (3.18). This prediction can be tested experimentally by measuring the volume of spine heads and boutons in cortical neuropil. In comparing the distribution of volume, our definition of a synapse as all synaptic contacts between two neurons should be remembered. If in addition, an experimental measurement of the filling fraction in the same neuropil is made, the theory can be tested in a way that involves no parameter fitting. This is because $\lambda$ can be calculated from the wiring volume and the filling fraction: from (3.23) and (3.31), we get that $\lambda =$

$-\log(1-f)/V_0$. To overcome the difficulty in measuring $V_\nu$ or $V_0$, one can alternatively measure the experimentally accessible quantities $f$ and $\langle V\rangle_{i>0}$ to determine $\lambda$. Then, from (3.30) and (3.31),

$$\lambda = \frac{-\log(f)}{(1-f)\langle V\rangle_{i>0}}. \tag{3.36}$$

However, these predictions are only approximate as the relative importance of maximizing information storage and minimizing conduction delays has not been characterized.

In Section 3.4, we predict the distribution of synaptic weight for arbitrary values of the volume-SNR exponent $\alpha$ (3.32), which can be compared to the experimentally observed synaptic weight distribution obtained in neocortical layer 5 neurons [129]. To perform such a comparison, we sort the input distribution synaptic weights into bins $[x_i - A_N(x_i), x_i + A_N(x_i)]$ and plot a histogram, see Figure 3-6. By performing a least squares fit of the logarithm of the EPSP distribution we find that the distribution is a stretched exponential with exponent 0.49. Recall that performing a least squares fit of the standard deviation of EPSP amplitude as a function of mean EPSP amplitude, Figure 3-3, yields a power law with exponent 0.38. Hence, $x/A_N \sim x^{0.62}$, and from (3.32), we find that $\alpha = 0.49/0.62 = 0.79$.

In Section 3.5, we had established a measure matching link from the distribution of synaptic weights and noise statistics to the synaptic cost function. The best power-law fit to the points in Figure 3-5B yields a sublinear cost function with power law exponent $\sim 0.48$. Recalling that $x/A_N \sim x^{0.62}$, we find that $\alpha = 0.48/0.62 = 0.77$. This estimate is consistent with that obtained from the discrete states model in the previous paragraph, thus validating the use of that model to approximate the continuous distribution of synaptic weights. The prediction of $\alpha$ can be tested directly by measuring the relationship between synaptic volume and weight and would be a major test of our theory.

In Section 3.6 we argued that discrete synaptic states could achieve information storage performance better than or almost as well as synapses with continuous weights. Then, the discreteness in the weight changes of individual synapses that have

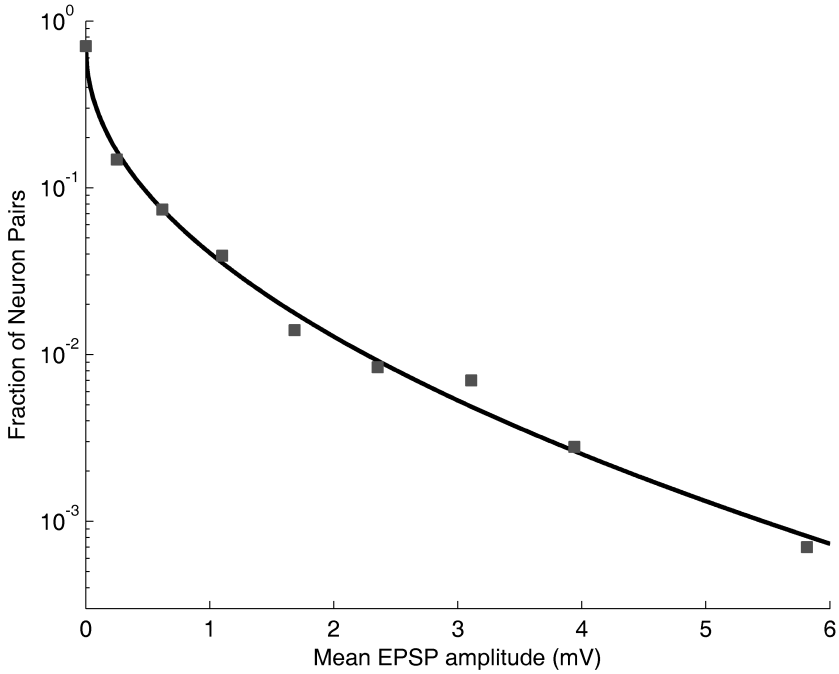


**Figure 3-6.** Distribution of synaptic weights sorted into bins whose width is given by two standard deviations for the corresponding EPSP amplitude. Squares represent the fraction of neuron pairs belonging to a bin centered on that synaptic weight. The line is a stretched exponential fit (i.e. power law in the log domain) with exponent 0.49.

been observed experimentally [132, 133, 134] seem consistent with maximizing information storage. However, it should be noted that this theoretical result implies that the distribution of synaptic weights among all of the synapses—not just the weight of an individual synapse—is discrete. The fact that such a discrete distribution has not been seen is not surprising because the theory relies on some simplifying assumptions, such as identical synaptic weight/volume relationship among all synapses. Since the relationship between synaptic weight and volume depends on the electrotonic distance from the synapse on a dendrite to the integration site [182], it can differ from synapse to synapse. Although the optimal solution is not known in this case, we speculate that, even if individual synapses were to have discrete states, these states would not be the same among all the synapses.

## 3.8 Discussion

In this chapter, we have applied information theoretic characterizations of optimal information storage to generate a unified theory of some structural characteristics of the mammalian brain. The main tool that was used was that the distribution of synaptic efficacies should be the capacity-achieving input distribution for the neural memory channel. The predictions and explanations are of a purely informational and physical nature; specific implementation issues such as error control codes or neural networks have been abstracted. Independent of the coding mechanism, however, the predictions and explanations that we make are required for optimal performance.

In some sense, the abstraction of the specific implementation can be considered as a weakness, since unknown mechanistic constraints have not been incorporated. There could be constraints due to mechanisms of storage and retrieval, as well as operation requirements on the network. Neural network models commonly assume specific mechanisms and yield information storage capacity estimates different from ours [122, 123, 124, 125, 126]. Interestingly, Brunel et al. [122] predict a distribution of synaptic weights similar to ours, although results such as this one may depend on the details of the neural network model at hand. Future research is likely to shed more

light on the biological mechanisms that shape and constrain information storage and retrieval.

As our analysis relies on optimizing information storage capacity, it is not applicable to brain regions for which information storage is not the main task. For example, synapses associated with early sensory processing, e.g. in the retina [114, 183], or those belonging to motorneurons [147, 149] may be large and reliable. This would be consistent with optimizing information transmission. In actuality, any given brain circuit probably contributes to both information storage and information transmission. Indeed, by applying our analysis in reverse, one could infer the role of a given circuit from its structural characteristics. In particular, different cortical layers may be optimized for a different combination of storage and processing.

Our formulation of synaptic memory in the optimal information storage framework implicitly casts each synapse—both potential and actual—as a channel usage; the total storage capacity being the number of synapses multiplied by the average storage capacity of each. Incidentally, this makes the storage capacity on the order of the number of synapses, which would correspond to an overall maximal storage capacity of several information kilobits for a neocortical L5 pyramidal neuron [184]. It is possible, however, that the synaptic information retrieval mechanism involves multiple read-out attempts from a single synapse. Since each channel usage is separated in space rather than in time, this does not increase the number of channel usages. Regardless, one may wonder what impact multiple read-out attempts would have on our analysis of information storage capacity.

It is known that the SNR increases approximately as the square root of the number of read-out trials for most forms of signal integration [185], so if the information stored in each synapse was retrieved using the same number of read-out attempts, this simply introduces a fixed multiplicative constant in (3.1). A fixed constant in (3.1) can simply be incorporated into the $V_\nu$ term and all of our results stand. On the other hand, if the number of read-out attempts is not fixed, but varies across different synapses, then it would cast much of our analysis into doubt. We point out, however, that multiple read-out attempts would lead to large time delays. As has been mentioned

throughout the chapter with reference to minimal accessory volume, time delay is also an important issue and may constrain the system. If information is used to control dynamical systems, it is known that large delay can be disastrous [40]. In addition, it is not clear how short-term plasticity caused by multiple read-out attempts would be overcome.

Other possible concerns arise from the lack of a true experimentally established input-output characterization of synaptic memory. Addressing this concern would require identification and description of the so-called engram, the physical embodiment of memory, which corresponds to the channel input, $X$ (Figure 3-1). In addition, it would necessitate a better characterization of the noise process which determines the input-output probability distribution, $p_{Y|X}(y|x)$. Description of the alphabet $\mathcal{X}$ would furthermore settle the question, alluded to in Section 3.6, of whether synapses are discrete-valued or continuously graded. In addition, we assumed in Section 3.5 that the channel input letter $x$ is given by the arithmetic mean of EPSPs observed in several trials. Alternatives to this assumption may alter the horizontal coordinate of points in Figure 3-5B.

Although our analysis relies on identifying synaptic noise with the variability of EPSP amplitude on read-out, the noise may come from other sources. The experiments that we have used have only been able to access a portion of the channel, the lower part of Figure 3-7. Perhaps, the main concern is that long-term memory storage at a synapse is open to perturbations due to active learning rules and ongoing neuronal activity [186], the so-called in situ noise (Figures 3-1 and 3-7). The longer the information is stored, the greater the perturbations caused by such processes (although see [187]). Amit and Fusi [188] have demonstrated that under mild assumptions, this noise restricts memory capacity significantly. Fusi, Drew, and Abbott [189] have recently proposed a solution with a cascade of discrete synaptic states with different transition probabilities, resulting in a form of metaplasticity that increases retention times in the face of ongoing activity. Presumably, other forms of metaplasticity may also help protect stored information from unwanted perturbations. In addition, the stability of physiological synaptic plasticity appears to depend critically on the details

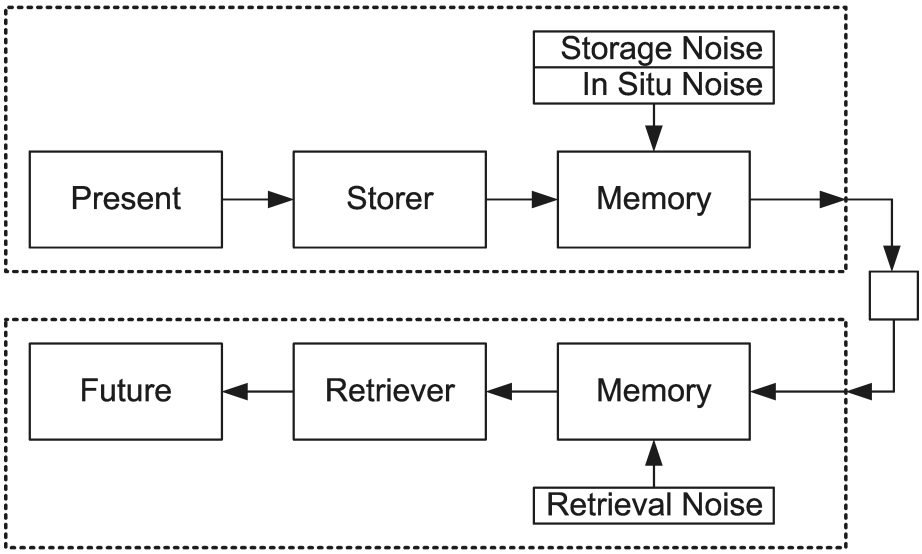


**Figure 3-7.** Partitioned neural memory system. Current experiments have only been able to characterize the portion of the channel from the unlabeled box to the retriever.

of activity patterns during and after the induction of plasticity [186], suggesting that specific biological mechanisms for the protection of stored information may exist.

Our theory can be modified to include sources of noise other than retrieval noise. For example, if in situ noise is quantified and turns out to be dominant, it can be used in the calculations presented in Sections 3.3-3.5. In fact, optimality of noisy synapses (Section 3.3) may be relevant to resolving the issue of protecting information from in situ noise. In general, a better understanding of the system functionality including characterization of storage, in situ, and retrieval noise should help specify $p_{Y|X}(y|x)$ in the future.

Finally, our contributions include not only many explanations and predictions of physical structures, but also the introduction of new methods to the study of memory in the brain. To develop the optimization principles, we have applied information theory results from Chapter 2 to the study of physical neural memory systems. Moreover, our application of the measure matching conditions to study the cost function (Section 3.5) appears to be the first instance where the measure matching characterization of optimality has been successfully applied to real system analysis, whether biological or human engineered. The measure matching problem inversion has wide applicability to the experimental study of information systems.

Chapter 4

# Source Coding for Nonsequential Sources

The importance of invariant representations for task relevant information storage was introduced in Chapter 1, and the notion of nonsequential semantics corresponding to permutation invariance was emphasized. Such invariance properties lead to fidelity criteria that are not single-letter but look at the entire source sequence collectively. Historically, the main focus of rate distortion theory has been on single-letter fidelity criteria, due to mathematical tractability. In Section 2.4.1, we were able to obtain a rate-distortion characterization with a non-single-letter fidelity criterion, and here we would like to do the same for nonsequential fidelity criteria.

When source sequences are invariant with respect to the fidelity criterion, a disjoint partitioning of the source sequences into equivalence classes is induced. Subsets of source sequences that are permutations of one another are the equivalence classes with respect to nonsequential fidelity criteria. These permutation-invariant equivalence classes are also known as composition classes and type classes. A set of class representatives is a subset of source sequences which contains exactly one element from each equivalence class. A general principle of source coding for invariant representations is to project the set of source sequences down to a set of class representatives, thereby reducing the dimension and the uncertainty of the source generating process without increasing the distortion with respect to the fidelity criterion.

In the original proof of the Slepian-Wolf theorem [190] (as opposed to the proof by random binning [191]), the set of source sequences is also partitioned into classes.

The partitioning of the source space is accomplished by providing side information to the decoder at a certain positive rate. The data sequence is then used to specify which equivalence class is to be used in the reconstruction. In contrast to the Slepian-Wolf scenario, the partitioning induced by the semantic fidelity criterion does not require any rate, and may be thought of as side information that comes for free. Moreover, in the invariant representation situation, the same element of the equivalence class need not be used as the class representative all the time [14].

## ■ 4.1 Nonsequential Source-Destination Semantics in Applications

Before detailing the results we can garner for information storage systems with nonsequential fidelity criteria, we review a few information storage applications where these semantics arise naturally. To tie things together with Chapter 3, the first application we present is human memory, though from the psychological perspective rather than the neurobiological perspective. In psychological experiments on recognition and recall, the nonsequential nature of the source is forced. In a recall experiment, a subject is asked to remember some number of symbols and later asked to recall as many as possible (without consideration of order). In recognition experiments, a subject is first shown a set, $\mathcal{G}$, of symbols selected from a set $\mathcal{U}$ and later is presented with a set $\mathcal{K}$, $\mathcal{G} \subset \mathcal{K} \subset \mathcal{U}$, from which the choice is made. Again, performance is measured without reference to ordering [192]. Presumably each subject will have a canonical representation of the set of symbols based on some personal total ordering principle. For example, one can imagine several ordering principles for the United States. Possible ordering principles include alphabetic: Alabama $<$ Alaska $< \cdots <$ Wyoming; geographic: Maine $<$ New Hampshire $< \cdots <$ Alaska; or historic: Delaware $<$ Pennsylvania $< \cdots <$ Hawaii. The reason that there are several ordering principles is that any curve through all members of a finite set generates a total order; in fact there are $|\mathcal{U}|!$ total orders for a finite set $\mathcal{U}$. Judd and Sutherland's psychology experiment design [192] actually uses a model of human memory as a channel with synchronization errors [24, 25, 26]. As delineated in Chapter 2, we will only consider channels where

there are the same number of outputs as inputs.

