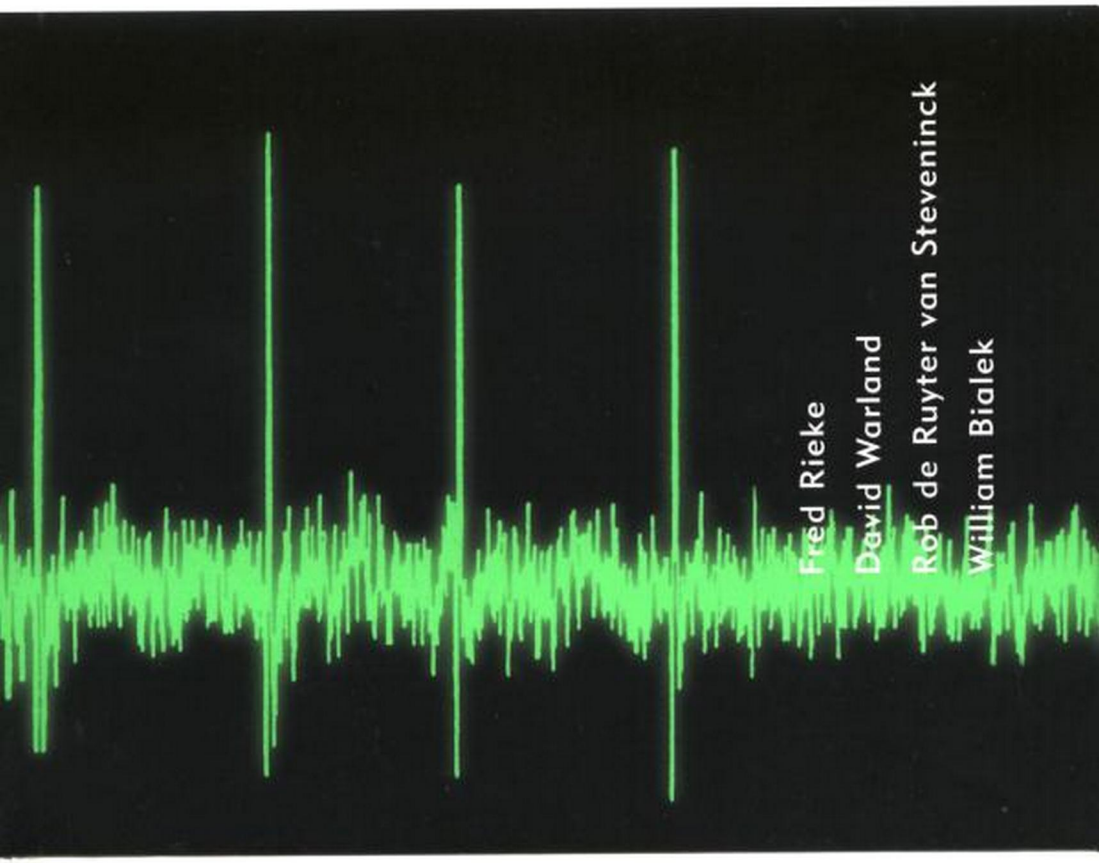
# SPIKES

## EXPLORING THE NEURAL CODE



Fred Rieke
David Warland
Rob de Ruyter van Steveninck
William Bialek

**Computational Neuroscience**
Terrence J. Sejnowski and Tomaso A. Poggio, editors

*Methods in Neuronal Modeling: From Synapses to Networks,* edited by Christof Koch and Idan Segev, 1989

*Neural Nets in Electric Fish,* Walter Heiligenberg, 1991

*The Computational Brain,* Patricia S. Churchland and Terrence J. Sejnowksi, 1992

*Dynamic Biological Networks: The Stomatogastric Nervous System,* edited by Ronald M. Harris-Warrick, Eve Marder, Allen I. Selverston, and Maurice Moulins, 1992

*The Neurobiology of Neural Networks,* edited by Daniel Gardner, 1993

*Large-Scale Neuronal Theories of the Brain,* edited by Christof Koch and Joel L. Davis, 1994

*The Theoretical Foundation of Dendritic Function: Selected Papers of Wilfrid Rall with Commentaries,* edited by Idan Segev, John Rinzel, and Gordon M. Shepherd, 1995

*Models of Information Processing in the Basal Ganglia,* edited by James C. Houk, Joel L. Davis, and David G. Beiser, 1995

*Spikes: Exploring the Neural Code,* Fred Rieke, David Warland, Rob de Ruyter van Steveninck, and William Bialek, 1997

# SPIKES
## EXPLORING THE NEURAL CODE

Fred Rieke, David Warland, Rob de Ruyter van Steveninck,
and William Bialek

A Bradford Book
The MIT Press
Cambridge, Massachusetts
London, England

First MIT Press paperback edition, 1999


This book was set in Times Roman by Windfall Software using ZzTEX and was printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Spikes : exploring the neural code / Fred Rieke . . . [ et al.].
p. cm. — (Computational neuroscience)
"A Bradford Book".
Includes bibliographical references and index.
ISBN 0-262-18174-6 (hc : alk. paper), 0-262-68108-0 (pb)
1. Neural transmission. 2. Sensory neurons. I. Rieke, Fred.
II. Series.
QP364.5.S66 1996
612.8—dc20 95-46161
CIP