Another application area for the nonsequential fidelity criteria, at least in a partial sense, is in the storage of computer programs or other descriptions of algorithms [193]. In the psychological realm, this can be thought of as memory of how to perform a task like tying a shoe. In computer programs, which can be thought of as descriptions of dynamical processes, some of the steps can be rearranged without any change in the behavior. For example the following two programs are functionally equivalent.

$$\begin{aligned} a &= 7 \\ b &= 8 \\ c &= a + b \\ d &= b - a \\ e &= c + d \end{aligned} \tag{4.1}$$

$$\begin{aligned} b &= 8 \\ a &= 7 \\ d &= b - a \\ c &= a + b \\ e &= c + d \end{aligned} \tag{4.2}$$

Similarly, in culinary recipes, instructions for fabricating objects, and deterministic control policies, there are often interchangeable steps.

The storage of databases of scientific data is another application where reordering does not affect utility. A fundamental assumption in science is that identical, unlinked, repeatable experiments yield results that follow an exchangeable joint distribution. Usually there is the stronger assumption that the results are i.i.d. Moreover, the order of the collected data is not used for any scientific purpose and is treated as an arbitrary property of data. To take an example, consider the EPSP

amplitude observations, depicted in Figure 3-5A, and used to calculate the horizontal coordinate of a data point in Figure 3-5B. Since the horizontal coordinate in the plot is the sample mean of the observation values, it is invariant to permutation of the data. In fact, if a sample consists of independent observations from the same distribution, any minimum variance unbiased estimator is invariant to permutations in the observations [194, p. 259].

Not only are databases of scientific data invariant to reordering of rows and columns, general databases often also share this feature. Massive tables with fixed-length records and fields often arise in such applications as logging of network traffic data in telephone systems and financial transactions in banking systems [195]. For these applications, several source coding algorithms based on reordering the rows, the columns, or both have been developed to significantly improve operational performance [196, 197], however general information theoretic results on optimal performance of such information storage systems have not been developed. That is the purpose of the remainder of the chapter.

## ■ 4.2 Nonsequential Semantics in Information Storage Problems

We will eventually formalize our notion of nonsequential semantics in terms of fidelity criteria, but for the current discussion, suffice it to say that when the retrieved vector $\vec{v}$ is a permuted version of the stored vector $\vec{u}$, there is no distortion. That is to say, the order of elements in the vectors is not considered when computing distortion. Due to the fidelity criteria that we adopt, source outputs are intrinsically robust to channel perturbations that cause permutations. Recall Shannon's statement [9], "if the source already has a certain redundancy and no attempt is made to eliminate it in matching to the channel, this redundancy will help combat noise." Judd and Sutherland model human memory in tests of recall and recognition as just such a random permutation channel [192]. Such a channel is also used as a rough model of neural communication [14, 17]. For concreteness, consider all alphabets to be binary sequences of length $k$ and consider the channel that performs an equiprobable random permutation. No

coding is used, so the source distribution is the channel input distribution. This distribution, expressed in terms of the Hamming weight, $w = w_H(u_1^k)$, is chosen to be

$$p_W(w) = \frac{\gamma^k(\gamma - 1)}{(\gamma^{k+1} - 1)\gamma^w}, \qquad w \in \{1, 2, \ldots, k\}, \tag{4.3}$$

for some $\gamma > 0$. Using the distortion function portion of the measure matching conditions (Theorem 8), the optimality-inducing distortion measure is found, in [14, 17], to be

$$d(\vec{u}, \vec{v}) = \begin{cases} 0, & \text{type}(\vec{u}) = \text{type}(\vec{v}) \\ \infty, & \text{type}(\vec{u}) \neq \text{type}(\vec{v}) \end{cases} \tag{4.4}$$

where the type$(\cdot)$ operator determines the type of its argument and is equivalent to the Hamming weight here. As seen, the fidelity criterion generated by this distortion function has the nonsequential semantics, as it depends only on the type, which is invariant to permutation.

Nonsequential semantics can also be used for the embedding of covert information. In particular, if the adversary does not realize that the destination does not care about the order of the cover text data, information can be embedded in the ordering. One can think of the covert ordering channel as a very degenerate form of a covert timing channel, as in Section 2.6.1. This method of information embedding is often useful for card tricks, Fitch's Five Card Trick being a prime example.[1] In the trick, covert communication is established between the magician and his/her assistant by encoding information in the order of cards. Since the order of cards is irrelevant in games such as poker and most instances where playing cards are used, an audience does not expect information to be embedded, thereby establishing the covert communication. If the deck is arranged in a particular way, based on de Bruijn cycles, a similar trick can be performed without an assistant: the encoding is already in the deck [200].

The information embedding problem and the problem of source coding with side information are closely related, and in fact duals of one another [201]. As we had

[1]This card trick and variations of it seem to have been popularized in the information theory community by Elwyn Berlekamp and Thomas Cover [198], among others. For extensions, achievability proofs, and information theoretic converses, see [199] and references therein.

mentioned in the opening of this chapter, source coding for systems where order is irrelevant can be thought of as a problem of source coding with access to free side information. The main focus of this chapter will be on the source coding problem. The approach that we take will be based on the rate matching interpretation of optimal information storage (Section 2.4) so we will be concerned with the tradeoff between rate and distortion. The storage mechanisms that we consider will sort the source sequence into a fixed order as a step before any other encoding. One can view the sorting procedure as a nonlinear filter that dissipates exactly the irrelevant entropy associated with ordering uncertainty.

## ■ 4.3 Separating Order and Value

Transform coding has been the workhorse of practical source coding [202], with a great bias towards linear, orthogonal transforms. The Karhunen-Loève Transform (KLT) is often considered as an optimal transform since it yields independent transform coefficients for Gaussian sources. For non-Gaussian sources, however, removing the orthogonality and linearity restrictions on the transform can often yield better source coding results, e.g. transformation to pseudophase and pseudomagnitude yields improvements for Gaussian scale mixture sources [203]. In fact, the pseudophase and pseudomagnitude representation yields independent components.

A *set* is a collection of objects in which neither order nor multiplicity have any significance. A *multiset* is a collection of objects in which order has no significance, but multiplicity is explicitly significant. A *sequence* is a collection of objects in which order and multiplicity both have explicit significance. Consider the source variables $U_1, U_2, \ldots, U_k$ drawn from the common alphabet $\mathcal{U}$ according to a joint distribution that is exchangeable. Denote the random $k$-tuple $U_1^k$ as $(U_i)_{i=1}^k$. A realization $(u_i)_{i=1}^k$ can be decomposed by filtering into a multiset of values, denoted $\{u_i\}_{i=1}^k$, and an order $j_k$. Here we consider sorting as a nonlinear transform that produces order and value as "transform coefficients." We use a collapsed two-line notation for permutations to

express the output of sorting.

$$(u_i)_{i=1}^k \xrightarrow{\text{sort}} \begin{pmatrix} i_1 & i_2 & \cdots & i_k \\ u_1 & u_2 & \cdots & u_k \end{pmatrix} = \begin{pmatrix} j_k \\ \{u_i\}_{i=1}^k \end{pmatrix}, \tag{4.5}$$

where the indices $i_1, \ldots, i_k$ are a permutation of the integers $1, \ldots, k$ for sets, but may have repetitions for multisets. The ordering is collapsed into a single variable $j_k$. Similarly, the values $u_1, \ldots, u_k$ are collapsed into a set $\{u_i\}_{i=1}^k$.

For exchangeable sources, the two transform coefficients, $J_k$ (order) and $\{U_i\}_{i=1}^k$ (value), are statistically independent chance variables. This can be expressed in information theoretic terms by decomposing the entropy into two independent parts. Define $H((U_i)_{i=1}^k)$ as entropy of the original sequence, $H(\{U_i\}_{i=1}^k)$ as multiset entropy, and $H(J_k)$ as the entropy of the chance variable representing order. In the statement of the theorem, we suppress subscripts used to denote length.

**Theorem 12.** *For exchangeable sources, the order and the value are independent and the sequence entropy can be decomposed into the order entropy and the value entropy:*

$$H((U)) = H(\{U\}) + H(J). \tag{4.6}$$

*Proof.*

$$\begin{aligned} H((U)) &\overset{(a)}{=} H((X)) + H(\{X\}) - H((X)|J) \\ &= H(\{X\}) + I((X); J) \\ &\overset{(b)}{=} H(\{X\}) + H(J) - H(J|(X)) \\ &= H(\{X\}) + H(J), \end{aligned} \tag{4.7}$$

where (a) follows from noting that $H(\{X\}) = H((X)|J)$ for exchangeable sources, since all orderings are equiprobable and uninformative about the value. The step (b) follows from the fact $H(J|(X)) = 0$ in general. The other steps are simple informational manipulations. □

Since the transform coefficients are independent, we can code them separately

without loss of optimality. In fact, we will ignore the order $J$ and only code the value. In a more general setting, a rate allocation favoring one or the other may be used, much as some subbands are favored in linear transform coding.

Since source coding considered in this chapter deals with a sorting preprocessing step to generate a multiset from a sequence, it behooves us to have a convenient representation of multisets. In particular, for the sequel, we want to deal with complete, minimal sufficient statistics of multisets rather than the multisets themselves. For a multiset, especially if drawn from a discrete alphabet, the number of occurrences of each $u \in \mathcal{U}$ fully specifies it. That is to say, the type is a sufficient statistic for the multiset for any arbitrary parameter estimation problem and is in fact minimal. In the continuous alphabet case, rather than working directly from types, we use the sequence in which the elements are in ascending order, a natural representative of a permutation-invariant equivalence class. This canonical sequence representation is equivalent to type representation and naturally leads to the framework of order statistics. Assuming that the parent alphabet consists of the real line, the basic distribution theory of order statistics can be used [204]. When the sequence of random variables $X_1, \ldots, X_n$ is arranged in ascending order as $X_{(1:n)} \leq \cdots \leq X_{(n:n)}$, $X_{(r:n)}$ is called the $r$th order statistic. It can be shown that the order statistics for exchangeable variates are complete, minimal sufficient statistics [204]. Since the order statistics are sufficient and complete statistics, an alternate proof of Theorem 12 can be constructed using Basu's Theorem on the independence of a sufficient, complete statistic for a parameter and any ancillary statistic whose distribution does not depend on that parameter. Here the parameter is the multiset, the sufficient statistic is the sequence of order statistics, and the ancillary statistic is the order. In contrast to discrete alphabets and real-valued scalar alphabets, for alphabets of vectors of real numbers there is no simple canonical form for expressing the minimal sufficient statistic, since there is no natural ordering of vectors [205].

## 4.4 Rate Distortion Theory for Growing Multisets

When order is irrelevant, the source coding problem reduces to a problem of coding multisets. Communicating a multiset of size $k$ should, of course, require less rate than communicating a vector of length $k$, since there is no need to distinguish between vectors that are permuted versions of each other. The question that arises is how much rate can be saved. Since there are $k!$ possible permutations, it would seem that a rate savings of $\log k!$ would enter in some way. If however, $\log k!$ turns out to be greater than $kR(0)$ (for a rate-distortion function with respect to a fidelity criterion that explicitly considers order), things would not make sense by the information inequality; there would need to be more to the story. In this section we present results for multisets that grow with the block length $k$, discussing cases with discrete and continuous alphabets separately.

### 4.4.1 Discrete Alphabets

In the finite situation, the multiset of values that we want to code can be represented by the type of the multiset. Consequently, we do not require an explicit order to be defined on the source alphabet. An alternative view is that an arbitrary curve through the alphabet can be used to define an order. As a result, the theory applies not just to alphabets of numerical scalars, but to alphabets of numerical vectors, language texts, images, and political subdivisions of republics. Since the multiset is equivalent to a type, we can use the type to define the word distortion measure and the fidelity criterion. The word distortion measure of length $k$ is defined to be

$$d_k(u_1^k, v_1^k) = \begin{cases} 0, & \text{type}(u_1^k) = \text{type}(v_1^k) \\ 1, & \text{type}(u_1^k) \neq \text{type}(v_1^k) \end{cases} \tag{4.8}$$

and the fidelity criterion is

$$F_{\text{val}_1} = \{d_k(u_1^k, v_1^k), k = 1, 2, \ldots\}. \tag{4.9}$$

This is not a single-letter fidelity criterion on the source, though it does have a frequency of error interpretation on the types. Due to this frequency of error interpretation, the $R(0)$ point is the entropy of the multiset, $H(\{U_i\}_{i=1}^k)$. Since there are only $k!$ possible orderings, the alphabet-size upper bound on $H(J_k)$ and Theorem 12 lead to the lower bound

$$H(\{U_i\}_{i=1}^k) \geq H((U_i)_{i=1}^k) - \log k!. \tag{4.10}$$

The lower bound is not tight due to the positive chance of ties among members of a multiset drawn from a discrete alphabet. If the chance of ties is small (if $|\mathcal{U}|$ is sufficiently large and $k$ is sufficiently small), the lower bound is a good approximation. As has been noted previously, one interpretation of sequence entropy reduction by order entropy to yield multiset entropy is of a multiset as an equivalence class of sequences. Since non-zero distortion only occurs when a sequence is represented by a sequence outside of its equivalence class, uncertainty within the class is allowable without incurring distortion.

Rather than bounding $H(\{U\})$, we can compute it exactly. If the multiset elements are drawn i.i.d., the distribution of types is given by a multinomial distribution derived from the distribution of the individual letters [206, Problem VI]. Suppose $u_i \in \mathcal{U}$ has probability $p_i$ and let $L_i$ be the number of occurrences of $u_i$ in $k$ independent trials. Then

$$\Pr[L_i = l_i] = \binom{k}{l_1, l_2, \ldots, l_{|\mathcal{U}|}} \prod_{i=1}^{|\mathcal{U}|} p_i^{l_i}, \qquad \text{for } i = 1, \ldots, |\mathcal{U}|, \tag{4.11}$$

for any type $(l_1, l_2, \ldots, l_{|\mathcal{U}|})$ of non-negative integers with sum $k$. Therefore,

$$H(\{U\}_{i=1}^k) = H(L_1, L_2, \ldots, L_{|\mathcal{U}|}; k), \tag{4.12}$$

where dependence on $k$ is explicitly notated. This was an exact calculation of $R_k(0)$ for a fixed $k$ and known distribution $p_i$. We are interested in $R(0)$ rather than $R_k(0)$ and would like to have a result that does not require the multiset generating process

to be i.i.d. As for other fidelity criteria, as in Section 2.4.1, finding an upper bound to the $R(0)$ point reduces to an enumeration problem, here the enumeration of types.

Denote the alphabet of distinct types as $\mathcal{L}(\mathcal{U}, k)$; its size may be computed and bounded through simple combinatorics [207]:

$$|\mathcal{L}(\mathcal{U}, k)| = \binom{k + |\mathcal{U}| - 1}{|\mathcal{U}| - 1} \leq (k+1)^{|\mathcal{U}|}. \tag{4.13}$$

Using (4.13) and noting that $|\mathcal{L}(\mathcal{U}, k)|/k$ is a monotonically decreasing bounded sequence of real numbers,

$$R(0) \leq \lim_{k\to\infty} \tfrac{1}{k} \log |\mathcal{L}(\mathcal{U}, n)| \leq \lim_{k\to\infty} \tfrac{1}{k} \log(k+1)^{|\mathcal{U}|} = 0. \tag{4.14}$$

**Theorem 13.** *$R(\Delta) = 0$ for any source with fidelity criterion $F_{\mathrm{val}_1}$.*

*Proof.* By the information inequality, $R(\Delta) \geq 0$. By (4.14), $R(0) \leq 0$, so $R(0) = 0$. Since $R(\Delta)$ is a non-increasing function, $R(\Delta) = 0$. □

Note that the theorem holds for any multiset, not just for multisets drawn i.i.d. In fact, if the multiset is drawn i.i.d., the bounding technique yields an upper bound that is quite loose. To achieve the bound with equality, each of the types would have to be equiprobable; however by the strong asymptotic equipartition property (AEP) [38], collectively, all non-strongly typical types will occur with probability as close to zero as desired.[2] The number of types in the strongly typical set is polynomial in $k$, so we cannot get a speed of convergence to $R_k(0)/k$ rate faster than $O(\log k/k)$. For clarity, we work through an example.

**Example 1.** *Consider a multiset of size $k$ whose elements are drawn i.i.d. from a Bernoulli distribution with parameter $p$. The $R_k(0)$ value is equal to the entropy of a binomial random variable with parameters $k$ and $p$. For large $k$, we can use the de Moivre approximation of the binomial, $\mathcal{N}(kp, kp(1-p))$ evaluated at the integers [206, pp. 243–259]. The entropy of this approximation is the same as the high rate*

[2] The weak AEP is not sufficient to make this claim.

*approximation of Gaussian entropy with interval size 1. In fact, a better asymptotic expression for this entropy is given in [208] as*

$$R_k(0) \sim \tfrac{1}{2}\log_2(2\pi ekp(1-p)) + \sum_{i\geq 1} a_i k^{-i}, \tag{4.15}$$

*for some known constants $a_i$. Thus the rate-distortion function is*

$$R(0) = \lim_{k\to\infty} \frac{R_k(0)}{k} = \lim_{k\to\infty} \frac{\log_2(2\pi ekp(1-p))}{2k} + \sum_{i\geq 1} a_i k^{-i-1} = 0. \tag{4.16}$$

*Evidently, the speed of convergence is $O(\log k/k)$, which is the same as the speed of convergence for the universally valid Theorem 13.*

Theorem 13 was proven for finite alphabets, however it can also be extended to countable alphabets using a different method of proof [209].

### ■ 4.4.2 Continuous Alphabets

Assume that the source alphabet consists of real scalars, with the usual ordering.[3] When source letters are drawn i.i.d. from a continuous alphabet, the associated multiset is almost surely a set. Rather than using types as natural representatives for sets, we use order statistics. Recalling that $U_{(r:k)}$ denotes the $r$th order statistic from a block of $k$, we define the word distortion measure to be

$$\rho_k(u_1^k, v_1^k) = \frac{1}{k}\sum_{i=1}^{k}(u_{(i:k)} - v_{(i:k)})^2, \tag{4.17}$$

and an associated fidelity criterion to be

$$F_{\text{val}_2} = \{\rho_k(u_1^k, v_1^k), k = 1, 2, \ldots\}. \tag{4.18}$$

Although not a single-letter fidelity criterion, it is single-letter mean square error on the block of order statistics. Just as we were able to show that $R(\Delta) = 0$ under the

[3]Recall that ordering of real vectors and other continuous sources without natural orders is problematic [205], as is defining multivariate quantile functions [210].

$F_{\text{val}_1}$ fidelity criterion, we would like to show that $R(\Delta) = 0$ under the $F_{\text{val}_2}$ fidelity criterion. We show exactly that result in this section, but restricting ourselves to i.i.d. generation and some mild conditions on the distribution from which the multiset elements are drawn.

The average incurred distortion as a function of rate, $\Delta(R)$ is the generalized inverse function of $R(\Delta)$ and is called the distortion rate function. For $F_{\text{val}_2}$, the incurred distortion for blocklength $k$ is

$$\Delta_k = \frac{1}{k} \sum_{i=1}^{k} \mathop{\mathrm{E}}_{p_{U,V}} \left[ \left( U_{(i:k)} - V_{(i:k)} \right)^2 \right]. \tag{4.19}$$

If we use no rate, then the best choice for the reconstruction letter is simply $v_{(i:k)} = \mathrm{E}_{p_U}[U_{(i:k)}]$ and the average incurred distortion reduces to

$$\Delta_k(R = 0) = \frac{1}{k} \sum_{i=1}^{k} \mathop{\mathrm{var}}_{p_U} \left[ U_{(i:k)} \right]. \tag{4.20}$$

Before proceeding with a general proof that $\Delta(R) = 0$ (and so $R(\Delta) = 0$), we give some examples.

**Example 2.** *Consider a set with elements drawn i.i.d. from the uniform distribution with support* $\left[-\sqrt{3}, \sqrt{3}\right]$. *Then the variance of the* $i$*th order statistic from a set of size* $k$ *is [211, Section IV.A]*

$$\mathop{\mathrm{var}}_{p_U} \left[ U_{(i:k)} \right] = \frac{12i\,(k - i + 1)}{(k+1)^2\,(k+2)}. \tag{4.21}$$

*The average variance is then*

$$\Delta_k(0) = \frac{1}{k} \sum_{i=1}^{k} \frac{12i\,(k - i + 1)}{(k+1)^2\,(k+2)} = \frac{2}{k+1}. \tag{4.22}$$

*The monotonically decreasing sequence of real numbers,* $\{\Delta_k(0)\}$*, is shown in Fig-*

*ure 4-1; the sequence clearly satisfies*

$$\lim_{k\to\infty} \Delta_k(0) = 0, \tag{4.23}$$

*so* $\Delta(0) = 0$.

**Example 3.** *Consider a set with elements drawn i.i.d. from the Gaussian distribution with mean zero and variance one. Then the variance of the ith order statistic from a set of size k can be found numerically [212] and used to compute the average variance. Figure 4-1 shows* $\Delta_k(0)$ *as a function of k. This bounded sequence of real numbers* $\{\Delta_k(0)\}$ *is monotonically decreasing and satisfies* $\lim_{k\to\infty} \Delta_k(0) = 0$ *[213], so* $\Delta(0) = 0$.

**Example 4.** *Consider a set with elements drawn i.i.d. from the exponential distribution with mean one. Then the variance of the ith order statistic from a set of size k is [204]*

$$\operatorname*{var}_{p_U}\left[U_{(i:k)}\right] = \sum_{m=k-i+1}^{k} m^{-2}. \tag{4.24}$$

*The average variance is then*

$$\Delta_k(0) = \frac{1}{k}\sum_{i=1}^{k}\sum_{m=k-i+1}^{k} m^{-2}. \tag{4.25}$$

*Again, this bounded, monotonically decreasing sequence of real numbers is shown in Figure 4-1 and has limit* 0 *for large k.*

The average variance for sets generated according to other distributions may be computed either directly or by noting that the average second moment of the ordered and unordered variates is the same, since it is simply a rearrangement. Then computation reduces to finding the average mean square of the order statistics. One particularly interesting way to do this is given in [213].

The general theorem on zero rate, zero distortion will be based on the quantile function of the i.i.d. process generating the set; this is the generalized inverse of the

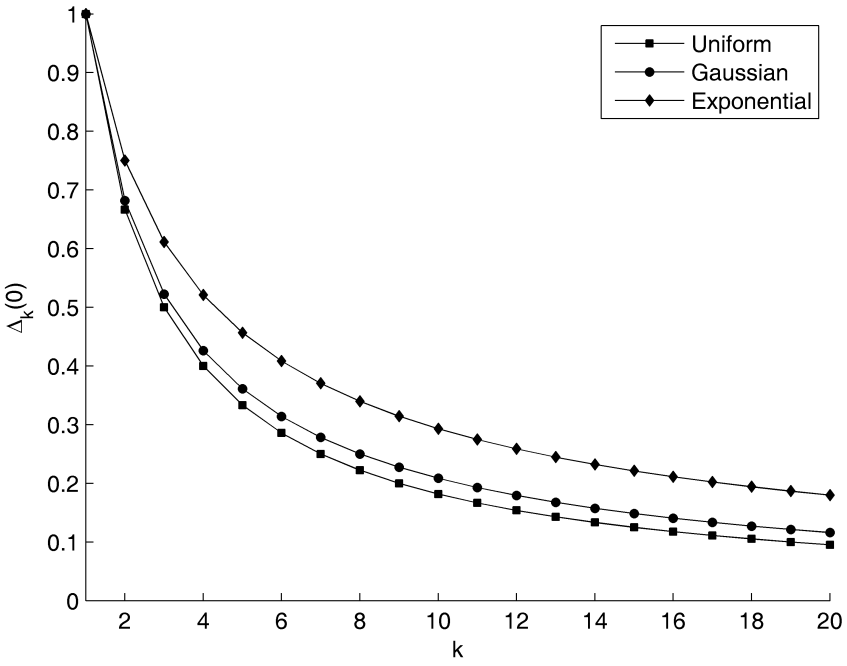


**Figure 4-1.** Distortion-rate function for some sources. $\Delta_k(0)$ for standard Gaussian shown with circles. $\Delta_k(0)$ for standard exponential shown with diamonds. $\Delta_k(0)$ for uniform shown with squares.

cumulative distribution function

$$Q(w) = F^{-1}(w) = \inf\{u : F_U(u) \geq w\}. \tag{4.26}$$

The empirical quantile function, defined in terms of order statistics is

$$Q_k(w) = U_{(\lfloor wk \rfloor + 1:k)} = F_k^{-1}(w), \tag{4.27}$$

where $F_k(\cdot)$ is the empirical distribution function. The quantile function $Q(\cdot)$ is continuous if and only if the distribution function has no flat portions in the interior, i.e. the density is strictly positive over its region of support except perhaps on isolated points. The main step of the proof will be a Glivenko-Cantelli like theorem for empirical quantile functions [214].

**Lemma 1.** *Let the sequence to be coded, $U_1, U_2, \ldots, U_k$, be generated in an i.i.d.*

*fashion according to $F_U(u)$ with associated quantile function $Q(w)$. Let $U_1$ satisfy*

$$\mathop{\mathrm{E}}_{p_U} |\min(U_1, 0)|^{1/\nu_1} < \infty \quad \textit{and} \quad \mathop{\mathrm{E}}_{p_U} (\max(U_1, 0))^{1/\nu_2} < \infty \tag{4.28}$$

*for some $\nu_1 > 0$ and $\nu_2 > 0$ and have continuous quantile function $Q(w)$. Then the sequence of distortion-rate values for the coding of size-$k$ sets drawn with the distribution of $U_1$ satisfy*

$$\lim_{k\to\infty} \Delta_k(R = 0) = 0. \tag{4.29}$$

*Proof.* For any nonnegative function $\omega$ defined on $(0, 1)$, define a weighted Kolmogorov-Smirnov like statistic

$$S_k(\omega) = \sup_{0<w<1} \omega(w) \left| Q_k(w) - Q(w) \right|. \tag{4.30}$$

For each $\nu_1 > 0$, $\nu_2 > 0$, and $w \in (0, 1)$, define the weight function

$$\omega_{\nu_1,\nu_2}(w) = w^{\nu_1}(1 - w)^{\nu_2}. \tag{4.31}$$

Assume that $Q$ is continuous, choose any $\nu_1 > 0$ and $\nu_2 > 0$, and define

$$\gamma = \limsup_{k\to\infty} S_k(\omega_{\nu_1,\nu_2}). \tag{4.32}$$

Then by a result of Mason [214], $\gamma = 0$ with probability 1 when (4.28) holds. Our assumptions on the generating process meet this condition, so $\gamma = 0$ with probability 1. This implies that

$$\limsup_{k\to\infty} |U_{(\lfloor wk \rfloor + 1)} - Q(w)| \leq 0 \text{ for all } w \in (0, 1) \text{ w.p.1}, \tag{4.33}$$

and since the absolute value is nonnegative, the inequality holds with equality. According to (4.33), for sufficiently large $k$, each order statistic takes a fixed value with probability 1. The bounded moment condition on the generating process, (4.28), implies a bounded moment condition on the order statistics. Almost sure convergence

to a fixed quantity, together with the bounded moment condition on the events of probability zero imply convergence in second moment of all order statistics. This convergence in second moment to a deterministic distribution implies that the variance of each order statistic is zero, and thus the average variance is zero. □

We have established that asymptotically in $k$, the point $(R = 0, \Delta = 0)$ is achievable, which leads to the following theorem.

**Theorem 14.** *Under fidelity criterion $F_{\text{val}_2}$, $R(\Delta) = 0$ for an i.i.d. source that meets the bounded moment condition (4.28) and has continuous quantile function.*

*Proof.* By the nonnegativity of the distortion function, $\Delta(R) \geq 0$. By Lemma 1, $\Delta(0) \leq 0$, so $\Delta(0) = 0$. Since $\Delta(R)$ is a non-increasing function, $\Delta(R) = 0$, and so $R(\Delta) = 0$. □

Due to the generality of the Glivenko-Cantelli like theorem that we used, the result will stand for a very large class of distortion measures. One only needs to ensure that the set of outcomes of probability zero are not problematic.

The $F_{\text{val}_2}$ problem is trivial in the same manner that the $F_{\text{val}_1}$ problem is trivial. No rate is needed to achieve zero distortion, i.e. no information need be stored on average to be able to retrieve information within the required fidelity. An interpretation of this phenomenon is that since the generating distribution is known at the destination, it is as if the source can just be simulated for the destination without having to store anything; this recreated source multiset will be the same as the original source multiset. Given these reductions to triviality for fidelity criteria defined on growing multisets, one might wonder what happens when the multiset sizes are fixed. This is the topic of the next three sections.

## ■ 4.5 Low-Rate Low-Dimension Quantization for Fixed Size Multisets

For fidelity criterion $F_{\text{val}_1}$, as $k$ increases, larger and larger multisets are used to compute the distortion. This trivializes the problem, in the sense that no rate is needed to achieve zero distortion. An alternate view of order irrelevance is to fix

the size of the multiset, at say $K$, and take larger and larger numbers of multisets. In effect, we now have a block source with a single-letter fidelity criterion. Using the previously defined word distortion on blocks of length $K$, (4.17), a new fidelity criterion is

$$F_{\text{val}_3} = \left\{ \frac{K}{k} \sum_{i=1}^{k/K} \rho_K(u_{iK-K+1}^{iK}, v_{iK-K+1}^{iK}), k = K, 2K, \ldots \right\}. \tag{4.34}$$

This is average mean square error on the block of order statistics. Notice that the fidelity criterion is defined only for words that have lengths that are multiples of the block size $K$.