10 9 8 7 6

*To our families*

# Contents

Series Foreword xi

Preface xiii

Acknowledgments xv

**Chapter 1**
**Introduction** 1

1.1 The classical results 2

1.2 Defining the problem 13

1.3 Central claims of this book 17

**Chapter 2**
**Foundations** 19

2.1 Characterizing the neural response 19

*2.1.1 Probabilistic responses and Bayes' rule* 21

*2.1.2 Rates, intervals, and correlations* 28

*2.1.3 Input/output analysis* 38

*2.1.4 Models for firing statistics* 49

2.2 Taking the organism's point of view 54

*2.2.1 Intervals in the signal and intervals between spikes* 55

*2.2.2 What can small numbers of spikes tell the brain?* 60

*2.2.3 Response-conditional ensembles* 63

2.3 Reading the code 76

*2.3.1 Why it might work* 76

*2.3.2 An experimental strategy* 88

*2.3.3 Qualitative features of a first test* 91

2.4 Summary 98

**Chapter 3**
**Quantifying information transmission** 101

3.1 Why information theory? 101

*3.1.1 Entropy and available information* 102

*3.1.2 Entropy of spike trains* 113

*3.1.3 Mutual information and the Gaussian channel* 121

*3.1.4 Time dependent signals* 127

3.2 Information transmission with spike trains 148

*3.2.1 Can we really "measure" information transmission?* 148

*3.2.2 Information transmission with discrete stimuli* 149

*3.2.3 Stimulus reconstruction and information rates* 156

3.3 Entropy and information with continuous stimuli 166

*3.3.1 Mechanical sensors in the cricket cercal system* 166

*3.3.2 Amphibian eyes and ears* 175

*3.3.3 Frogs and frog calls* 181

3.4 Summary 187

**Chapter 4**
**Reliability of computation** 189

4.1 Reliability of neurons and reliability of perception 189

*4.1.1 Historical background* 190

*4.1.2 Photon counting* 193

*4.1.3 Auditory discrimination* 204

*4.1.4 Motion discrimination in monkey vision* 214

4.2 Hyperacuity 221

*4.2.1 Where is the limit?* 221

*4.2.2 Experiments with single neurons* 228

*4.2.3 Temporal hyperacuity* 231

4.3 Motion processing in the fly visual system 235

*4.3.1 Limits to discrimination* 235

*4.3.2 Discrimination experiments with H1* 238

*4.3.3 Continuous estimation* 247

4.4 Summary 253



**Chapter 5**
**Directions** 255

5.1 Arrays of neurons 255

5.2 Natural signals 261

5.3 Optimal coding and computation 267

**Epilogue**
**Homage to the single spike** 279

**Appendix**
**Mathematical asides** 281

A.1 Rates as expectation values 281

A.2 Two-point functions 285

A.3 Wiener kernels 289

A.4 Poisson model I 295

A.5 Poisson model II 301

A.6 Estimation from independent responses 305

A.7 Conditional mean as optimal estimator 307

A.8 Practical calculations of reconstruction filters 311

*A.8.1 The "acausal-shifted" calculation* 311

*A.8.2 Power series expansions of the $K_n$* 313

A.9 Entropy of Gaussian distributions 316

A.10 Approximating the entropy of spike trains 317

A.11 Maximum entropy and spike counts 319

A.12 The Gaussian channel 327

A.13 Gaussians and maximum entropy 332

A.14 Wiener–Khinchine theorem 338

A.15 Maximizing information transmission 340

A.16 Maximum likelihood 348

A.17 Poisson averages 351

A.18 Signal to noise ratios with white noise 355

A.19 Optimal filters 362

References 369

Index 389

# Series Foreword

Computational neuroscience is an approach to understanding the information content of neural signals by modeling the nervous system at many different structural scales, including the biophysical, the circuit, and the systems levels. Computer simulations of neurons and neural networks are complementary to traditional techniques in neuroscience. This book series welcomes contributions that link theoretical studies with experimental approaches to understanding information processing in the nervous system. Areas and topics of particular interest include biophysical mechanisms for computation in neurons, computer simulations of neural circuits, models of learning, representation of sensory information in neural networks, systems models of sensory-motor integration, and computational analysis of problems in biological sensing, motor control, and perception.

Terrence J. Sejnowski
Tomaso A. Poggio

# Preface

This is a book about the way in which the nervous system represents or encodes sensory signals. Our approach to the problem of neural coding is motivated by a desire for quantitative analysis. In particular, we would like to describe the performance of neurons on an absolute scale, making precise the intuitive notion that these cells are telling the brain something about the sensory world.

The neural code is the subject of a vast literature, and we must make clear at the outset that what follows is not an encyclopedic guide to that literature. On the contrary, we try to focus on a small number of questions. We develop these questions in section 1.2, and in section 1.3 we make three claims about the answers. Subsequent chapters are devoted to the ideas and data that surround and support these three claims.

The questions we are asking about the neural code are phrased within a framework provided by ideas from probability and statistics, information theory, and the analysis of signals and noise. In the background are analogies between these ideas and ideas from statistical mechanics and thermodynamics. We contend that these concepts provide not only a natural language for talking about the issues, but also a set of concrete tools for the design and analysis of experiments.

This approach will appeal, we expect, to the growing numbers of physicists, mathematicians and engineers who are becoming involved in problems related to neural computation. But if this group is our only audience we will have failed. Since we are writing about the design and analysis of real experiments on real neurons, we must address ourselves to the community of experimental neurobiologists. It is, therefore, essential that we explore the connection of the different theoretical constructs to the quantities that are measured in the laboratory. As part of this effort, we try to use experimental data from real

neurons as an illustration of each new idea or mathematical tool, and as a result most of the figures in the text are constructed from real data.

In trying to make precise links between theory and experiment, we want to distinguish between ideas that are essential to the general discussion and the details or calculations that could be skipped on a first reading. Unlike experimental results or the results of computer simulations, all statements of mathematical fact can be checked by the reader with pen and paper. Ultimately, this checking tests understanding and builds intuition. On the other hand, a text that repeatedly points out all the wonderful places you can test your understanding can be a bit tiresome. As a compromise, we have collected many of the calculations into "mathematical asides" which are placed at the end of the text. Our hope is that many readers will find their way into these asides, but that by pulling the details aside we have left a main text that can be read and enjoyed.

# Acknowledgments

We have had the good fortune to collaborate with many different people, each of whom has helped shape our thinking about the problems discussed in this book. Most of their names appear somewhere in the text, but collecting them here gives us the opportunity to recall the pleasures of collaboration, and to say thank you once again to D. A. Bodnar, R. R. Capranica, R. H. Carlson, M. DeWeese, W. Hare, R. Koberle, M. Landolfa, S. B. Laughlin, B. P. M. Lenting, G. Lewen, E. R. Lewis, H. A. K. Mastebroek, M. Meister, J. P. Miller, R. Miller, K. T. Moortgat, W. G. Owen, M. Potters, D. L. Ruderman, S. Smirnakis, N. Socci, S. Strong, J. Vrieslander, W. Yamada, W. H. Zaagman, and A. Zee.

Convincing yourself that something is true or interesting is sometimes depressingly easy. We have been fortunate to have many friends and colleagues who are less easily convinced, and our arguments—er, discussions—with them have played an important role in sharpening our ideas and presentation. Not having gone to law school, we have found no way to make them responsible for the remaining gaps in our thinking, so we suggest that you use these ideas at your own risk. Again, collecting the names here reminds us of many enjoyable conversations, and we offer our thanks to J. J. Atick, H. B. Barlow, D. A. Baylor, W. Bruno, E. J. Chichilnisky, M. Crair, S. DeVries, H. Duifhuis, B. Edin, J. H. van Hateren, P. I. M. Johannesma, J. S. Joseph, A. J. Libchaber, L. Kruglyak, J. W. Kuiper, J. N. Onuchic, D. Seligson, A. Simon, D. G. Stavenga, and G. Zweig.

There are many times while writing a book when one wonders whether the result will be worth the effort. It is a great help to have colleagues who ask for copies of the current draft, at least creating the illusion that there are many people out there who are interested in the book. We don't know, of course, how many of these enthusiastic colleagues read the book, and we can only

assume that many are too polite to tell us what they really think. A few people, however, were not always so reserved, and we thank them for comments that had a direct effect on the final outcome: H. B. Barlow, C. D. Bialek, K. Miller, and T. J. Sejnowski. Several of the anonymous referees recruited by The MIT Press provided us with amazingly detailed reviews, and these were also a great help. We know that not everyone who offered advice will be happy with our responses, not least because different people offered conflicting advice. We have tried to rethink each section of the text that triggered comment, and through this process each comment had a real impact on the text. We hope that other authors will enjoy the help of equally thoughtful reviewers.

We have learned that there is a surprisingly large difference between being almost done with a book and being done. Our patient families have also noticed this large difference. We are therefore happy to thank those kind people who read and commented upon the "almost final" version of the manuscript: D. A. Baylor, M. Kvale, and K. Miller. They performed a great service in a remarkably short time.

During the writing of this book. F. R. held postdoctoral fellowships at The University of Chicago and at Stanford University, supported in part by grants from the National Institutes of Health, and D. W. was a postdoctoral fellow at Harvard University, supported in part by grants from the Office of Naval Research. We thank D. A. Baylor, M. Meister, and E. A. Schwartz for their support and encouragement, even when the manuscript took time from the real work in the laboratory. R. d. R. v. S. and W. B. have been at the NEC Research Institute, and we have all benefited from NEC's commitment to the continuing support of basic research. We especially thank J. A. Giordmaine for his interest in the project and for his efforts at making NECI an enjoyable and productive place to think and work.

We would accomplish very little, and certainly would not have finished this book, were it not for two of our colleagues, Mary Anne Rich and Allan Schweitzer not only helped us with matters related to this text, but also kept other things running while we were absorbed in writing. For these efforts, and for their persistent good humor, we offer our most sincere thanks.

This book has its origins in conversations with our editor at The MIT Press, Fiona Stevens. In the intervening years—we won't admit how many—she has encouraged us, cajoled us, commiserated with us, and never lost patience. We are grateful for all her help in bringing this text to press.