For low rates and coding one set at a time, we can find optimal mean square error quantizers through the Lloyd-Max optimization procedure [215]. The quantizers generated in this way are easy to implement for practical source coding, and also provide an upper bound on the rate-distortion function. Designing the quantizers requires knowledge of the distributions of order statistics, which can be derived from the distribution of the unordered variates [204]. For $U_1, \ldots, U_K$ that are drawn i.i.d. according to the cumulative distribution function $F_U(u)$, the marginal cumulative distribution function of $U_{(r:K)}$ is given in closed form by

$$F_{(r:K)}(u) = \sum_{i=r}^{K} \binom{K}{i} F_U^i(u) \left[1 - F_U(u)\right]^{K-i} = I_{F_U(u)}(r, K-r+1), \tag{4.35}$$

where $I_p(a, b)$ is the incomplete beta function. Subject to the existence of the density $f_U(u)$ of the original generating process, the marginal density of $U_{(r:K)}$ is

$$f_{(r:K)}(u) = \frac{1}{B(r, K-r+1)} \left[1 - F_U(u)\right]^{K-r} F_U^{r-1}(u) f_U(u), \tag{4.36}$$

where $B(a, b)$ is the beta function. The joint density of all $K$ order statistics is

$$f_{(1:K)\cdots(K:K)}(u_1, \ldots, u_K) = \begin{cases} K! \prod_{i=1}^{K} f_U(u_i), & u_1^K \in \mathfrak{R}; \\ 0, & \text{else.} \end{cases} \tag{4.37}$$

The region of support, $\mathfrak{R} = \{u_1^K : u_1 \leq \cdots \leq u_K\}$, is a convex cone that occupies $(1/K!)$th of $\mathbb{R}^K$. The order statistics also have the Markov property [204] with transition probability

$$f_{U_{(r+1:K)}|U_{(r:K)}=x}(y) = (K-r)\left[\frac{1-F_U(y)}{1-F_U(x)}\right]^{K-r-1}\frac{f_U(y)}{1-F_U(x)}, \quad \text{for } y > x. \tag{4.38}$$

In the usual setup, the sorting filter is applied first and then further source coding is performed. Since sorting quantized numbers is easier than sorting real-valued numbers, we would prefer to be able to interchange the operations. Based on the form of the joint distribution of order statistics, (4.37), we can formulate a statement about when sorting and quantization can be interchanged without loss of optimality. If the order statistics are to be quantized individually using scalar quantization, then interchange without loss can be made in all cases [216]. Scalar quantization, however, does not take advantage of the Markovian dependence among elements to be coded. We consider coding the entire set together, referring to $K$ as the dimension of the order statistic vector quantizer.

If the representation points for an MSE-optimal ($R$ rate, $K$ dimension) order statistic quantizer are the intersection of the representation points for an MSE-optimal $(R + \log K!, K)$ quantizer for the unordered variates and $\mathfrak{R}$, then we can interchange sorting and quantization without loss of optimality. This condition can be interpreted as a requirement of permutation polyhedral symmetry on the quantizer of the unordered variates. This form of symmetry requires that there are corresponding representation points of the unordered variate quantizer in each of the $K!$ convex cones that partition $\mathbb{R}^K$ on the basis of permutation. The polyhedron with vertices that are corresponding points in each of the $K!$ convex cones is a permutation polyhedron. In fact, the distortion performance of the MSE-optimal $(R, K)$ order statistic quantizer is equal to the distortion performance of the best $(R + \log K!, K)$ unordered quantizer constrained to have the required permutation symmetry. An example where the symmetry condition is met is for the standard bivariate Gaussian distribution shown in Figure 4-2.

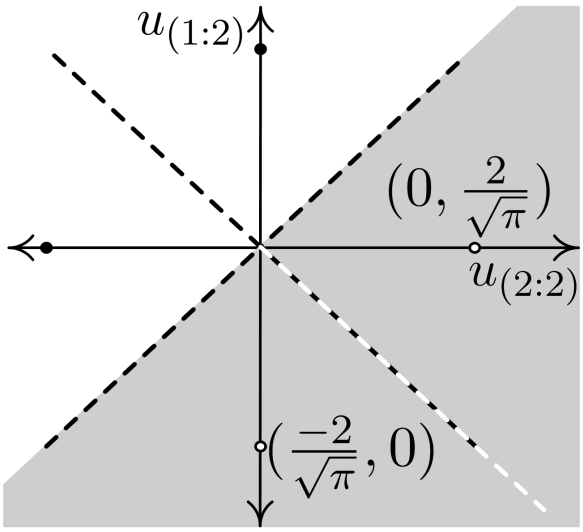


**Figure 4-2.** Quantization for bivariate standard Gaussian order statistics. Optimal one-bit quantizer (white) achieves $(R = 1, \Delta = \frac{2\pi-4}{\pi})$. Optimal two-bit quantizer (black) for unordered variates achieves $(R = 2, \Delta = \frac{2\pi-4}{\pi})$. Since representation points for order statistic quantizer are the intersection of the cone (shaded) and the representation points for the unordered quantizer, the distortion performance is the same.

## ■ 4.6 High-Rate Quantization Theory for Fixed Size Multisets

Based on the basic distributional properties of order statistics, (4.36)–(4.38), the differential entropies of order statistics can be derived. The individual marginal differential entropies are

$$h(U_{(r:K)}) = \int f_{(r:K)}(u) \log f_{(r:K)}(u) du, \tag{4.39}$$

where no particular simplification is possible unless the generating process is specified. The average marginal differential entropy, however, can be expressed in terms of the differential entropy of the generating process and a constant that depends only on $K$ [217]:

$$\bar{h}(U_{(1:K)}, \ldots, U_{(K:K)}) = \frac{1}{K} \sum_{i=1}^{K} h(U_{(i:K)}) = h(U_1) - \log K - \frac{1}{K} \sum_{i=1}^{K} \log \binom{K-1}{i-1} + \frac{K-1}{2}. \tag{4.40}$$

The subtractive constant is positive and increasing in $K$, and not dependent on the generating distribution. The individual conditional differential entropies, as derived

in Appendix 4.A, are

$$h(U_{(r+1:K)}|U_{(r:K)}) = -\log(K-r) - N_h(K) + N_h(K-r) + 1 - \frac{1}{K-r} \qquad (4.41)$$
$$- \frac{K!}{\Gamma(K-r)\Gamma(r)} \int_{-\infty}^{\infty} \int_{x}^{\infty} f_U(y) \log(y) \left[1 - F_U(y)\right]^{K-r-1} dy F_U^{r-1}(x) f_U(x) dx,$$

where $\Gamma(\cdot)$ is the gamma function and $N_h(k) = \sum_{m=1}^{k} 1/m$ is the harmonic number. As in the individual marginal case, further simplification of this expression requires the generating distribution to be specified. Again, as in the marginal case, the total conditional differential entropy can be expressed in terms of the generating process differential entropy and a constant that depends only on $K$. Due to Markovianity, the sum of the individual conditional differential entropies is in fact the joint differential entropy:

$$h(U_{(1:K)}, \ldots, U_{(K:K)}) = h(U_{(1:K)}) + \sum_{i=1}^{K-1} h(U_{(i+1:K)}|U_{(i:K)}) = Kh(U_1) - \log K!. \quad (4.42)$$

Notice that an analogous statement (4.10) was a lower bound in the discrete alphabet case; equality holds in the continuous case since there are no ties.

High-rate quantization results follow easily from the differential entropy calculations. High-rate quantization theory is an alternative to rate distortion theory and provides a different way of characterizing the tradeoff between rate and distortion in the rate matching interpretation of optimal information storage. The rate and distortion performance of a particular form of asymptotic, deterministic source codes is approximated in the high-rate (low-distortion) limit.

To develop results, we introduce four quantization schemes in turn, measuring performance under fidelity criterion $F_{\text{val}_3}$. In particular, we sequentially introduce a *shape advantage*, a *memory advantage*, and a *space-filling advantage* as in [218].[4] As a baseline, take the naïve scheme of directly uniform scalar quantizing the randomly ordered sequence with quantization step size $\epsilon$. The average rate and distortion per

[4]Note that vector quantizer advantages are discussed in terms of distortion for fixed rate in [218], but we discuss some of these advantages in terms of rate for fixed distortion.

source symbol of the naïve scheme are $R_1 = h(U_1) - \log \epsilon$, and $\Delta_1 = \epsilon^2/12$. Now uniform scalar quantize the deterministically ordered sequence (the order statistics). This changes the shape of the marginal distributions that we are quantizing, and thus we get a shape advantage. The average rate per source symbol for this scheme is

$$R_2 = \bar{h}(U_{(1:K)}, \ldots, U_{(K:K)}) - \log \epsilon = R_1 - \log K - \frac{1}{K} \sum_{i=1}^{K} \log \binom{K-1}{i-1} + \frac{K-1}{2}. \quad (4.43)$$

The distortion is the same as the naïve scheme, $\Delta_2 = \Delta_1$. As a third scheme, scalar quantize the order statistics sequentially, using the previous order statistic as a form of side information. We assume that the previous order statistics are known exactly to both the storer and retriever. Since the order statistics form a Markov chain, this single-letter sequential transmission exploits all available memory advantage. The rate for this scheme is

$$R_3 = \tfrac{1}{K} h(U_{(1:K)}, \ldots, U_{(K:K)}) - \log \epsilon = R_1 - \tfrac{1}{K} \log K!. \quad (4.44)$$

Again, $\Delta_3 = \Delta_1$. Finally, the fourth scheme would vector quantize the entire sequence of order statistics collectively. Since we have exploited all shape and memory advantages, the only thing we can gain is space-filling gain. The rate is the same as the third scheme, $R_4 = R_3$, however the distortion is less. This distortion reduction is a function of $K$, is related to the best packing of polytopes, and is not known in closed form for most values of $K$; see Table I of [218] and more recent work on packings. We denote the distortion as $\Delta_4 = \Delta_1 / G(K)$, where $G(K)$ is a function greater than unity. The performance improvements of these schemes are summarized in Table 4.6. Notice that all values in Table 4.6 depend only on the set length $K$ and not on the distribution that is used to generate the sets.

We have introduced several quantization schemes and calculated their performance in the high-rate limit. It was seen that taking the fidelity criterion into account when designing the source coder resulted in rate savings that did not depend on the distribution according to which the source sequence was generated. These rate

| | Rate Reduction ($-$) | Distortion Reduction ($\times$) |
|---|---|---|
| Scheme 1 | 0 | 1 |
| Scheme 2 (s) | $\log K + \frac{1}{K}\sum_{i=1}^{K} \log \binom{K-1}{i-1} - \frac{K-1}{2}$ | 1 |
| Scheme 3 (s,m) | $\log K!/K$ | 1 |
| Scheme 4 (s,m,f) | $\log K!/K$ | $1/G(K)$ |

**Table 4.1.** Comparison between the Scheme 1 and several other quantization schemes. The symbols (s),(m), and (f) denote shape, memory, and space-filling advantages.

savings can be quite significant for large blocklengths $K$.

## ■ 4.7 Rate Distortion Theory for Fixed Size Multisets

Let us return to rate-distortion problems as in Section 4.4, but now under fixed size multiset fidelity criteria like $F_{\mathrm{val}_3}$. First consider discrete alphabets, before returning to the continuous alphabet, $F_{\mathrm{val}_3}$ problem.

### ■ 4.7.1 Discrete Alphabets

For discrete alphabets, we define a fidelity criterion for fixed size multisets in the same manner as for $F_{\mathrm{val}_3}$.

$$F_{\mathrm{val}_4} = \left\{ \frac{K}{k} \sum_{i=1}^{k/K} d_K(u_{iK-K+1}^{iK}, v_{iK-K+1}^{iK}), k = K, 2K, \ldots \right\}. \tag{4.45}$$

The word distortion measure used to define the fidelity criterion is (4.8). The $F_{\mathrm{val}_4}$ notion of fidelity casts the problem into a frequency of error framework on the types. Assuming that the multisets to be coded are independent and identically distributed, this is simply an i.i.d. discrete source with error frequency distortion, so the reverse waterfilling solution of Erokhin [219] applies. The rate-distortion function is given parametrically as

$$\begin{aligned}
\Delta_\theta &= 1 - S_\theta + \theta(N_\theta - 1) \\
R_\theta &= -\sum_{l:p(l)>\theta} p(l) \log p(l) + (1 - \Delta_\theta) \log(1 - \Delta_\theta) + (N_\theta - 1)\theta \log \theta,
\end{aligned}$$

where $N_\theta$ is the number of types whose probability is greater than $\theta$ and $S_\theta$ is the sum of the probabilities of these $N_\theta$ types. The parameter $\theta$ goes from 0 to $p(l^\ddagger)$ as $\Delta$ goes from 0 to $\Delta_{max} = 1 - p(l^\dagger)$; the most probable type is denoted $l^\dagger$ and the second most probable type is denoted $l^\ddagger$. If the letters within the multisets are also i.i.d., the probability values needed for the reverse waterfilling characterization are computed using the multinomial distribution.

Only the most probable source types are used in the representation alphabet. It is known that the probability of type class $l$ drawn i.i.d. from the parent $p_U$ is bounded as follows [207]:

$$\tfrac{1}{|\mathcal{L}|} 2^{-kD(p_l \| p_U)} \leq \Pr[l] \leq 2^{-kD(p_l \| p_U)}, \tag{4.46}$$

where $p_l$ is a probability measure derived by normalizing the type $l$. The multiset types used in the representation alphabet are given by the type classes in the typical set

$$T_{p_U}^{\epsilon(\theta)} = \{l : D(p_l \| p_U) \leq \epsilon(\theta)\}. \tag{4.47}$$

Since multiset sources are successively refinable under error frequency distortion [220], scalable coding would involve adding types into the representation alphabet.

In addition to $F_{\mathrm{val}_4}$, we can define other fidelity criteria that reduce the multiset rate-distortion problem to well-known discrete memoryless source rate-distortion problems. As a simple example, consider multisets of length $K = 2$ and consisting of i.i.d. equiprobable binary elements. Then there are three letters in the alphabet of types: $\{0, 0\}$, $\{0, 1\}$, and $\{1, 1\}$, which can be represented by their Hamming weights, $\{0, 1, 2\}$. The probabilities of these three letters are $\{1/4, 1/2, 1/4\}$. Define the word distortion function

$$\delta(u_1^2, v_1^2) = \left| w_H(u_1^2) - w_H(v_1^2) \right|. \tag{4.48}$$

The fidelity criterion is

$$F_{\mathrm{val}_5} = \left\{ \frac{2}{k} \sum_{i=1}^{k/2} \delta(u_{2i-1}^{2i}, v_{2i-1}^{2i}), k = 2, 4, \ldots \right\}. \tag{4.49}$$

This is a single-letter fidelity criterion on the Hamming weights and is in fact the well-studied problem known as the Gerrish problem [77, Problem 2.8]. One can easily generate equivalences to other known problems as well.

### 4.7.2 Continuous Alphabets

It is quite difficult to obtain the full rate-distortion function for the $F_{\text{val}_3}$ fidelity criterion, however upper and lower bounds may be quite close to each other for particular source distributions. As an example, consider the rate-distortion function for the independent bivariate standard Gaussian distribution that was considered in Figure 4-2. The rate-distortion function under $F_{\text{val}_3}$ is equivalent to the rate-distortion function for the order statistics under the mean square error fidelity criterion. The Shannon lower bound is simply

$$R_{SLB}(\Delta) = \log(1/\Delta), \tag{4.50}$$

the Gaussian rate-distortion function under the mean square error fidelity criterion, reduced by $\log K!$ bits (one bit). Note that since the order statistic source cannot be written as the sum of two independent processes, one of which has the properties of a Gaussian with variance $\Delta$,[5] the Shannon lower bound is loose everywhere [221], though it becomes asymptotically tight in the high-rate limit.