Here I must refer to the previous Waynflete Lectures given by Professor E. D. Adrian, on *The Physical Background of Perception*, because the results of physiological investigations seem to me in perfect agreement with my suggestion about the meaning of reality in physics. The messages which the brain receives have not the least similarity with the stimuli. They consist in pulses of given intensities and frequencies, characteristic for the transmitting nerve-fiber, which ends at a definite place of the cortex. All the brain 'learns' (I use here the objectionable language of the 'disquieting figure of a little hobgoblin sitting up aloft in the cerebral hemisphere') is a distribution or 'map' of pulses. From this information it produces the image of the world by a process which can metaphorically be called a consummate piece of combinatorial mathematics: it sorts out of the maze of indifferent and varying signals invariant shapes and relations which form the world of ordinary experience.

M. Born (1949)

# Chapter 1
# Introduction

Two friends, one living in the city and the other on the family farm, describe to one another the experiences of everyday life. The farmer conjures up pastoral images, acres of wheat swaying in a gentle breeze, the sweet smells of spring, and the songs of the birds. The city dweller recounts scenes of thousands of people emerging from the train station, the inescapable odors of traffic, and the throbbing beat of a street musician's drums. It would seem that these sensory experiences are as different as one could imagine, yet they share with all our sensory experiences one crucial feature: In each case, our perception of the world is constructed out of the raw data sent to our brains by our sensory nerves, and in each case these data come in the same standard form—as sequences of identical voltage pulses called action potentials or "spikes."

When we see, we are not interpreting the pattern of light intensity that falls on our retina; we are interpreting the pattern of spikes that the million cells of our optic nerve send to the brain. When we hear, we are not interpreting the patterns of amplitude and frequency modulation that characterize the acoustic waveform; we are interpreting the patterns of spikes from roughly thirty thousand auditory nerve fibers. All the myriad tasks our brains perform in the processing of incoming sensory signals begin with these sequences of spikes. When it comes time to act on the results of these computations, the brain sends out sequences of spikes to the motor neurons. Spike sequences are the language for which the brain is listening, the language the brain uses for its internal musings, and the language it speaks as it talks to the outside world.

If spikes are the language of the brain, we would like to provide a dictionary. We would like to understand the structure of this dictionary, perhaps even providing the analog of a thesaurus. We would like to know if, as in language, there are notions of context that can influence the meaning of the individual words. And of course we would like to know whether our use of the linguistic analogy makes sense. We must travel a long road even to give these questions a precise formulation. We begin at the beginning, more than two centuries ago.

## 1.1 THE CLASSICAL RESULTS

Our understanding of how the sensory world is represented in the electrical activity of the sensory nerves is limited, first and foremost, by our ability to record this activity. Indeed, the history of experiments on the electrical activity of nerves is intertwined with the history of electrical measurements more generally. The science of electricity as we understand it today began with Galvani and Volta in the 1700s (Pera 1986). Galvani observed that the muscles of a frog could be made to twitch when touched with a piece of metal, and he believed that the metal evoked "animal electricity" in the muscle. Volta suspected that the electricity was generated at the contact point itself, and that similar effects should be observable from a contact between different inorganic materials. Volta was right, and the pursuit of his ideas led him to what we now call a Voltaic pile, the first real battery. The fact that electricity was not the special provenance of animals was one of the first nails in the coffin of vitalism.

Galvani and Volta made macroscopic measurements. Their biological preparations consisted of large hunks of muscle—often the entire muscle—not what we now know to be the single muscle fibers or motor neurons that make up these tissues. The notion that the body is constructed from cells emerged only through the efforts of the nineteenth-century microscopists, which culminated in the beautiful observations of Ramón y Cajal on the cellular nature of the brain itself (Cajal 1909–11). As a more microscopic picture of the nervous system began to take shape, it seemed natural to ask how the activity of individual cells might relate to our perceptions. Müller developed the doctrine of specific nerve energies, according to which the identity of a sensory stimulus is represented by the fact that certain nerve fibers, and not others, are activated by that stimulus (Boring 1942). Helmholtz provided evidence for this view in his analysis of the inner ear, arguing that cells at different locations along the cochlear spiral are sensitive to different frequencies of sound (Helmholtz 1885). These discussions from the late nineteenth century form the foundation for much of our current thinking about the nervous system. When we read about a computational map in the cortex (Knudsen, du Lac, and Esterly 1987), where an array of neurons decomposes incoming signals according to the values of different component features, we are reminded of Helmholtz, who realized that the array of auditory nerve fibers would decompose sound into its component frequencies.

Testing the ideas of Helmholtz and Müller requires the direct observation of electrical activity in *individual* sensory neurons, not just the summed activity



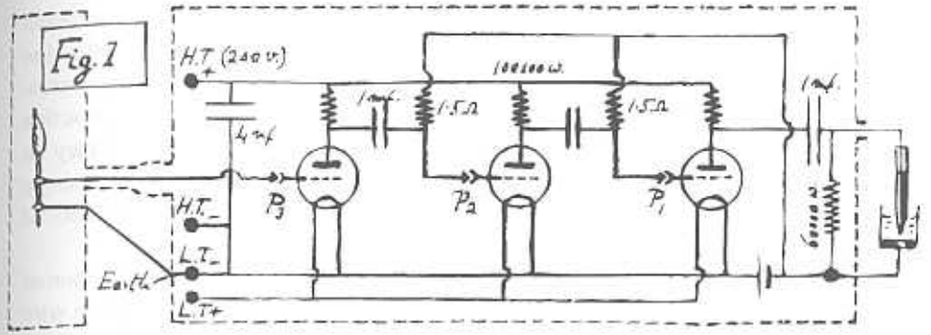



**Figure 1.1**
Schematic of Adrian's apparatus for recording the electrical activity in a nerve fiber. The fiber itself is at the far left. Adrian placed the fiber across two electrodes and measured the difference in the voltage at these two points along the nerve. The signal was amplified and used to control a mercury column, at the far right. Records were obtained by scanning a piece of film behind the mercury column, an example of which is shown in Fig. 1.3. Redrawn from Adrian (1926).

of a nerve bundle. But the electrical signals from individual cells are very small, at least when seen by an observer outside the cell. To pick up these small signals required a new method of low noise amplification, and this was provided in the first decade of this century by the vacuum tube. Using these new devices at Cambridge University, Lucas (1917) built instruments which allowed the recording of microvolt signals in bandwidths of several kiloHertz. We should remember that these experiments predate the oscilloscope, so even the display of submillisecond signals posed a significant problem. The solution to this problem, together with a general schematic of the instruments, is shown in Fig. 1.1. Lucas, sadly, died young, and the task of using these instruments fell to E. D. Adrian. In the space of roughly a decade, Adrian learned much of what we know to this day about the problem of neural coding. Independently, H. K. Hartline made many of the same discoveries. We follow the line of reasoning laid out by Adrian, and return shortly to some special features of Hartline's observations.

The classic early work of Adrian is contained, primarily, in a series of papers published in 1926 (Adrian 1926; Adrian and Zotterman 1926a, 1926b). Adrian summarized these results and their implications in a (still) very readable monograph, *The Basis of Sensation* (1928). One can trace the evolution of Adrian's thinking in two subsequent books (Adrian 1932, 1947).

Adrian's experiments established three fundamental facts about the neural code. First, he saw that individual sensory neurons produce stereotyped action potentials, or spikes. This is the all-or-none law, which had already been established for muscles and motor neurons: Incoming stimuli either produce action potentials, which propagate long distances along the cell's axon, or they do not; there are no intermediate signaling mechanisms. This means that a single neuron can provide information to the brain only through the arrival times of the spikes.

To make Adrian's observations a bit clearer, we look at a modern version of the same experiment. In Fig. 1.2 we show raw data from a fine tungsten wire electrode which has been placed close to a single neuron in the brain of a fly; the voltage at this electrode is measured relative to that at a reference electrode placed in the body fluids. Although the trace is noisy, there are clear, stereotyped events that can be isolated by appropriate filtering. These are the action potentials or spikes produced by this neuron and seen from outside the cell. The observation of all-or-none responses raises several questions:[1] Why does the nervous system choose this mode of communication? How is the stereotyped action potential waveform selected and stabilized? Is this mechanism universal?

Action potential propagation is an active process—the cell expends energy to produce and transmit a spike, and the energy expenditure increases the farther the spike must travel. In the absence of active processes, the electrical properties of cell membranes are such that a pulse starting at one end of a cell would spread and decay rather than propagating at constant velocity, and the characteristic decay lengths are on the order of one millimeter (Hodgkin and Rushton 1946). Therefore, passive mechanisms are inadequate for sending signals over long distances, such as the roughly one meter from your fingertips to your spinal cord, or even from one area of the cortex to a neighboring area; action potentials provide the means for such long distance communication. On the other hand, cells that send signals only over short distances, such as within the retina or even across the body of a small animal, need not generate action potentials and can, instead, operate entirely with "graded" voltage responses to sensory stimuli (Roberts and Bush 1981); we will see examples of this more continuous mode of neural signalling in section 3.1.4.

1. The experiments and theoretical developments which provided the answers to these questions are by now classic chapters in the history of neuroscience (Aidley 1989). We provide only a brief summary, but we encourage the reader to look at the original papers, as well as the lovely text by Katz (1966). Some of the history is recounted in the essays collected for the one-hundredth anniversary of the Physiological Society (Hodgkin et al. 1977).



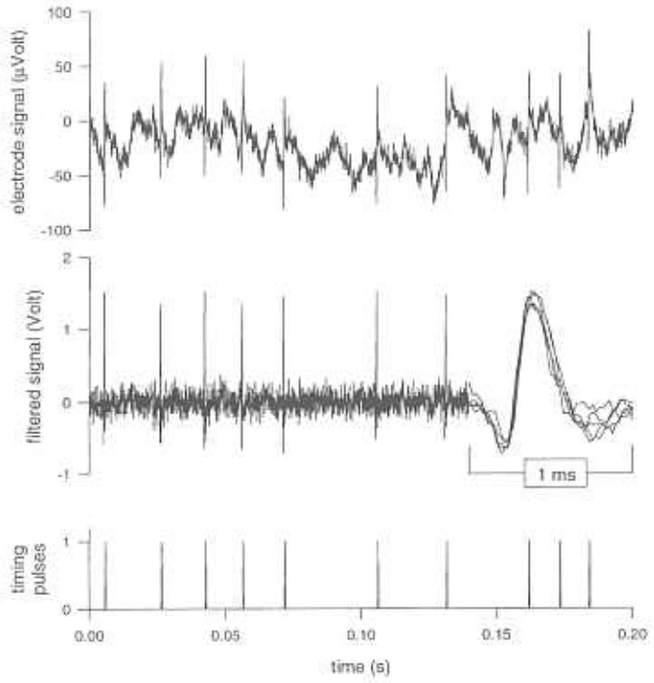



**Figure 1.2**
All-or-none coding by action potentials. Each action potential generated by the cell has a similar shape. Thus action potentials are the elementary units of the neural code. The top panel shows the difference between the voltage recorded with a fine tungsten wire placed near a cell in the fly's brain and that recorded with a reference electrode placed in the body fluid. The middle panel shows the same voltage after band-pass filtering to separate the relatively high frequency components in the action potential from low frequency noise; after filtering, the shapes of individual action potentials are quite similar. At the right, five action potentials are shown overlaid on an expanded time scale. This gives an impression of the shape and of the reproducibility of the time course. The bottom panel shows timing pulses generated electronically by a threshold discriminator circuit.

The local circuit properties of a cell membrane include active elements, conductances that are modulated by voltage changes and are electrically in series with power supplies (or, effectively, batteries) that are maintained by ion pumps; these pumps in turn are powered by chemical energy from the cell's metabolism. Hodgkin and Huxley (1952a, 1952b, 1952c) analyzed the electrical dynamics of the cell membrane in the giant axon of squid, and showed that these dynamics could be described with relatively simple phenomenological models of conductances that depend on voltage and are selective for different ions. When these local, active elements are assembled into a long cable, such as the axon, the nonlinear dynamics of the conductances select a stereotyped pulse which can propagate at constant velocity, while all other voltage changes eventually decay; the great triumph of this work was to show that this pulse has a shape and speed essentially identical to the observed action potentials (Hodgkin and Huxley 1952d). The mathematics of pulse selection has its roots in the nineteenth century, but a complete theory came much later (Aronson and Weinberger 1978), and the Hodgkin–Huxley equations continue to provide the inspiration for interesting mathematics and physics problems.

Although their analysis was purely phenomenological, the form of the Hodgkin–Huxley equations suggested a microscopic picture in which the conductances selective for different ions correspond to different molecular elements, or channels, in the membrane, and the modulations of the conductance correspond to transitions among discrete states of these channel molecules. Continuing advances in low noise amplification made it possible to resolve the electrical noise generated by spontaneous transitions among the different channel states, and finally to detect the currents flowing through single channel molecules (Sakmann and Neher 1983). Measurements on the properties of individual channel molecules, together with the techniques of modern molecular biology, have made it possible to identify a great diversity of channel types (Hille 1992), but these studies also demonstrate that many features of channel structure and function are strongly conserved throughout the animal kingdom (Jan and Jan 1994). This universality of mechanism at the molecular level harks back to Adrian's observations on the universality of spike encoding. Over the years, Adrian and his colleagues recorded the activity of sensory neurons from an enormous variety of different sensory systems in different animals. Although the quantitative details vary from neuron to neuron, it seems that the principles are universal.

The second of Adrian's fundamental observations was that, in response to a static stimulus such as a continuous load on a stretch receptor, the rate of spiking increases as the stimulus becomes larger. The raw data from Adrian's



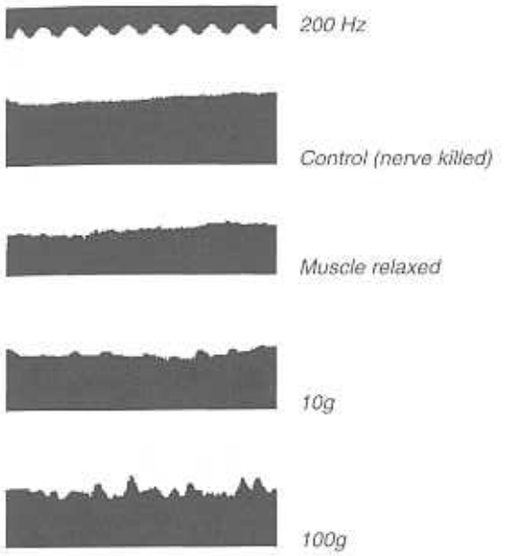



**Figure 1.3**
Firing rate as a function of stimulus strength, adapted from Adrian and Zotterman (1926a). The spikes in these panels are visible as the fluctuations riding on the black-white interface. A time marker is shown on top. Adrian and Zotterman measured the relation between the force applied to a muscle and the firing rate in a stretch receptor embedded in the muscle. Different forces were generated by hanging weights with different masses from the muscle. This type of experiment established that the frequency of firing in sensory neurons increased with increasing stimulus strength.

original demonstration of this principle is shown in Fig. 1.3, and a quantitative analysis is shown in Fig. 1.4a. Thus the rate, or frequency, of spikes indicates the intensity of the stimulus. To be a bit more precise, the number of spikes in a fixed time window following the onset of a static stimulus represents the intensity of that stimulus. This is the idea of *rate coding*.

The third of Adrian's discoveries was that if a static stimulus is continued for a very long time, the spike rate begins to decline, as illustrated in Fig. 1.4b. This is called *adaptation*, although this term is also used more generally to describe a dependence of the neural response on the history of stimulation. Adrian suggested that this physiological phenomenon corresponds to perceptual phenomena wherein we become gradually unaware of constant stimuli.