The covariance matrix of the Gaussian order statistics can be computed in closed form as

$$\Lambda = \begin{bmatrix} 1 - 1/\pi & 1/\pi \\ 1/\pi & 1 - 1/\pi \end{bmatrix}, \tag{4.51}$$

with eigenvalues 1 and $1 - 2/\pi$. Reverse waterfilling yields the Shannon upper bound

$$R_{SUB}(\Delta) = \begin{cases} \frac{1}{2}\log\left(\frac{2-4/\pi}{\Delta}\right) + \frac{1}{2}\log\left(\frac{2}{\Delta}\right), & 0 \le \Delta \le 2 - 4/\pi \\ \frac{1}{2}\log\left(\frac{1}{\Delta - 1 + 2/\pi}\right), & 2 - 4/\pi \le \Delta \le 2 - 2/\pi \\ 0, & \Delta \ge 2 - 2/\pi. \end{cases} \tag{4.52}$$

[5] Even though $U_{(1:2)} = \frac{1}{2}(U_1 + U_2) - \frac{1}{2}|U_1 - U_2|$ and $U_{(2:2)} = \frac{1}{2}(U_1 + U_2) + \frac{1}{2}|U_1 - U_2|$, and the first terms are Gaussian, the troublesome part is the independence.

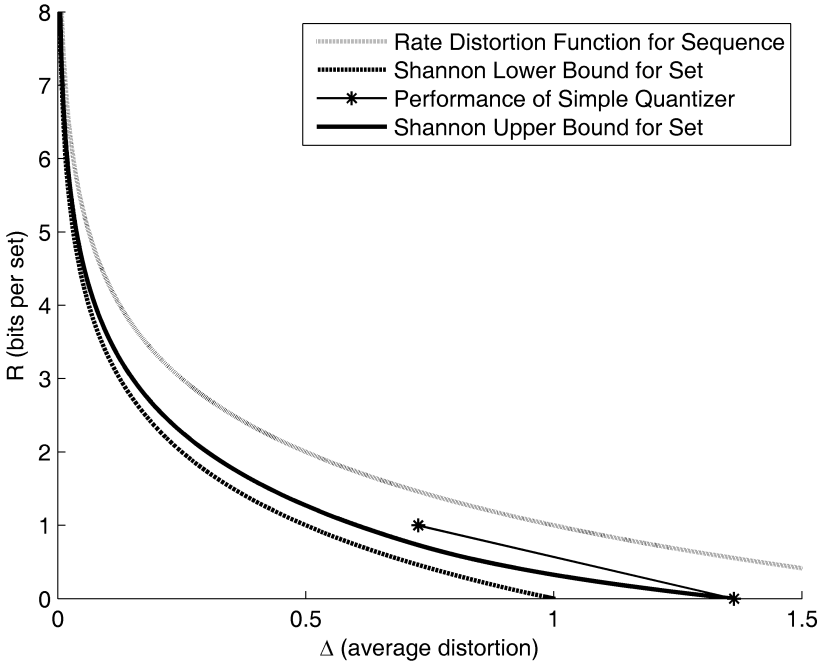


**Figure 4-3.** Shannon upper and lower bounds for the Gaussian order statistic rate-distortion function. The point achievable by single set code of Figure 4-2 is also shown connected to the zero rate point, which is known to be tight. Note that rate is not normalized per source letter.

This bound is tight at the point achieved by zero rate. Since the Gaussian order statistics for $K = 2$ have small non-Gaussianity, the Shannon lower bound and the Shannon upper bound are close to each other, as shown in Figure 4-3. For moderately small distortion values, we can estimate the rate-distortion function quite well.

## ■ 4.8 Closing Remarks

The goal of this chapter has been to present some rate matching implications of source semantics where order of elements is irrelevant. Such semantics arise in numerous information storage applications, as well as numerous applications in information transmission [222]. Separation of order and value and disregard for order is a coding strategy that is similar to the separation of phase and magnitude for source coding [223]. The nonlinear sorting filter that dissipates irrelevant entropy may allow a great savings in the rate required for storage. Overall, in the end-to-end information storage problem, one can say that correct recognition and discrimination of the invariant

properties that need to be stored with high fidelity and those properties that need not be stored at all can significantly reduce the rate, and thereby greatly reduce the resources that are needed for storage.

## Appendix

### 4.A Conditional Differential Entropy

The conditional differential entropy of the order statistics may be derived from the joint density and conditional density of the order statistics. For conciseness, denote the $r$th order statistic $U_{(r:n)}$ as $U_{(r)}$, and the cumulative distribution function and density function of the i.i.d. generating process, $F_U(\cdot)$ and $f_U(\cdot)$, as $F(\cdot)$ and $f(\cdot)$ respectively. The term *distribution-free* is used to describe quantities that do not depend on $F$ or $f$, only on $n$ and $r$. It is known [204] that the joint density between two consecutive order statistics is

$$f_{U_{(r)},U_{(r+1)}}(x,y) = \frac{n!}{(r-1)!(n-r-1)!} F^{r-1}(x) f(x) f(y) \left[1-F(y)\right]^{n-r-1}, \; y > x \tag{4.53}$$

and the conditional density is

$$f_{U_{(r+1)}|U_{(r)}}(y|x) = \frac{(n-r)f(y)}{1-F(x)} \left[\frac{1-F(y)}{1-F(x)}\right]^{n-r-1}, \; y > x. \tag{4.54}$$

Thus the conditional differential entropy is, by definition,

$$h(U_{(r+1)}|U_{(r)}) = -\int_{-\infty}^{\infty}\int_{x}^{\infty} f_{U_{(r)},U_{(r+1)}}(x,y) \log f_{U_{(r+1)}|U_{(r)}}(y|x)\, dy\, dx. \tag{4.55}$$

Substituting in and splitting the logarithm, we get

$$\begin{aligned} h = -\int_{-\infty}^{\infty}\int_{x}^{\infty} f_{U_{(r)},U_{(r+1)}} &\left\{\log(n-r) + (r-n)\log\left[1-F(x)\right]\right. \\ &\left. + \log f(y) + (n-r-1)\log\left[1-F(y)\right]\right\} dy\, dx \\ \triangleq -t_1 - t_2 - t_3 - t_4, \end{aligned} \tag{4.56}$$

where the $t_i$ are the four additive terms into which the integral may be split. Using the fact that

$$\int_x^\infty f(y)\left[1-F(y)\right]^{n-r-1}dy = \frac{\left[1-F(x)\right]^{n-r}}{(n-r)} \tag{4.57}$$

and that

$$\int_{-\infty}^\infty \left[1-F(x)\right]^{n-r}F^{r-1}(x)f(x)dx = B(r,n-r+1), \tag{4.58}$$

defined in terms of the beta function $B(\cdot,\cdot)$, which is a distribution-free quantity, we find that $t_1$ is also distribution-free and simply

$$t_1 = \log(n-r). \tag{4.59}$$

Using (4.57) and the additional fact that

$$\int_{-\infty}^\infty \left[1-F(x)\right]^{n-r}F^{r-1}(x)f(x)\log\left[1-F(x)\right]dx = \frac{-(n-r)!\Gamma(r)\left[N_h(n)-N_h(n-r)\right]}{n!}, \tag{4.60}$$

where $\Gamma(\cdot)$ is the Gamma function and $N_h(s)=\sum_{m=1}^s 1/m$ is the harmonic number, we get that

$$t_2 = (n-r)\left[N_h(n)-N_h(n-r)\right]. \tag{4.61}$$

This is also a distribution-free quantity. Note that the harmonic number can also be expressed as

$$N_h(s) = \gamma + \psi_0(s+1), \tag{4.62}$$

where $\gamma$ is the Euler-Mascheroni constant and $\psi_0(\cdot)$ is the digamma function.

We use the fact that

$$\int_x^\infty f(y)\left[1-F(y)\right]^{n-r-1}\log\left[1-F(y)\right]dy = -\frac{\left[1-F(x)\right]^{n-r}\left\{1+(r-n)\log\left[1-F(x)\right]\right\}}{(n-r)^2} \tag{4.63}$$

along with (4.58) and (4.60) to get that

$$t_4 = -1 + \frac{1}{n-r} - (n-r-1)\left[N_h(n) - N_h(n-r)\right]. \tag{4.64}$$

Again, this is distribution-free. Unfortunately, $t_3$ is not distribution-free, nor does it seem to be expressible in closed form and is

$$t_3 = \frac{n!}{\Gamma(n-r)\Gamma(r)} \int_{-\infty}^{\infty} \int_{x}^{\infty} f(y)\log(y)\left[1-F(y)\right]^{n-r-1} dy F^{r-1}(x) f(x) dx. \tag{4.65}$$

So combining the additive terms,

$$\begin{aligned} h = &-\log(n-r) - (n-r)\left[N_h(n) - N_h(n-r)\right] + (n-r-1)\left[N_h(n) - N_h(n-r)\right] \\ &+ 1 - \frac{1}{n-r} - \frac{n!}{\Gamma(n-r)\Gamma(r)} \int_{-\infty}^{\infty} \int_{x}^{\infty} f(y)\log(y)\left[1-F(y)\right]^{n-r-1} dy F^{r-1}(x) f(x) dx, \end{aligned} \tag{4.66}$$

which simplifies to

$$\begin{aligned} h(U_{(r+1)}|U_{(r)}) = &-\log(n-r) - N_h(n) + N_h(n-r) + 1 - \frac{1}{n-r} \\ &- \frac{n!}{\Gamma(n-r)\Gamma(r)} \int_{-\infty}^{\infty} \int_{x}^{\infty} f(y)\log(y)\left[1-F(y)\right]^{n-r-1} dy F^{r-1}(x) f(x) dx. \end{aligned} \tag{4.67}$$

# Chapter 5

# Conclusion

In this ultimate chapter, we briefly recapitulate the main results of the thesis, also interpreting some results in terms of a hierarchy of communication problems that has recently been formulated [40]. Although the attentive reader has surely envisaged areas where the work presented in the thesis may be extended, we also explicitly mention some of these areas of possible future work, mainly focusing on the issue of delay.

## 5.1 Recapitulation

The objective of information storage is to store and then retrieve information with high fidelity and low resource consumption. The notion of fidelity is determined by the task or set of tasks for which the stored information is to be used upon retrieval. The notion of resources is determined by the costs and constraints of the physical information storage medium that is to be used. The semantic and physical problem of information storage is changed into a mathematical one by appropriate mathematization of the fidelity and resource criteria. Several characterizations of optimal information storage systems, based on different system properties, can be stated, all essentially versions of the information storage theorem. Moreover, there are several ways of designing optimal information systems that have certain robustness properties superior to others. All of the characterizations are based on matching system parameters in some way. In some cases, the design of certain system components are determined by other fixed components. For example, the designs of written alphabets as stored for humans are thought to be matched to the fixed human visual system

retriever for information storage through writing on paper [224], which in turn is fixed for the design of the optical character recognition retriever [225].

Although there are several characterizations of information storage systems in terms of various system parameters, the fundamental performance parameters are the average incurred distortion, $\Delta$, and average incurred cost, $B$. Developing a hierarchy of information storage problems allows interpretation of which problems are harder than others. Since $B$ and $\Delta$ are fundamental, these should be used to form the hierarchy. One of the earliest investigations of ordering the difficulty of storage problems used discrete memoryless channels themselves as the system resource over which to form a partial order [226]. Similarly, in [40], the channel itself is used to define a partial order on storage problems, but the fact that more fundamental resource parameters could be used is noted. Within this Sahai-Mitter hierarchy, our work is concerned only with a single family of storage problems, the family of estimation problems with distortion constraints. Though, as will be mentioned in Section 5.2, perhaps other families of storage problems should also be considered. Parametric comparison across families seems arduous, but comparison within a single family of problems can be parameterized by performance criteria such as $\Delta$ and $B$. Assuming fixed resource and fidelity criteria, the parameter $\Delta$ or $B$ fully characterizes the difficulty of the storage problem within the family of estimation problems with distortion constraints. A storage problem is said to be harder than another storage problem if the minimum cost needed to solve (in the sense of achieving $\Delta$) the first problem is more than that for the second problem. Since the cost-distortion curve, as defined by the intersection of (2.9) and (2.10), is a monotonic function, we can equivalently say that a storage problem with a smaller $\Delta$ is intrinsically harder than a storage problem with a larger $\Delta$. This gives a partial order on the difficulty of storage problems, parameterized by $\Delta$. Changing the resource criterion, e.g. changing $V_0$ in Chapter 3, or changing the fidelity criterion, as in Chapter 4, may change the absolute hardness of problems in an affine scaling sense, but does not change the partial ordering of problems. More drastic changes to the resource or fidelity criteria can change the partial ordering; in fact the partial ordering can be reversed with suitable modification to

the cost function or the distortion function. The hierarchy casts system resources $B$ as the central determinant of the hardness of problem that a particular storage system can solve. Although not central to our development, the hierarchy provides an interpretation of how changes in the problem statement make things easier or harder.

Having defined and characterized optimal information storage in Chapter 2, we used the characterization to investigate scientific questions on the structure of the mammalian brain. Our procedure can be understood in terms of an analogy with radio communication. We used optimality theory to say what properties an optimal system should have, extended an antenna out into the world to observe what the signaling scheme used in practice was, and then compared. Comparison of the predictions made by our unified theory and experimental results showed that what is observed and what is predicted are qualitatively and in some cases quantitatively the same, though further experimental results would be needed to further confirm or falsify the predictions. Chapter 3 was mainly concerned with the synaptic memory channel, whereas Chapter 4 turned to the question of how to represent memories in a useful way. In a sense, we addressed the question of what we really want to store, not what we think we want to store. The results of the chapter demonstrated that if the source has nonsequential semantics, what we really need to store is much less than what we naïvely might think we need to store: the problem is easier in the hierarchy.

## ■ 5.2 Future Directions

Further experimental verification of our theoretical predictions on the physical structure of the mammalian brain, as given in Chapter 3, stands out as worthy of further investigation. A full input-output characterization of synapses would help refine the system model and the predictions made by our theoretical framework. Determining the distribution of synaptic volume would require segmentation and volume calculation of synapses from electron micrographs and is a non-trivial machine vision problem. Determining the actual relationship between synaptic weight and synaptic volume would require a joint electrophysiology and electron microscopy experiment.

Since the experimental data that we have used is from rodents of the family Muridae, another experimental verification to pursue is cross-species validation. It is known that primates and particularly humans have more voluminous neocortex than other mammals [227]; a question that might be posed is whether this implies that humans have more bandwidth and thus have more total storage capacity, or is the volume taken up by something else.

Our work in Chapter 4 can also be extended in numerous ways. An assumption in basic information theory is that the storer and the retriever share knowledge of a code, which is essentially the same as sharing a probabilistic model for messages. This shared knowledge is what caused the triviality results to arise. One can ask what happens when neither the storer nor the retriever know the model of the source and it is necessary to use universal source coding schemes that learn the model of the source as they go on. It is clear that problems will not trivialize as easily. We have no results for lossy coding of sets of real-valued vector sources, as might be the model for compression of certain numerical databases, providing another topic for further investigation. Finally, investigation of other semantically significant yet mathematically tractable fidelity criteria calls out for attention.

Shannon's puzzling epigraph, "we may have knowledge of the past but cannot control it; we may control the future but have no knowledge of it" [37] comments on the separation between past and future, but does not explicitly mention that with an information storage system, knowledge of the past can actually help control the future. In fact, whether the dynamical system to be controlled is a chemical plant, a universal source coder, or an animal, it seems that knowledge of the past is requisite for controlling the future. This thesis has not explicitly incorporated time delay into the optimality criteria, which were simply the distortion and the cost, $(\Delta, B)$. Delay, however, is certainly an important issue in the design of information systems and in many ways determines the utility of retrieved information. Although information can remain in the channel for an indefinite period of time, when the need arises for its retrieval, the retrieval should be fast, so the delay requirement is asymmetric.