As we have tried to find a precise modern formulation for the problem of neural coding, we have been struck by the extent to which the ideas of Adrian

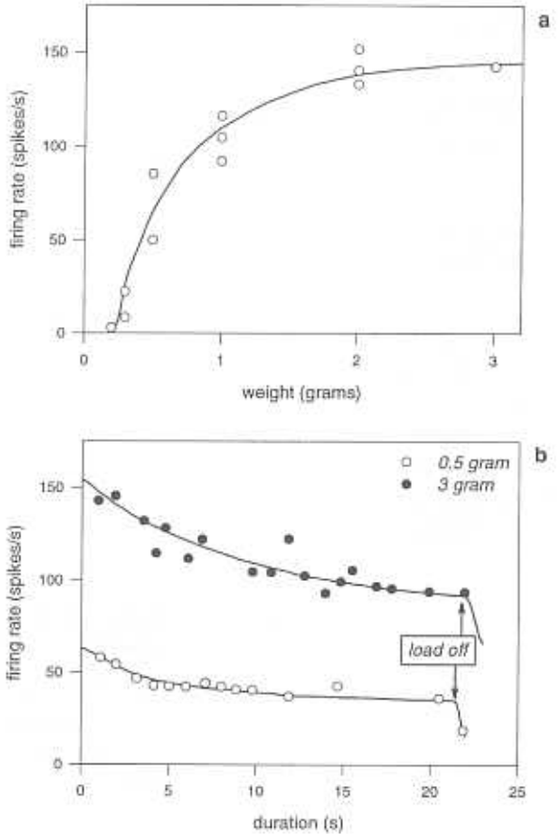



**Figure 1.4**
Rate coding and adaptation. (a) Average firing rate of a stretch receptor as a function of the weight applied to the muscle, in an experiment similar to that of Fig. 1.3. (b) Decrease in firing rate with time following the onset of a static stimulus at $t = 0$, adapted from Adrian (1926). This desensitization, or *adaptation*, is a general property of neural coding.



and Hartline have formed the paradigm for subsequent exploration of the nervous system. On the one hand this must mean that their early experiments captured essential and universal features of the neural code. On the other hand one must worry that, in following this single line of ideas, some crucial points may have been missed.

In the first experiments on single sensory neurons the stimulus was often defined by a single parameter. This parameter, such as the load on a stretch receptor, was held fixed while the stimulus was on. But naturally occurring stimuli are defined by a much larger number of parameters. In vision, for example, a small region of the visual field may be described by its overall luminance, but also by its contrast relative to the background, the size and shape of any features in the region, the positions and orientations of such features, their color, depth, and so on. By analogy with the Adrian–Hartline observations on spike rate as a function of stimulus intensity, one can plot the responses of a visual neuron as a function of these multiple parameters. This leads to the notion of *feature selectivity*, in which the cell's response depends most strongly on a small number of parameters and is maximal at some optimum value of these parameters.

Precursors to the notion of feature selectivity can be found in the work of Hartline and collaborators, who studied the responses of single neurons from the compound eyes of the horseshoe crab *Limulus polyphemus*. In addition to reproducing Adrian's results concerning rate coding, Hartline found that the stimulus whose strength was coded by one neuron reflected the difference between the light intensity at the location of that cell and the intensity at neighboring cells. Thus the crab retina has an enhanced response to spatial contrast or edges. Hartline, Ratliff, and coworkers suggested that this enhancement is connected to the perceptual phenomenon of Mach bands, shown schematically in Fig. 1.5. The unraveling of the retinal circuitry responsible for contrast enhancement led to a long sequence of now classic papers (Ratliff 1974).

The concept of feature selectivity was clearly enunciated by Barlow (1953a, 1953b), who was Adrian's student. Recording from retinal ganglion cells in the frog, he showed that the response of these cells to a spot of light at first grows with the area of the spot, but then declines if the spot exceeds a critical size, as summarized in Fig. 1.6a. The portion of the visual world that can influence the activity of a neuron is called the receptive field of that cell, and Barlow's results can be described as a "center–surround" organization of the receptive field: spots within a small region (the center) excite the cell, but spots just outside this region (in the surround) inhibit the cell (Fig. 1.6b).

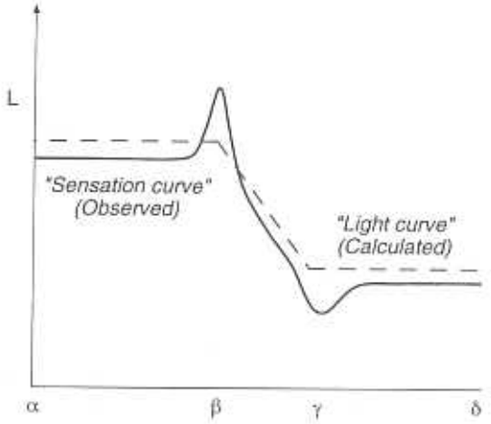



**Figure 1.5**
Mach bands at the edge of a shadow. The "Light curve" is based on physical calculations of the luminance at the edge of a shadow. The point $\alpha$ is in the fully illuminated space, the point $\beta$ is at the outer edge of the shadow, the point $\gamma$ is at the inner edge of the shadow, and the point $\delta$ is in the full shadow. The thicker line represents the apparent luminance, or "Sensation curve" actually observed. The maximum and minimum of these curves correspond to the light and dark Mach bands that arise from differencing mechanisms in the visual system that enhance contrast. Redrawn from Ratliff (1974).

To a good approximation these receptive fields are circularly symmetric. Essentially identical receptive fields were found in cat retinal ganglion cells by Kuffler (1953). In many cases the excitation and inhibition are balanced so that spatially uniform stimulation produces no response. Another interpretation is that these cells are tuned to objects of a given apparent size, perhaps that of the bugs the frog is hunting. The picture of frog retinal ganglion cells as specialized "bug detectors" was emphasized by Lettvin and coworkers (1959). In the limiting case this view presents sensory neurons as yes/no devices, signaling the presence or absence of certain elementary features.

The importance of feature selectivity was strongly supported by the observations of Kuffler's colleagues Hubel and Wiesel (1962). They found that many cells in cat visual cortex are selective not only for the size of objects (e.g., the width of a bar) but also for their orientation. As in the Barlow–Kuffler experiments, Hubel and Wiesel observed this selectivity by counting the number of spikes the cell produced in response to the presentation of a static stimulus or in response to the motion of the stimulus through the cell's receptive field. Hubel and Wiesel presented a scenario for how this orientation selectiv-



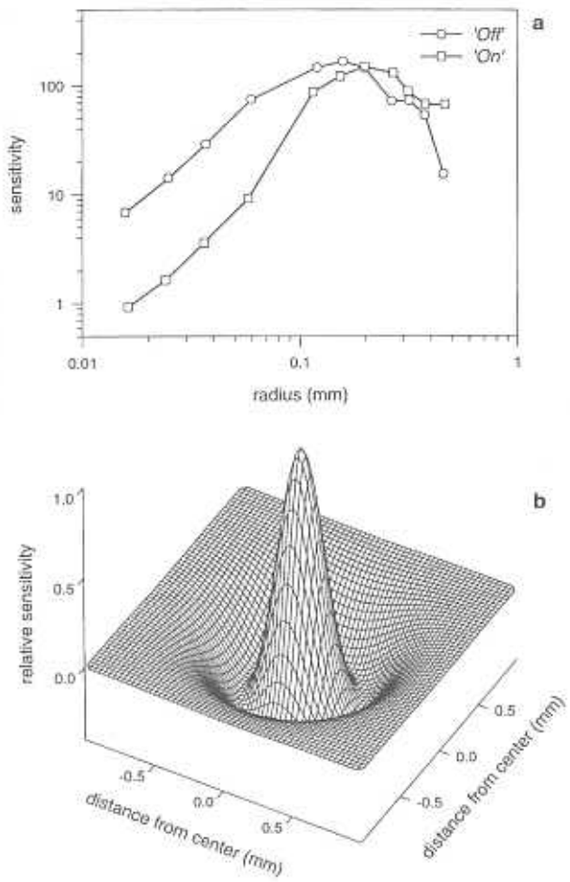



**Figure 1.6**
Center–surround receptive fields in retinal ganglion cells. (a) Sensitivity of retinal ganglion cells in the frog as a function of the radius of the light stimulus; sensitivity is defined as the light intensity required to elicit a fixed number of spikes. As the stimulus size is increased, the sensitivity initially increases, but then begins to decrease when the stimuli are larger than 0.2 mm in radius. This behavior was seen in both "on" ganglion cells, which respond to an increase in light intensity in the central region of their receptive field, and in "off" ganglion cells, which respond to a decrease in light intensity. (b) Receptive field organization suggested by Barlow to explain measurements such as those in (a). Light falling within the central excitatory region of the cell's receptive field causes an increase in the number of spikes, while light falling in the inhibitory surround causes a decrease in the number of spikes, indicated here as a negative sensitivity. Maximal response to a spot of light is achieved when the stimulus just covers the entire receptive field center.