With the pyramid of biological and technological storage systems now available

to people, as alluded to in the first paragraph of the thesis, it seems that delay is a main limiting factor. In terms of the Sahai-Mitter hierarchy, the joint biological and technological memory system can use far more resources than a biological system alone and can thus solve much harder problems. The question is now whether the correct information can be retrieved quickly. Pyramidal memory systems are commonly used in computer systems, with many levels of memory with varying speeds and sizes. The fastest memories are more expensive per bit than the slower memories and are thus made smaller [228]. Since each type of memory storage medium has a different capacity per unit cost as well as delay property, it would seem that some optimization of size and distance could be made.

Related to the issue of delay is the ability to search for useful stored information, which is perhaps the most significant problem tackled in computer science research on databases and information storage systems. In computer systems, concepts of temporal and spatial locality, as well as prefetching are used to try to predict what stored information will become useful in the near future, thus hiding some of the latency associated with searching [228]. Since these are very simple methods of predicting the need for particular information in some upcoming time, better prediction would allow even more of the latency concerns to be hidden. It is believed that the neural systems of animals are quite good at predicting what information will be needed, so the search is not an issue, only propagation speed is.

In Chapter 3, we had incorporated the time delay constraint in an ad hoc manner, using previous results on signal propagation speed [165, 166] to bound the minimal size of axons and dendrites. A better theoretical approach would involve optimizing delay, distortion, and cost simultaneously. In the Sahai-Mitter hierarchy, by using a third performance measure, we would leave the family of estimation problems with distortion constraints and enter a family of delay-limited estimation problems with distortion constraints. Since there would be an additional constraint, problems in this family would be harder than or equal to problems in the previous family. If uncoded storage, of the type discussed in Section 2.6, could be used in all cases, then the delay requirement would not be of much concern since the receiving process would involve

no delay; only signal propagation speed would be of concern. It is not clear whether there would be a simple parameterization of the difficulty of problems, nor whether there would be simple characterizations of optimal storage, but it is definitely worthy of further investigation.

Optimal information storage and retrieval systems preserve information from the present for use in the future. By consuming more resources, better performance is possible, but by correct recognition of what needs to be preserved, less resources may allow equivalent performance. The importance of information preservation can not be emphasized enough; as people say, those who do not learn from history are doomed to repeat it.

# Bibliography


[1] L. Brillouin, *Science and Information Theory*, 2nd ed. New York: Academic Press, 1962.

[2] J. Hawkins and S. Blakeslee, *On Intelligence*. New York: Times Books, 2004.

[3] J. A. V. Bates, "Significance of information theory to neurophysiology," in *London Symposium on Information Theory*, Sept. 1950, printed as IRE PGIT-1, Feb. 1953.

[4] G. A. Miller, "Human memory and the storage of information," *IRE Trans. Inform. Theory*, vol. IT-2, no. 3, pp. 129–137, Sept. 1956.

[5] F. Rieke, D. Warland, R. de Ruyter van Steveninck, and W. Bialek, *Spikes: Exploring the Neural Code*. Cambridge, MA: The MIT Press, 1997.

[6] A. Borst and F. E. Theunissen, "Information theory and neural coding," *Nat. Neurosci.*, vol. 2, no. 11, pp. 947–957, Nov. 1999.

[7] S. Verdú, "On channel capacity per unit cost," *IEEE Trans. Inform. Theory*, vol. 36, no. 5, pp. 1019–1030, Sept. 1990.

[8] ——, "Spectral efficiency in the wideband regime," *IEEE Trans. Inform. Theory*, vol. 48, no. 6, pp. 1319–1343, June 2002.

[9] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–656, July/Oct. 1948.

[10] S. R. Kulkarni, "Problems of computational and information complexity in machine vision and learning," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, May 1991.

[11] N. Wiener, *The Extrapolation, Interpolation and Smoothing of Stationary Time Series*. New York: John Wiley & Sons, Inc., 1949.

[12] J. A. O'Sullivan, R. E. Blahut, and D. L. Snyder, "Information-theoretic image formation," *IEEE Trans. Inform. Theory*, vol. 44, no. 6, pp. 2094–2123, Oct. 1998.

[13] P. M. Woodward, *Probability and Information Theory, With Applications to Radar*. New York: McGraw-Hill, 1953.

[14] B. Rimoldi, "Beyond the separation principle: A broader approach to source-channel coding," in *4th International ITG Conference on Source and Channel Coding*, Berlin, Jan. 2002.

[15] R. M. Gray, *Entropy and Information Theory*. New York: Springer-Verlag, 1990.

[16] D. J. Sakrison, "On the role of the observer and a distortion measure in image transmission," *IEEE Trans. Commun.*, vol. COM-25, no. 11, pp. 1251–1267, Nov. 1977.

[17] M. Gastpar, B. Rimoldi, and M. Vetterli, "To code, or not to code: Lossy source-channel communication revisited," *IEEE Trans. Inform. Theory*, vol. 49, no. 5, pp. 1147–1158, May 2003.

[18] M. Gastpar, "To code or not to code," Ph.D. dissertation, École Polytechnique Fédérale de Lausanne, Switzerland, Jan. 2003.

[19] L. R. Varshney, "Engineering theory and mathematics in the early development of information theory," in *2004 IEEE Conference on the History of Electronics*, Bletchley Park, England, June 2004.

[20] R. R. Kline, "What is information theory a theory of?: Boundary work among information theorists and information scientists in the United States and Britain during the Cold War," in *The History and Heritage of Scientific and Technological Information Systems*, W. B. Rayward and M. E. Bowden, Eds. Medford, NJ: American Society of Information Science and Technology and the Chemical Heritage Foundation, 2004, pp. 15–28.

[21] S. Verdú, "Fifty years of Shannon theory," *IEEE Trans. Inform. Theory*, vol. 44, no. 6, pp. 2057–2078, Oct. 1998.

[22] R. J. McEliece, *The Theory of Information and Coding: A Mathematical Framework for Communication*, ser. Encyclopedia of Mathematics and Its Applications. London: Addison-Wesley, 1977, vol. 3.

[23] H. S. Nussbaum, "Source coding and adaptive data compression for communication networks," Ph.D. dissertation, University of California, Los Angeles, Los Angeles, CA, Dec. 1976.

[24] R. L. Dobrushin, "Shannon's theorems for channels with synchronization errors," *Probl. Inf. Transm.*, vol. 3, no. 4, pp. 11–26, Oct.-Dec. 1967.

[25] S. Z. Stambler, "Memoryless channels with synchronization errors: The general case," *Probl. Inf. Transm.*, vol. 6, no. 3, pp. 223–229, July-Sept. 1970.

[26] R. Ahlswede and J. Wolfowitz, "Channels without synchronization," *Adv. Appl. Prob.*, vol. 3, no. 2, pp. 383–403, Autumn 1971.

[27] A. Sahai and S. K. Mitter, "Source coding and channel requirements for unstable processes," *IEEE Trans. Inform. Theory*, 2005, submitted.

[28] S. Vembu, S. Verdú, and Y. Steinberg, "The source-channel separation theorem revisited," *IEEE Trans. Inform. Theory*, vol. 41, no. 1, pp. 44–54, Jan. 1995.

[29] S. Verdú and T. S. Han, "A general formula for channel capacity," *IEEE Trans. Inform. Theory*, vol. 40, no. 4, pp. 1147–1157, July 1994.

[30] C. E. Shannon, "Communication in the presence of noise," *Proc. IRE*, vol. 37, no. 1, pp. 10–21, Jan. 1949.

[31] M. Unser, "Sampling—50 years after Shannon," *Proc. IEEE*, vol. 88, no. 4, pp. 569–587, Apr. 2000.

[32] M. Vetterli, P. Marziliano, and T. Blu, "Sampling signals with finite rate of innovation," *IEEE Trans. Signal Processing*, vol. 50, no. 6, pp. 1417–1428, June 2002.

[33] L. R. Varshney and S. D. Servetto, "A distributed transmitter for the sensor reachback problem based on radar signals," in *Advances in Pervasive Computing and Networking*, B. K. Szymanski and B. Yener, Eds. Boston: Kluwer Academic Publishers, 2005, pp. 225–245.

[34] J. Barros and S. D. Servetto, "Network information flow with correlated sources," *IEEE Trans. Inform. Theory*, vol. 52, no. 1, pp. 155–170, Jan. 2006.

[35] A. D. Murugan, P. K. Gopala, and H. El Gamal, "Correlated sources over wireless channels: Cooperative source-channel coding," *IEEE J. Select. Areas Commun.*, vol. 22, no. 6, pp. 988–998, Aug. 2004.

[36] B. Ananthasubramaniam and U. Madhow, "Virtual radar imaging for sensor networks," in *The 3rd International Symposium on Information Processing in Sensor Networks (IPSN'04)*, Berkeley, CA, Apr. 2004.

[37] C. E. Shannon, "Coding theorems for a discrete source with a fidelity criterion," in *IRE National Convention Record, Part 4*, 1959, pp. 142–163.

[38] R. W. Yeung, *A First Course in Information Theory.* New York: Kluwer Academic/Plenum Publishers, 2002.

[39] C. E. Shannon, "The zero error capacity of a noisy channel," *IRE Trans. Inform. Theory*, vol. IT-2, no. 3, pp. 8–19, Sept. 1956.

[40] A. Sahai and S. Mitter, "The necessity and sufficiency of anytime capacity for stabilization of a linear system over a noisy communication link Part I: scalar systems," *IEEE Trans. Inform. Theory*, 2005, submitted.

[41] T. S. Han, *Information-Spectrum Methods in Information Theory.* Berlin: Springer, 2003.

[42] R. E. Totty and G. C. Clark, "Reconstruction error in waveform transmission," *IEEE Trans. Inform. Theory*, vol. IT-13, no. 2, pp. 336–338, Apr. 1967.

[43] S. S. Pradhan, J. Chou, and K. Ramchandran, "Duality between source coding and channel coding and its extension to the side information case," *IEEE Trans. Inform. Theory*, vol. 49, no. 5, pp. 1181–1203, May 2003.

[44] V. Bush, *Operational Circuit Analysis.* New York: John Wiley & Sons, Inc., 1929.

[45] R. H. Cannon, Jr., *Dynamics of Physical Systems.* New York: McGraw-Hill Book Company, 1967.

[46] J. R. Pierce, *An Introduction to Information Theory: Symbols, Signals and Noise*, 2nd ed. Dover, 1980.

[47] L. Galvani, "De viribus electricitatis in motu musculari commentarius," *Bon. Sci. Art. Inst. Acad. Comm.*, vol. 7, pp. 363–418, 1791.

[48] G. Basalla, *The Evolution of Technology.* New York: Cambridge University Press, 1988.

[49] M. Piccolino, "Animal electricity and the birth of electrophysiology: The legacy of Luigi Galvani," *Brain Res. Bull.*, vol. 46, no. 5, pp. 381–407, July 1998.

[50] C. Matteucci, *Traité des Phénomènes Électro-physiologiques des Animaux.* Paris: Fortin, Masson, 1844.

[51] R. W. Lucky, "A bit is a bit is a bit?" *IEEE Spectr.*, p. 15, July 1994.

[52] A. Sahai, "Anytime information theory," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, Feb. 2001.

[53] P. Elias, A. Feinstein, and C. E. Shannon, "A note on the maximum flow through a network," *IRE Trans. Inform. Theory*, vol. IT-2, no. 4, pp. 117–119, Dec. 1956.

[54] A. R. Lehman and E. Lehman, "Complexity classification of network information flow problems," in *Proceedings of the Fifteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA'04)*, New Orleans, Jan. 2004, pp. 142–150.

[55] V. Kawadia and P. R. Kumar, "A cautionary perspective on cross-layer design," *IEEE Wireless Commun. Mag.*, vol. 12, no. 1, pp. 3–11, Feb. 2005.

[56] R. Ahlswede, N. Cai, S.-Y. R. Li, and R. W. Yeung, "Network information flow," *IEEE Trans. Inform. Theory*, vol. 46, no. 4, pp. 1204–1216, July 2000.

[57] L. R. Ford, Jr. and D. R. Fulkerson, "Maximal flow through a network," *Can. J. Math.*, vol. 8, pp. 399–404, 1956.

[58] L. Lovász and M. D. Plummer, *Matching Theory*. Amsterdam: Elsevier Science Publishers, 1986.

[59] S. S. Pradhan, S. Choi, and K. Ramchandran, "A graph-based framework for transmission of correlated sources over multiple access channels," *IEEE Trans. Inform. Theory*, 2006, submitted.

[60] T. Berger and W. C. Yu, "Rate-distortion theory for context-dependent fidelity criteria," *IEEE Trans. Inform. Theory*, vol. IT-18, no. 3, pp. 378–384, May 1972.

[61] L. J. Guibas and A. M. Odlyzko, "String overlaps, pattern matching, and nontransitive games," *J. Comb. Theory, Ser. A*, vol. 30, no. 2, pp. 183–208, Mar. 1981.

[62] B. Winterfjord, "Binary strings and substring avoidance," Master's thesis, Chalmers University of Technology and Göteborg University, May 2002.

[63] S. Kak, "The golden mean and the physics of aesthetics." [Online]. Available: http://arxiv.org/pdf/physics/0411195

[64] A. N. Singh, "On the use of series in Hindu mathematics," *Osiris*, vol. 1, pp. 606–628, Jan. 1936.

[65] T. M. Green, "Recurrent sequences and Pascal's triangle," *Math. Mag.*, vol. 41, no. 1, pp. 13–21, Jan. 1968.

[66] R. V. L. Hartley, "Transmission of information," *Bell Syst. Tech. J.*, vol. 7, pp. 535–563, July 1928.

[67] C. Pimentel and B. F. Uchôa-Filho, "A combinatorial approach to finding the capacity of the discrete noiseless channel," *IEEE Trans. Inform. Theory*, vol. 49, no. 8, pp. 2024–2028, Aug. 2003.

[68] K. A. S. Immink, P. H. Siegel, and J. K. Wolf, "Codes for digital recorders," *IEEE Trans. Inform. Theory*, vol. 44, no. 6, pp. 2260–2299, Oct. 1998.

[69] K. A. S. Immink, "Runlength-limited sequences," *Proc. IEEE*, vol. 78, no. 11, pp. 1745–1759, Nov. 1990.

[70] S. Shamai (Shitz) and S. Verdú, "The empirical distribution of good codes," *IEEE Trans. Inform. Theory*, vol. 43, no. 3, pp. 836–846, May 1997.

[71] T. Weissman and E. Ordentlich, "The empirical distribution of rate-constrained source codes," *IEEE Trans. Inform. Theory*, vol. 51, no. 11, pp. 3718–3733, Nov. 2005.

[72] J. L. Massey, "Joint source and channel coding," in *Communication Systems and Random Process Theory*, J. K. Skwirzynski, Ed. Alphen aan den Rijn, Holland: Sijthoff & Noordhoff, 1978, pp. 279–293.

[73] T. Berger, "Living information theory," in *2002 IEEE International Symposium on Information Theory*, Lausanne, Switzerland, July 2002, (Shannon Lecture).

[74] D. F. Eldridge, "A special application of information theory to recording systems," *IEEE Trans. Audio*, vol. 11, no. 1, pp. 3–6, Jan. 1963.

[75] N. Singla, J. A. O'Sullivan, R. S. Indeck, and Y. Wu, "Iterative decoding and equalization for 2-D recording channels," *IEEE Trans. Magn.*, vol. 38, no. 5, pp. 2328–2330, Sept. 2002.