ity could be built out of center–surround neurons in lower levels of the visual system, making explicit the intuitive notion that higher percepts are built out of elementary features. Finally, they found that neighboring neurons are tuned to neighboring orientations, so that feature selectivity is mapped over the surface of the cortex. This notion of cortical mapping, presaged by Mountcastle's (1957) observations on the responses of cells in the somatosensory cortex, revealed order amid the seemingly impenetrable mass of cortical circuitry. This discovery led, in turn, to the investigation of how this order develops out of the more amorphous circuitry of the embryonic brain. The ideas of feature selectivity, cortical maps, and self-organization of maps during development have dominated the exploration of cortex ever since (Hubel and Wiesel 1977).

If we return to the original Adrian–Hartline experiments on sensory neurons, we see that one could extend the description of the neural code in two very different directions. One direction is to study the coding of multiparameter stimuli, which has been followed extensively in the exploration of the visual system. A second direction is to use stimuli with realistic time dependencies. In a natural environment, sensory inputs are not broken into discrete presentations, and they are not simply turned on and off. More complex dynamic signals have been used in the study of the auditory system, where the main issues concern recognition and classification of temporal waveforms. But even in these experiments there is a tendency to approximate real dynamic signals with more elaborate but still essentially stationary signals. For example, the coding of vowel sounds has often been studied using continuous, periodic stimuli whose power spectra approximate those of real vowels.

A primary concern in this book is to understand how the nervous system represents signals with realistic time dependencies. The problem of coding for nearly static stimuli is very different from the problems faced by the brain under more natural conditions. In particular, the focus on time dependent signals forces us to think about the significance of much smaller numbers of spikes. But we are getting ahead of ourselves.

Do the ideas of rate coding, feature selectivity, and cortical mapping tell us what we want to know about the neural code? Certainly the fact that neurons in deeper layers of the brain are selective for more complex features tells us something about the kinds of computations that are carried out as sensory signals are passed from one stage of processing to the next, although it is dangerous to take a hierarchical or sequential view of sensory processing too literally. The idea of rate coding leaves open the question of whether other features of the spike train—generally grouped under the catch phrase *timing*—carry meaningful information, and indeed this question has been central to



many discussions of neural coding. The idea of mapping leads us to think about the representation of the sensory world in arrays of neurons; it also leads to the concepts of ensemble or population coding, which are active topics of current research.

The classical results on the neural code suggest many avenues for exploration. We cannot do justice to all the different paths taken by different investigators. In the following section we hope to make precise a more limited set of questions, which, with luck, we can answer in the space of the remaining text.

## 1.2 DEFINING THE PROBLEM

What would it mean to say that we "understand" the neural code in a particular region of the nervous system? How do we quantify the notion that the spike train of a single cell "conveys information" about the sensory world? In what sense is a particular sequence of spikes the "right answer" to some computational problem faced by the brain? We search for sharper versions of these questions by forcing ourselves to adopt a more precise and more mathematical language. In talking about the nervous system we routinely make colloquial use of terms such as code, information, and reliability. All of these words can be given precise mathematical definitions, and we hope, through the remainder of the text, to convince the reader that these definitions provide a clear guide to the design and analysis of new experiments. In striving for precision we shall see the emergence of some new ideas. We begin, however, by revisiting an old idea, the homunculus.

The homunculus is an often derided concept in discussions of the brain. We recall that this metaphor conjures up a little man—or, in a lovely variant by Michael Land and Simon Laughlin (Fig. 1.7), a little fly—who observes the responses of his own sensory neurons and finally forms the percepts that the organism experiences. The problem with this picture is that it never gets to the essence of what it means to perceive and to experience the world. On the other hand, as explorers of the nervous system we place ourselves, inevitably, in the position of the homunculus—we observe the responses of sensory neurons and try to decide what these responses could mean to the organism. This problem of assigning meaning to the activity of sensory neurons is the central issue in our discussion of the neural code.

It is easy to imagine that the task of the homunculus is trivial—after all, he just watches a projected image of the world as it flashes through the brain. But this projected image is *encoded* in the patterns of action potentials generated by the sensory neurons. It is not at all clear what the homunculus would have

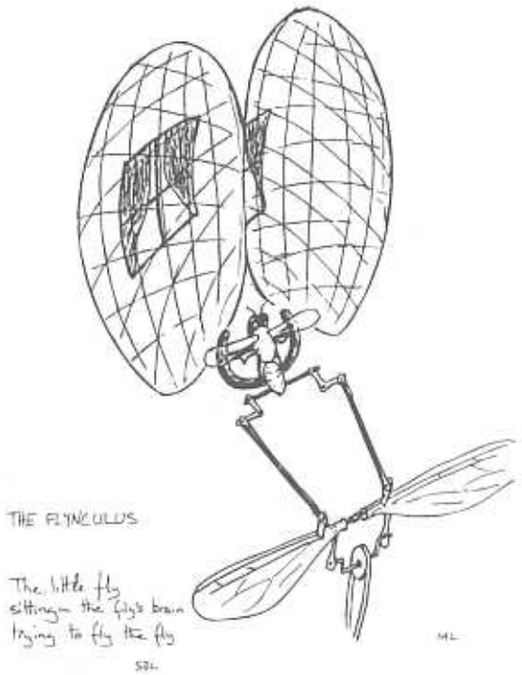



**Figure 1.7**
The Flynculus. Doodle by M. F. Land, quotation, "The little fly sitting in the fly's brain trying to fly the fly," from S. B. Laughlin, with permission.

to do, even in principle, to make sense out of these encoded data. We propose that "understanding the neural code" means that we would know how to make sense out of the bewildering array of spike trains streaming in from the sense organs: If we understand the code, we can function as the homunculus.

When we ask what a spike train means, or what it can tell us about the world, we need to set some boundaries for the question, or, equivalently, a context for the answer. If we live in a world with only two possible sensory stimuli, we can ask how the homunculus could best use the spike train data to make a decision about which stimulus in fact occurred. This decision rule would constitute a complete understanding of the neural code, assuming that the world offers just two possible signals.

In many psychophysical discrimination experiments (Green and Swets 1966), a world of two alternatives is created artificially, and the subject must



solve the problem of choosing between these alternatives. This binary decision problem provides a convenient context for asking questions about the reliability of our perceptions, and we shall see that it is also useful for investigating the reliability of neurons. But it is not enough to build a homunculus that functions in a world of two alternatives; we want to ask our question about the meaning of spike trains in a context that approaches the complexity of the natural world.