[76] T. J. Goblick, Jr., "Theoretical limitations on the transmission of data from analog sources," *IEEE Trans. Inform. Theory*, vol. IT-11, no. 4, pp. 558–567, Oct. 1965.

[77] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Englewood Cliffs, NJ: Prentice-Hall, Inc., 1971.

[78] T. Berger and J. D. Gibson, "Lossy source coding," *IEEE Trans. Inform. Theory*, vol. 44, no. 6, pp. 2693–2723, Oct. 1998.

[79] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," in *Advances in Neural Information Processing Systems 18*, Y. Weiss, B. Schölkopf, and J. Platt, Eds. Cambridge, MA: MIT Press, 2006.

[80] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge: Cambridge University Press, 2004.

[81] V. Anantharam and S. Verdú, "Bits through queues," *IEEE Trans. Inform. Theory*, vol. 42, no. 1, pp. 4–18, Jan. 1996.

[82] J. J. Kaye, M. K. Levitt, J. Nevins, J. Golden, and V. Schmitt, "Communicating intimacy one bit at a time," in *CHI '05 Extended Abstracts on Human Factors in Computing Systems*. New York: ACM Press, 2005, pp. 1529–1532.

[83] J. Giles and B. Hajek, "An information-theoretic and game-theoretic study of timing channels," *IEEE Trans. Inform. Theory*, vol. 48, no. 9, pp. 2455–2477, Sept. 2002.

[84] J. T. Brassil, S. Low, and N. F. Maxemchuk, "Copyright protection for the electronic distribution of text documents," *Proc. IEEE*, vol. 87, no. 7, pp. 1181–1196, July 1999.

[85] S. H. Low and N. F. Maxemchuk, "Capacity of text marking channel," *IEEE Signal Processing Lett.*, vol. 7, no. 12, pp. 345–347, Dec. 2000.

[86] S. Verdú, "The exponential distribution in information theory," *Probl. Inf. Transm.*, vol. 32, no. 1, pp. 100–111, Jan.-Mar. 1996.

[87] I. C. Abou-Faycal, M. D. Trott, and S. Shamai (Shitz), "The capacity of discrete-time memoryless Rayleigh-fading channels," *IEEE Trans. Inform. Theory*, vol. 47, no. 4, pp. 1290–1301, May 2001.

[88] I. Csiszár, "On the error exponent of source-channel transmission with a distortion threshold," *IEEE Trans. Inform. Theory*, vol. IT-28, no. 6, pp. 823–828, Nov. 1982.

[89] Y. Zhong, F. Alajaji, and L. L. Campbell, "On the joint source-channel coding error exponent for discrete memoryless systems: Computation and comparison with separate coding," Queens University, Mathematics and Engineering Communications Laboratory, Tech. Rep., Dec. 2005.

[90] R. G. Gallager, *Information Theory and Reliable Communication*. New York: John Wiley & Sons, 1968.

[91] M. S. Pinsker, V. V. Prelov, and S. Verdú, "Sensitivity of channel capacity," *IEEE Trans. Inform. Theory*, vol. 41, no. 6, pp. 1877–1888, Nov. 1995.

[92] M. H. M. Costa, "Writing on dirty paper," *IEEE Trans. Inform. Theory*, vol. IT-29, no. 3, pp. 439–441, May 1983.

[93] A. S. Cohen and A. Lapidoth, "The Gaussian watermarking game," *IEEE Trans. Inform. Theory*, vol. 48, no. 6, pp. 1639–1667, June 2002.

[94] I. Csiszár and P. C. Shields, "Information theory and statistics: A tutorial," *Found. Trends Commun. Inform. Theory*, vol. 1, no. 4, pp. 420–527, Dec. 2004.

[95] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, 3rd ed. Budapest: Akadémiai Kiadó, 1997.

[96] M. F. Bear, B. W. Connors, and M. A. Paradiso, *Neuroscience: Exploring the Brain*, 3rd ed. Lippincott Williams & Wilkins, 2006.

[97] R. M. Alexander, *Optima for Animals*. London: Edward Arnold, 1982.

[98] G. A. Parker and J. Maynard Smith, "Optimality theory in evolutionary biology," *Nature*, vol. 348, no. 6296, pp. 27–33, Nov. 1990.

[99] E. R. Weibel, *Symmorphosis*. Cambridge: Harvard University Press, 2000.

[100] D. B. Chklovskii and A. A. Koulakov, "Maps in the brain: What can we learn from them?" *Annu. Rev. Neurosci.*, vol. 27, pp. 369–392, July 2004.

[101] S. Ramón y Cajal, *La Textura del Sistema Nerviosa del Hombre y los Vertebrados*. Springer, 1899.

[102] C. Allen and C. F. Stevens, "An evaluation of causes for unreliability of synaptic transmission," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 91, no. 22, pp. 10 380–10 383, Oct. 1994.

[103] N. A. Hessler, A. M. Shirke, and R. Malinow, "The probability of transmitter release at a mammalian central synapse," *Nature*, vol. 366, no. 6455, pp. 569–572, Dec. 1993.

[104] P. Isope and B. Barbour, "Properties of unitary granule cell→Purkinje cell synapses in adult rat cerebellar slices," *J. Neurosci.*, vol. 22, no. 22, pp. 9668–9678, Nov. 2002.

[105] A. Mason, A. Nicoll, and K. Stratford, "Synaptic transmission between individual pyramidal neurons of the rat visual cortex *in vitro*," *J. Neurosci.*, vol. 11, no. 1, pp. 72–84, Jan. 1991.

[106] M. Raastad, J. F. Storm, and P. Andersen, "Putative single quantum and single fibre excitatory postsynaptic currents show similar amplitude range and variability in rat hippocampal slices," *Eur. J. Neurosci.*, vol. 4, no. 1, pp. 113–117, Oct. 1992.

[107] C. Rosenmund, J. D. Clements, and G. L. Westbrook, "Nonuniform probability of glutamate release at a hippocampal synapse," *Science*, vol. 262, no. 5134, pp. 754–757, Oct. 1993.

[108] R. J. Sayer, M. J. Friedlander, and S. J. Redman, "The time course and amplitude of EPSPs evoked at synapses between pairs of CA3/CA1 neurons in the hippocampal slice," *J. Neurosci.*, vol. 10, no. 3, pp. 826–836, Mar. 1990.

[109] C. Koch, *Biophysics of Computation: Information Processing in Single Neurons*. New York: Oxford University Press, 1999.

[110] M. S. Goldman, "Enhancement of information transmission efficiency by synaptic failures," *Neural Comput.*, vol. 16, no. 6, pp. 1137–1162, June 2004.

[111] W. B. Levy and R. A. Baxter, "Energy-efficient neuronal computation via quantal synaptic failure," *J. Neurosci.*, vol. 22, no. 11, pp. 4746–4755, June 2002.

[112] A. Manwani and C. Koch, "Detecting and estimating signals over noisy and unreliable synapses: Information-theoretic analysis," *Neural Comput.*, vol. 13, no. 1, pp. 1–33, Jan. 2001.

[113] A. Zador, "Impact of synaptic unreliability on the information transmitted by spiking neurons," *J. Neurophysiol.*, vol. 79, no. 3, pp. 1219–1229, Mar. 1998.

[114] S. B. Laughlin, R. R. de Ruyter van Steveninck, and J. C. Anderson, "The metabolic cost of neural information," *Nat. Neurosci.*, vol. 1, no. 1, pp. 36–41, May 1998.

[115] W. B. Levy and R. A. Baxter, "Energy efficient neural codes," *Neural Comput.*, vol. 8, no. 3, pp. 531–543, Apr. 1996.

[116] S. Schreiber, C. K. Machens, A. V. M. Herz, and S. B. Laughlin, "Energy-efficient coding with discrete stochastic events," *Neural Comput.*, vol. 14, no. 6, pp. 1323–1346, June 2002.

[117] R. Sarpeshkar, "Analog versus digital: Extrapolating from electronics to neurobiology," *Neural Comput.*, vol. 10, no. 7, pp. 1601–1638, Oct. 1998.

[118] C. Holmgren, T. Harkany, B. Svennenfors, and Y. Zilberter, "Pyramidal cell communication within local networks in layer 2/3 of rat neocortex," *J. Physiol.*, vol. 551, no. 1, pp. 139–153, Aug. 2003.

[119] H. Markram, J. H. R. Lübke, M. Frotscher, A. Roth, and B. Sakmann, "Physiology and anatomy of synaptic connections between thick tufted pyramidal neurones in the developing rat neocortex," *J. Physiol.*, vol. 500, no. 2, pp. 409–440, Apr. 1997.

[120] P. J. Sjöström, G. G. Turrigiano, and S. B. Nelson, "Rate, timing, and cooperativity jointly determine cortical synaptic plasticity," *Neuron*, vol. 32, no. 6, pp. 1149–1164, Dec. 2001.

[121] T. W. Parks and J. H. McClellan, "Chebyshev approximation for nonrecursive digital filters with linear phase," *IEEE Trans. Circuit Theory*, vol. CT-19, no. 2, pp. 189–194, Mar. 1972.

[122] N. Brunel, V. Hakim, P. Isope, J.-P. Nadal, and B. Barbour, "Optimal information storage and the distribution of synaptic weights: perceptron versus Purkinje cell," *Neuron*, vol. 43, no. 5, pp. 745–757, Sept. 2004.

[123] E. Gardner, "Maximum storage capacity in neural networks," *Europhys. Lett.*, vol. 4, pp. 481–485, Sept. 1987.

[124] R. J. McEliece, E. C. Posner, E. R. Rodemich, and S. S. Venkatesh, "The capacity of the Hopfield associative memory," *IEEE Trans. Inform. Theory*, vol. IT-33, no. 4, pp. 461–482, July 1987.

[125] C. M. Newman, "Memory capacity in neural network models: rigorous lower bounds," *Neural Netw.*, vol. 1, no. 3, pp. 223–238, 1988.

[126] E. T. Rolls and A. Treves, *Neural Networks and Brain Function*. Ocford: Oxford University Press, 1998.

[127] L. G. Valiant, "Memorization and association on a realistic neural model," *Neural Comput.*, vol. 17, no. 3, pp. 527–555, Mar. 2005.

[128] H. Markram, "A network of tufted layer 5 pyramidal neurons," *Cereb. Cortex*, vol. 7, no. 6, pp. 523–533, Sept. 1997.

[129] S. Song, P. J. Sjöström, M. Reigl, S. Nelson, and D. B. Chklovskii, "Highly nonrandom features of synaptic connectivity in local cortical circuits," *PLoS Biol.*, vol. 3, no. 3, pp. 0507–0519, Mar. 2005.

[130] A. M. Thomson and A. P. Bannister, "Interlaminar connections in the neocortex," *Cereb. Cortex*, vol. 13, no. 1, pp. 5–14, Jan. 2003.

[131] A. M. Thomson, D. C. West, Y. Wang, and A. P. Bannister, "Synaptic connections and small circuits involving excitatory and inhibitory neurons in layers 2-5 of adult rat and cat neocortex: triple intracellular recordings and biocytin labelling *in vitro*," *Cereb. Cortex*, vol. 12, no. 9, pp. 936–953, Sept. 2002.

[132] J. Lisman, "Long-term potentiation: outstanding questions and attempted synthesis," *Proc.-R. Soc. Lond., Biol. Sci.*, vol. 358, no. 1432, pp. 829–842, Apr. 2003.

[133] D. H. O'Connor, G. M. Wittenberg, and S. S.-H. Wang, "Graded bidirectional synaptic plasticity is composed of switch-like unitary events," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 102, no. 27, pp. 9679–9684, July 2005.

[134] C. C. H. Petersen, R. C. Malenka, R. A. Nicoll, and J. J. Hopfield, "All-or-none potentiation at CA3-CA1 synapses," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 95, no. 8, pp. 4732–4737, Apr. 1998.

[135] D. O. Hebb, *The Organization of Behavior: A Neuropsychological Theory*. New York: Wiley, 1949.

[136] J. L. McGaugh, "Memory–a century of consolidation," *Science*, vol. 287, no. 5451, pp. 248–251, Jan. 2000.

[137] M. A. Lynch, "Long-term potentiation and memory," *Physiol. Rev.*, vol. 84, no. 1, pp. 87–136, Jan. 2004.

[138] R. G. M. Morris, "Long-term potentiation and memory," *Proc.-R. Soc. Lond., Biol. Sci.*, vol. 358, no. 1432, pp. 643–647, Apr. 2003.

[139] H. Kasai, M. Matsuzaki, J. Noguchi, N. Yasumatsu, and H. Nakahara, "Structure–stability–function relationships of dendritic spines," *Trends Neurosci.*, vol. 26, no. 7, pp. 360–368, July 2003.

[140] M. Matsuzaki, G. C. R. Ellis-Davies, T. Nemoto, Y. Miyashita, M. Iino, and H. Kasai, "Dendritic spine geometry is critical for AMPA receptor expression in hippocampal CA1 pyramidal neurons," *Nat. Neurosci.*, vol. 4, no. 11, pp. 1086–1092, Nov. 2001.

[141] V. N. Murthy, T. Schikorski, C. F. Stevens, and Y. Zhu, "Inactivity produces increases in neurotransmitter release and synapse size," *Neuron*, vol. 32, no. 4, pp. 673–682, Nov. 2001.

[142] Z. Nusser, R. Lujan, G. Laube, J. D. B. Roberts, E. Molnar, and P. Somogyi, "Cell type and pathway dependence of synaptic AMPA receptor number and variability in the hippocampus," *Neuron*, vol. 21, no. 3, pp. 545–559, Sept. 1998.

[143] T. Schikorski and C. F. Stevens, "Quantitative ultrastructural analysis of hippocampal excitatory synapses," *J. Neurosci.*, vol. 17, no. 15, pp. 5858–5867, Aug. 1997.

[144] Y. Takumi, V. Ramírez-León, P. Laake, E. Rinvik, and O. P. Ottersen, "Different modes of expression of AMPA and NMDA receptors in hippocampal synapses," *Nat. Neurosci.*, vol. 2, no. 7, pp. 618–624, July 1999.

[145] M. Matsuzaki, N. Honkura, G. C. R. Ellis-Davies, and H. Kasai, "Structural basis of long-term potentiation in single dendritic spines," *Nature*, vol. 429, no. 6993, pp. 761–766, June 2004.

[146] Q. Zhou, K. J. Homma, and M.-M. Poo, "Shrinkage of dendritic spines associated with long-term depression of hippocampal synapses," *Neuron*, vol. 44, no. 5, pp. 749–757, Dec. 2004.

[147] J. P. Pierce and L. M. Mendel, "Quantitative ultrastructure of Ia boutons in the ventral horn: scaling and positional relationships," *J. Neurosci.*, vol. 13, no. 11, pp. 4748–4763, Nov. 1993.

[148] L. C. Streichert and P. B. Sargent, "Bouton ultrastructure and synaptic growth in a frog autonomic ganglion," *J. Comp. Neurol.*, vol. 281, no. 1, pp. 159–168, Mar. 1989.

[149] M. B. L. Yeow and E. H. Peterson, "Active zone organization and vesicle content scale with bouton size at a vertebrate central synapse," *J. Comp. Neurol.*, vol. 307, no. 3, pp. 475–486, May 1991.

[150] C. Cherniak, M. Changizi, and D. W. Kang, "Large-scale optimization of neuron arbors," *Phys. Rev. E*, vol. 59, no. 5, pp. 6001–6009, May 1999.

[151] D. B. Chklovskii, "Synaptic connectivity and neuronal morphology: Two sides of the same coin," *Neuron*, vol. 43, no. 5, pp. 609–617, Sept. 2004.

[152] A. Hsu, Y. Tsukamoto, R. G. Smith, and P. Sterling, "Functional architecture of primate cone and rod axons," *Vis. Res.*, vol. 38, no. 17, pp. 2539–2549, Sept. 1998.

[153] G. Mitchison, "Neuronal branching patterns and the economy of cortical wiring," *Proc.-R. Soc. Lond., Biol. Sci.*, vol. 245, no. 1313, pp. 151–158, Aug. 1991.

[154] N. Kalisman, G. Silberberg, and H. Markram, "The neocortical microcircuit as a *tabula rasa*," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 102, no. 3, pp. 880–885, Jan. 2005.

[155] H. J. Koester and D. Johnston, "Target cell-dependent normalization of transmitter release at neocortical synapses," *Science*, vol. 308, no. 5723, pp. 863–866, May 2005.

[156] R. A. Silver, J. Lübke, B. Sakmann, and D. Feldmeyer, "High-probability uniquantal transmission at excitatory synapses in barrel cortex," *Science*, vol. 302, no. 5652, pp. 1981–1984, Dec. 2003.

[157] S. Cash and R. Yuste, "Linear summation of excitatory inputs by CA1 pyramidal neurons," *Neuron*, vol. 22, no. 2, pp. 383–394, Feb. 1999.

[158] D. B. Chklovskii, B. W. Mel, and K. Svoboda, "Cortical rewiring and information storage," *Nature*, vol. 431, pp. 782–788, Oct. 2004.

[159] P. Poirazi, T. Brannon, and B. W. Mel, "Arithmetic of subthreshold synaptic summation in a model CA1 pyramidal cell," *Neuron*, vol. 37, no. 6, pp. 977–987, Mar. 2003.

[160] A. Polsky, B. W. Mel, and J. Schiller, "Computational subunits in thin dendrites of pyramidal cells," *Nat. Neurosci.*, vol. 7, no. 6, pp. 621–627, June 2004.

[161] J. M. Bekkers and C. F. Stevens, "Quantal analysis of EPSCs recorded from small numbers of synapses in hippocampal cultures," *J. Neurophysiol.*, vol. 73, no. 3, pp. 1145–1156, Mar. 1995.

[162] J. del Castillo and B. Katz, "Quantal components of the end-plate potential," *J. Physiol.*, vol. 124, no. 3, pp. 560–573, June 1954.

[163] P. J. Sjöström, G. G. Turrigiano, and S. B. Nelson, "Neocortical LTD via coincident activation of presynaptic NMDA and cannabinoid receptors," *Neuron*, vol. 39, no. 4, pp. 641–654, Aug. 2003.

[164] A. A. Faisal, J. A. White, and S. B. Laughlin, "Ion-channel noise places limits on the miniaturization of the brain's wiring," *Curr. Biol.*, vol. 15, no. 12, pp. 1143–1149, June 2005.

[165] D. B. Chklovskii, T. Schikorski, and C. F. Stevens, "Wiring optimization in cortical circuits," *Neuron*, vol. 34, no. 3, pp. 341–347, Apr. 2002.

[166] Q. Wen and D. B. Chklovskii, "Segregation of the brain into gray and white matter: A design minimizing conduction delays," *PLoS Comput. Biol.*, vol. 1, no. 7, pp. 0617–0630, Dec. 2005.

[167] M. Jimbo and K. Kunisawa, "An iteration method for calculating the relative capacity," *Inf. Control*, vol. 43, no. 2, pp. 216–223, Nov. 1979.

[168] E. Uysal-Biyikoglu and A. El Gamal, "On adaptive transmission for energy efficiency in wireless data networks," *IEEE Trans. Inform. Theory*, vol. 50, no. 12, pp. 3081–3094, Dec. 2004.

[169] W. L. Root, "Estimates of $\epsilon$ capacity for certain linear communication channels," *IEEE Trans. Inform. Theory*, vol. IT-14, no. 3, pp. 361–369, May 1968.

[170] A. N. Kolmogorov and V. M. Tihomirov, "$\epsilon$-entropy and $\epsilon$-capacity of sets in functional spaces," *Uspekhi Mat. Nauk*, vol. 14, pp. 3–86, 1959.

[171] A. Papoulis and S. U. Pillai, *Probability, Random Variables and Stochastic Processes*, 4th ed. New York: McGraw-Hill, 2002.

[172] V. Balasubramanian, D. Kimber, and M. J. Berry, II, "Metabolically efficient information processing," *Neural Comput.*, vol. 13, no. 4, pp. 799–815, Apr. 2001.

[173] G. G. de Polavieja, "Reliable biological communication with realistic constraints," *Phys. Rev. E*, vol. 70, no. 6, p. 061910, Dec. 2004.

[174] S. Arimoto, "An algorithm for computing the capacity of arbitrary discrete memoryless channels," *IEEE Trans. Inform. Theory*, vol. IT-18, no. 1, pp. 14–20, Jan. 1972.

[175] R. E. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Trans. Inform. Theory*, vol. IT-18, no. 4, pp. 460–473, July 1972.

[176] J. G. Smith, "The information capacity of amplitude- and variance-constrained scalar Gaussian channels," *Inf. Control*, vol. 18, no. 3, pp. 203–219, Apr. 1971.

[177] A. Tchamkerten, "On the discreteness of capacity-achieving distributions," *IEEE Trans. Inform. Theory*, vol. 50, no. 11, pp. 2773–2778, Nov. 2004.

[178] M. C. Gursoy, H. V. Poor, and S. Verdú, "The noncoherent Rician fading channel—part I: structure of the capacity-achieving input," *IEEE Trans. Wireless Commun.*, vol. 4, no. 5, pp. 2193–2206, Sept. 2005.

[179] J. Huang and S. P. Meyn, "Characterization and computation of optimal distributions for channel coding," *IEEE Trans. Inform. Theory*, vol. 5, no. 7, pp. 2336–2351, July 2005.

[180] D. Guo, S. Shamai (Shitz), and S. Verdú, "Mutual information and minimum mean-square error in Gaussian channels," *IEEE Trans. Inform. Theory*, vol. 51, no. 4, pp. 1261–1282, Apr. 2005.

[181] A. Stepanyants, P. R. Hof, and D. B. Chklovskii, "Geometry and structural plasticity of synaptic connectivity," *Neuron*, vol. 34, no. 2, pp. 275–288, Apr. 2002.

[182] J. C. Magee and E. P. Cook, "Somatic EPSP amplitude is independent of synapse location in hippocampal pyramidal neurons," *Nat. Neurosci.*, vol. 3, no. 9, pp. 895–903, Sept. 2000.

[183] P. Sterling and G. Matthews, "Structure and function of ribbon synapses," *Trends Neurosci.*, vol. 28, no. 1, pp. 20–29, Jan. 2005.

[184] V. Braitenberg and A. Schüz, *Cortex: Statistics and Geometry of Neuronal Connectivity*. New York: Springer-Verlag, 1998.

[185] J. V. Harrington, "An analysis of the detection of repeated signals in noise by binary integration," *IRE Trans. Inform. Theory*, vol. 1, no. 1, pp. 1–9, Mar. 1955.

[186] Q. Zhou and M.-M. Poo, "Reversal and consolidation of activity-induced synaptic modifications," *Trends Neurosci.*, vol. 27, no. 7, pp. 378–383, July 2004.

[187] W. C. Abraham, B. Logan, J. M. Greenwood, and M. Dragunow, "Induction and experience-dependent consolidation of stable long-term potentiation lasting months in the hippocampus," *J. Neurosci.*, vol. 22, no. 21, pp. 9626–9634, Nov. 2002.

[188] D. J. Amit and S. Fusi, "Learning in neural networks with material synapses," *Neural Comput.*, vol. 6, no. 5, pp. 957–982, Sept. 1994.

[189] S. Fusi, P. J. Drew, and L. F. Abbott, "Cascade models of synaptically stored memories," *Neuron*, vol. 45, no. 4, pp. 599–611, Feb. 2005.

[190] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inform. Theory*, vol. IT-19, no. 4, pp. 471–480, July 1973.

[191] T. M. Cover, "A proof of the data compression theorem of Slepian and Wolf for ergodic sources," *IEEE Trans. Inform. Theory*, vol. IT-21, no. 2, pp. 226–228, Mar. 1975.

[192] B. R. Judd and N. S. Sutherland, "The information content of nonsequential messages," *Inf. Control*, vol. 2, no. 4, pp. 315–332, Dec. 1959.

[193] S. A. Savari, "Compression of words over a partially commutative alphabet," *IEEE Trans. Inform. Theory*, vol. 50, no. 7, pp. 1425–1441, July 2004.

[194] C. R. Rao, *Linear Statistical Inference and Its Applications*. New York: John Wiley & Sons, Inc., 1965.

[195] A. L. Buchsbaum, D. F. Caldwell, K. W. Church, G. S. Fowler, and S. Muthukrishnan, "Engineering the compression of massive tables: an experimental approach," in *Proceedings of the Eleventh Annual ACM-SIAM Symposium on Discrete Algorithms (SODA'00)*, San Francisco, Jan. 2000, pp. 175–184.

[196] A. L. Buchsbaum, G. S. Fowler, and R. Giancarlo, "Improving table compression with combinatorial optimization," *J. ACM*, vol. 50, no. 6, pp. 825–851, Nov. 2003.

[197] S. Vucetic, "A fast algorithm for lossless compression of data tables by reordering," in *Proceedings of the Data Compression Conference (DCC 2006)*, Snowbird, Utah, Mar. 2006, p. 469.

[198] T. M. Cover, "Poker hands and channels with known state," in *IEEE Commun. Theory Workshop*, Aptos, CA, 1999.

[199] N. Do, "Mathellaneous: A mathemagical card trick," *Australian Mathematical Society Gazette*, vol. 32, no. 1, pp. 9–15, Mar. 2005.

[200] P. Diaconis, "Mathematics and magic tricks," Clay Mathematics Institute, Cambridge, MA, Apr. 2006.

[201] R. J. Barron, B. Chen, and G. W. Wornell, "The duality between information embedding and source coding with side information and some applications," *IEEE Trans. Inform. Theory*, vol. 49, no. 5, pp. 1159–1180, May 2003.

[202] V. K. Goyal, "Theoretical foundations of transform coding," *IEEE Signal Processing Mag.*, vol. 18, no. 5, pp. 9–21, Sept. 2001.

[203] D. E. Ba and V. K. Goyal, "Nonlinear transform coding: Polar coordinates revisited," in *Proceedings of the Data Compression Conference (DCC 2006)*, Snowbird, Utah, Mar. 2006, p. 438.

[204] H. A. David and H. N. Nagaraja, *Order Statistics*, 3rd ed. Hoboken, NJ: Wiley-Interscience, 2003.

[205] V. Barnett, "The ordering of multivariate data," *J. R. Stat. Soc. Ser. A. Gen.*, vol. 139, no. 3, pp. 318–355, 1976.

[206] A. de Moivre, *The Doctrine of Chances: or, A Method of Calculating the Probabilities of Events in Play*, 3rd ed. London: A. Millar, 1756.

[207] I. Csiszár, "The method of types," *IEEE Trans. Inform. Theory*, vol. 44, no. 6, pp. 2505–2523, Oct. 1998.

[208] P. Jacquet and W. Szpankowski, "Entropy computations via analytic depoissonization," *IEEE Trans. Inform. Theory*, vol. 45, no. 4, pp. 1072–1081, May 1999.

[209] A. Orlitsky, N. P. Santhanam, K. Viswanathan, and J. Zhang, "Limit results on pattern entropy," *IEEE Trans. Inform. Theory*, vol. 52, no. 7, July 2006.

[210] R. Serfling, "Quantile functions for multivariate analysis: Approaches and applications," *Stat. Neerl.*, vol. 56, no. 2, pp. 214–232, May 2002.

[211] V. K. Goyal, S. A. Savari, and W. Wang, "On optimal permutation codes," *IEEE Trans. Inform. Theory*, vol. 47, no. 7, pp. 2961–2971, Nov. 2001.

[212] A. E. Sarhan and B. G. Greenberg, "Estimation of location and scale parameters by order statistics by singly and doubly censored samples," *Ann. Math. Stat.*, vol. 27, no. 2, pp. 427–451, June 1956.

[213] J. G. Saw and B. Chow, "The curve through the expected values of ordered variates and the sum of squares of normal scores," *Biometrika*, vol. 53, no. 1/2, pp. 252–255, June 1966.

[214] D. M. Mason, "Some characterizations of almost sure bounds for weighted multidimensional empirical distributions and a Glivenko-Cantelli theorem for sample quantiles," *Z. Wahrscheinlichkeitstheor. verw. Geb.*, vol. 59, no. 4, pp. 505–513, May 1982.

[215] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Boston: Kluwer Academic Publishers, 1992.

[216] P. P. Gandhi, "Optimum quantization of order statistics," *IEEE Trans. Signal Processing*, vol. 45, no. 9, pp. 2153–2159, Sept. 1997.

[217] K. M. Wong and S. Chen, "The entropy of ordered sequences and order statistics," *IEEE Trans. Inform. Theory*, vol. 36, no. 2, pp. 276–284, Mar. 1990.

[218] T. D. Lookabaugh and R. M. Gray, "High-resolution quantization theory and the vector quantizer advantage," *IEEE Trans. Inform. Theory*, vol. 35, no. 5, pp. 1020–1033, Sept. 1989.

[219] V. Erokhin, "$\varepsilon$-entropy of a discrete random variable," *Theory Probab. Appl.*, vol. 3, no. 1, pp. 97–100, 1958.

[220] W. H. R. Equitz and T. M. Cover, "Successive refinement of information," *IEEE Trans. Inform. Theory*, vol. 37, no. 2, pp. 269–275, Mar. 1991.

[221] A. M. Gerrish and P. M. Schultheiss, "Information rates of non-Gaussian processes," *IEEE Trans. Inform. Theory*, vol. IT-10, no. 4, pp. 265–271, Oct. 1964.

[222] L. R. Varshney and V. K. Goyal, "Toward a source coding theory for sets," in *Proceedings of the Data Compression Conference (DCC 2006)*, Snowbird, Utah, Mar. 2006, pp. 13–22.

[223] W. A. Pearlman and R. M. Gray, "Source coding of the discrete Fourier transform," *IEEE Trans. Inform. Theory*, vol. IT-24, no. 6, pp. 683–692, Nov. 1978.

[224] M. A. Changizi, Q. Zhang, H. Ye, and S. Shimojo, "The structures of letters and symbols throughout human history are selected to match those found in objects in natural scenes," *The American Naturalist*, vol. 167, no. 5, pp. E117–E139, May 2006.

[225] J. J. Leimer, "Design factors in the development of an optical character recognition machine," *IRE Trans. Inform. Theory*, vol. IT-8, no. 2, pp. 167–171, Feb. 1962.

[226] C. E. Shannon, "A note on a partial ordering for communication channels," *Inf. Control*, vol. 1, no. 4, pp. 390–397, Dec. 1958.

[227] D. A. Clark, P. P. Mitra, and S. S.-H. Wang, "Scalable architecture in mammalian brains," *Nature*, vol. 411, no. 6834, pp. 189–193, May 2001.

[228] D. A. Patterson and J. L. Hennessy, *Computer Organization & Design: The Hardware/Software Interface*, 2nd ed. San Francisco: Morgan Kaufmann Publishers, Inc., 1998.