Under natural conditions, the stimulus that will appear in the next brief time window is not known to us in advance. Instead the stimulus is chosen from an infinite set of alternatives. On the other hand, these alternatives are not all equally likely. While there are blue spruce trees, green trees do not suddenly turn blue (or red or yellow either). Natural stimuli develop in time, and these dynamics have some underlying regularity or structure. This structure has a deterministic component, as when a leaf falls downward according to Newton's laws. But since we do not know all the forces that shape the dynamics of sensory stimuli, some aspects of these stimuli are unpredictable, as when the falling leaf is deflected by a gust of wind. The result is that natural signals are presented to us at random, but these signals have correlations that reflect their origins in deterministic physical processes.

Rather than inhabiting a world of two alternatives, we thus inhabit a world of random but correlated time dependent signals. The time dependence is crucial, because it means that we cannot wait forever to decide what we are looking at. Not only does biology press for quick decisions—we must catch our prey and not be caught by predators—the physics of our environment is such that any simple averaging for long periods of time will average away the very signals that interest us. The task of the homunculus, then, is not to create a static image of the sensory world from the input spike trains, but rather to give a sort of running commentary or simultaneous translation. We emphasize that this running commentary need not be, and most likely cannot be, a comprehensive reconstruction of the world around us.

To give meaning to the spike trains nonetheless requires that we recreate at least some aspects of the continuous time dependent world that is encoded in discrete sequences of spikes. From our experience in the laboratory we know that when forced to interpret rapidly changing signals we are very susceptible to noise; usually we try to combat noise by averaging in time or by averaging over repeated presentations of the same signal. But the homunculus is not free to set arbitrary averaging times, and he certainly cannot ask for a second, identical copy of the immediate past. On the contrary, the homunculus (and

the animal as well!) has to reach conclusions about the world from just one example of the spike train in each of his sensory neurons.

In generating a running commentary on the meaning of spike trains we shall have to deal with whatever level of noise is present in these data. Ideally, our interpretation of the spike trains should be as reliable as possible given this noise, and the statistically sophisticated homunculus would report confidence levels on his estimates of what is happening in the world. If understanding the neural code means building a homunculus, we can compare two different candidate homunculi—two different candidates for the structure of the neural code—by comparing the accuracy of their inferences about events in the sensory world.

We are closing in, then, on a more precise definition of the problems in understanding the neural code. We place ourselves in the position of the homunculus, monitoring the spike trains of sensory neurons as stimuli vary in time along some unknown trajectory. We must generate a running commentary on the identity of these stimuli, using only the spike train data as input. Our inferences about events in the world will have some limited accuracy, and we shall have to quantify this accuracy. Out of many possible homunculi, there is one that tells us as much as possible about the world given the noise in the spike train data itself. The performance of this best homunculus will reflect a compromise between averaging in time to combat noise and responding quickly to keep up with the dynamics of the world, and we shall have to be precise about these time scales.

The construction of a complete homunculus, or even the complete flynculus of Fig. 1.7, is a daunting task. In the fly, visual signals stream in along thousands of parallel paths reflecting the array of lenses in the compound eye, and in ourselves and our primate cousins the corresponding numbers are three orders of magnitude larger. There are a few special cases, such as the moths discussed in section 4.1.1 (Roeder 1963), for which it might be possible to monitor all of the spike trains that encode one sensory modality, but in general this is hopeless. As we have noted, however, there is a long tradition of trying to make sense out of the responses of single neurons, always recognizing that one cell can tell us about only a small piece of the sensory environment. In this tradition, most of this book is about the problem of an impoverished homunculus who looks at the spike train of just one neuron at a time; we take a brief look at the problem of multiple neurons in section 5.1. We thus have a clear question, amenable to experimental investigation: What can the spike train of this one neuron tell us about events in the world?



## 1.3 CENTRAL CLAIMS OF THIS BOOK

Nearly seventy years ago, Adrian summarized the first generation of experiments on neural coding (Adrian 1928). We have argued that, even today, this classic work contains a large fraction of what we know about the language of the brain. Forty years later, Perkel and Bullock (1968) provided an encyclopedic summary of the state of the field, a handbook of diverse candidate coding strategies in different systems. What can we add after all these years?

We believe that there has been substantial progress in both the formulation and the resolution of three major issues regarding coding by single neurons. These three points form the core of our presentation:

*1. Representation of time-dependent signals.* In a variety of sensory systems, single neurons produce on the order of one spike per characteristic time of stimulus variations—a sparse temporal representation. This is in direct contradiction to a simple, intuitive implementation of the rate coding idea, since the rate is an average quantity not available from a single spike. Sparse temporal codes can be decoded by simple algorithms, even when the encoding is a complex nonlinear process. Thus the problem of *decoding*—the problem solved by our homunculus—may be simpler than the classical problem of encoding.

*2. Information rates and coding efficiency.* The focus on signals with realistic time dependencies leads to the demonstration that single neurons can transmit large amounts of information, on the order of several bits per spike. In at least one case, signals with more natural temporal correlations are coded more efficiently, so that the spike train provides more information with roughly the same number of spikes. These high rates come close to saturating the fundamental physical limits to information transmission.

*3. Reliability of computation.* Understanding the reliability of the nervous system requires that we understand the code which the system uses to represent the answers to its computational problems; the study of neural coding is thus tied to much broader issues of neural computation. In several systems there is agreement between at least two of three fundamental quantities: The reliability of behavior, the reliability of single neurons, and the fundamental physical limits to reliability imposed by noise in the sense data itself. It is clear that the approach to the physical limits is closest for the more natural tasks of processing time-dependent signals.

These three ideas provide, we hope, a clear answer to the questions formulated in section 1.2. Decoding the spike train provides a literal construction of

the "running commentary" that we require from the homunculus, the measurement of information transmission rates quantifies how much our impoverished homunculus can tell us by looking at just one neuron, and the observations on reliability place this information on a meaningful scale relative to the capabilities of the whole organism.

In exploring these three issues, we will refer to experimental results from many different systems, obtained by many different groups over a period of several decades. The common thread running through these diverse studies is the attempt to *quantify* the behavior of neurons, specifically under conditions that approximate the function of the nervous system in the life of the organism. Much of the text is also concerned with methodology, reviewing several theoretical approaches that have been proposed as guides to the design and analysis of quantitative experiments. Many of our readers may reasonably wonder whether the effort of building up this more mathematical framework will be rewarded. One reason for persevering is that the quantitative analysis of neural coding leads to surprising results. As devices for transmitting and processing information, neurons are doing much more than one might have expected, and in a precise sense they are doing almost as much as is physically possible. Even simple quantitative questions—how many spikes carry a meaningful signal?—have surprising answers. Thus we claim that the results of a quantitative approach are sufficiently extreme that they begin to alter our qualitative conception of how the nervous system works.

# Chapter 2
# Foundations

How do we quantify the behavior of spiking neurons? What would a model of the neural response look like, and how would we decide that this model was sufficient or complete? In this chapter we review several different methods for the quantitative characterization of neural responses and discuss the relation of these different characterizations to the problems of coding outlined in the introduction. We hope to provide a view that combines mathematical ideas about encoding and decoding with an intuition from physiology and ethology about the time scales involved in neural signal processing. In particular, these natural time scales are such that it can be quite easy to decode the neural response. This allows us to understand the neural code in the practical sense that we can say what the spikes from a given neuron mean about signals in the outside world.

## 2.1 CHARACTERIZING THE NEURAL RESPONSE

In the early experiments of Adrian and Hartline, the response of a neuron was measured by counting the number of spikes in a fixed time window following the onset of the stimulus. In modern experiments, one typically repeats the same stimulus many times and averages over these repeated presentations. The first thing one sees in such experiments—before averaging—is that the spike train is not identical on each trial, so that there appears to be an element of randomness in the neural response, an example of which is shown in Fig. 2.1. The observation of random responses raises several questions: How do we quantify the degree of randomness in the neural response? What are the origins of this randomness, and how does it limit the reliability of information transmission and computation in the nervous system? What is the correct empirical characterization of the neural response given that this response is not completely reproducible from trial to trial? We turn first to this last question.

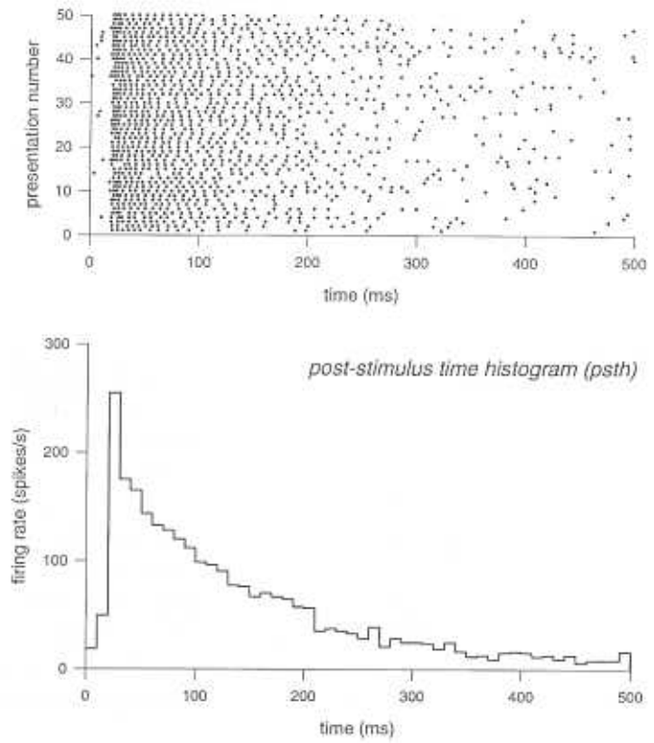



**Figure 2.1**
Variability of neural responses and construction of the average response. The top panel shows a raster plot of 50 individual spike trains in response to a stimulus at $t = 0$. Each dot in the raster plot marks the time of occurrence of a single spike. In this case, spikes are recorded extracellularly from the movement sensitive neuron H1 in the fly visual system, as in figure 1.2. The visual pattern seen by the fly makes a step motion at $t = 0$, creating a brief impulse of nonzero angular velocity. We see that the spike trains in response to repeated presentations of the same stimulus are not identical. A count of the average number of spikes in each bin (10 ms in this case) following stimulus presentation, and normalization to the number of presentations and the bin size, produces the post-stimulus time histogram, or psth, shown in the bottom panel. Normalized in this way, the psth gives the firing rate—or probability per unit time of firing, $r(t)$—as a function of time. The delay before the peak in the firing rate is due to delays in the visual receptors and in the synapses between the receptors and H1.



### 2.1.1 Probabilistic responses and Bayes' rule

Understanding the neural code means understanding the relationship between spike trains and real events in the sensory world. We would like to have this understanding in the form of a guide to the homunculus—a set of rules that gives meaning to the spike trains in much the same way that a bilingual dictionary gives meaning to the words of a foreign language. One possibility is that each distinct event in the sensory world triggers a unique spike train response, and conversely every spike train represents a unique event in the world. In practice one does not find this sort of uniqueness, because repeated presentations of the same stimulus lead to different spike trains, as shown in Fig. 2.1. The dictionary for understanding the neural code thus cannot consist of a simple list that gives a one-to-one mapping of spike trains into sensory stimuli.

Instead of a one-to-one mapping, each sensory stimulus is assigned, apparently at random, to one of many possible spike trains. Describing the neural response and building our dictionary requires that we quantify the extent of randomness. More generally our dictionary must be written in a language that goes beyond a simple list of correspondences. The appropriate language is provided by probability theory, and probabilistic ideas provide the unifying theme for all of our subsequent discussion.

In experiments like those of Fig. 2.1, the experimenter chooses some particular time dependent sensory stimulus, which we call $s(t)$, and then examines the spike trains produced in response to repeated presentations of this stimulus. Since there is no unique response, the most we can say is that there is some probability of observing each of the different possible responses. This is a conditional probability distribution, because we are talking about the probability of observing a particular spike train *given* that we present some stimulus $s(t)$. We can describe the spike train in terms of the arrival times of each spike, $t_1, t_2, \cdots, t_N$, and we will abbreviate this list of times as $\{t_i\}$. Then our notation for the conditional probability of the spike train given the stimulus is $P[\{t_i\}|s(t)]$.

In formulating the problem of the homunculus, we emphasized that the real world does not consist of just a few alternative stimuli. On the contrary, stimuli are chosen at random from an infinite set of possibilities, although these random signals have a (perhaps complex) correlation structure. To make this idea precise, we say that signals are chosen from some probability distribution, which we write as $P[s(t)]$. The actual functional form of this distribution embodies all of the structure in the world, such as the persistence of sensory qualities over time and the smoothness of motion. To refer to this structure

we shall say that the probability distribution $P[s(t)]$ defines an *ensemble of signals.*

If signals are chosen at random and the neuron has an element of randomness in its response, then the most complete description of the neuron in the sensory world would be to give the *joint* distribution of signals and spike trains, $P[\{t_i\}, s(t)]$. This distribution measures the likelihood that, in the course of an experiment or in the life of the animal, we will observe both the stimulus $s(t)$ and the spike train $\{t_i\}$. In the usual picture of stimulus and response, stimuli are chosen from $P[s(t)]$ and presented to the neuron, which responds with spikes at times $t_1, t_2, \cdots, t_N$ drawn from the conditional distribution $P[\{t_i\}|s(t)]$. This picture corresponds to the mathematical decomposition of the joint distribution into the conditional distribution multiplied by the prior distribution for the stimuli,

$$P[\{t_i\}, s(t)] = P[\{t_i\}|s(t)] \times P[s(t)]. \tag{2.1}$$

The distribution of stimuli $P[s(t)]$ is called the *prior distribution* because it embodies our knowledge, prior to any observations of the spike train, that signals will be chosen in accord with a certain statistical structure.

Although Eq. (2.1) captures our intuition about the cell responding to the stimulus, it does not represent the only point of view on the system. In particular, it is not the point of view appropriate for our homunculus. The homunculus sees one example of the spike train $\{t_i\}$ and must say something about the stimulus $s(t)$. From his point of view, it is the spike train that has been chosen at random from a distribution we shall call $P[\{t_i\}]$. Since there is no unique stimulus that can be placed in correspondence with this spike train, the most the homunculus can tell us is that some stimuli are more likely than others, given the observed spike train. But this statement can be quantified by another conditional distribution, the distribution of signals given the spike train, $P[s(t)|\{t_i\}]$. Just as the prior distribution $P[s(t)]$ defines the ensemble of signals, the conditional distribution $P[s(t)|\{t_i\}]$ defines the *response–conditional ensemble,* which we shall discuss in more detail in section 2.2.3. In the same way that we can think about stimuli generating spike trains, giving us Eq. (2.1), we can think about spike trains leading to inferences about stimuli, giving us

$$P[\{t_i\}, s(t)] = P[s(t)|\{t_i\}] \times P[\{t_i\}]. \tag{2.2}$$

When we pick up a bilingual dictionary we notice that there are two parts. The first, for example, translates from Dutch to English, the second from English to Dutch. The two languages can be brought into correspondence, but this can be done from two points of view. For signals and spike trains



this correspondence has a probabilistic element, but again there must be two equivalent points of view. The "spike speaker" needs to translate into stimuli, and looks up the distribution $P[s(t)|\{t_i\}]$, while the "stimulus speaker" needs to translate into spikes and so looks up $P[\{t_i\}|s(t)]$. As in the dictionary, we can make a list of symbols in the two "languages" and make correspondences, but correspondences are really two-headed arrows. In the probabilistic context this bidirectionality of translation is the statement that Eq's. (2.1) and (2.2) are both decompositions of the same joint probability distribution. Hence the two decompositions must be related to one another, and the two conditional distributions are therefore also related:

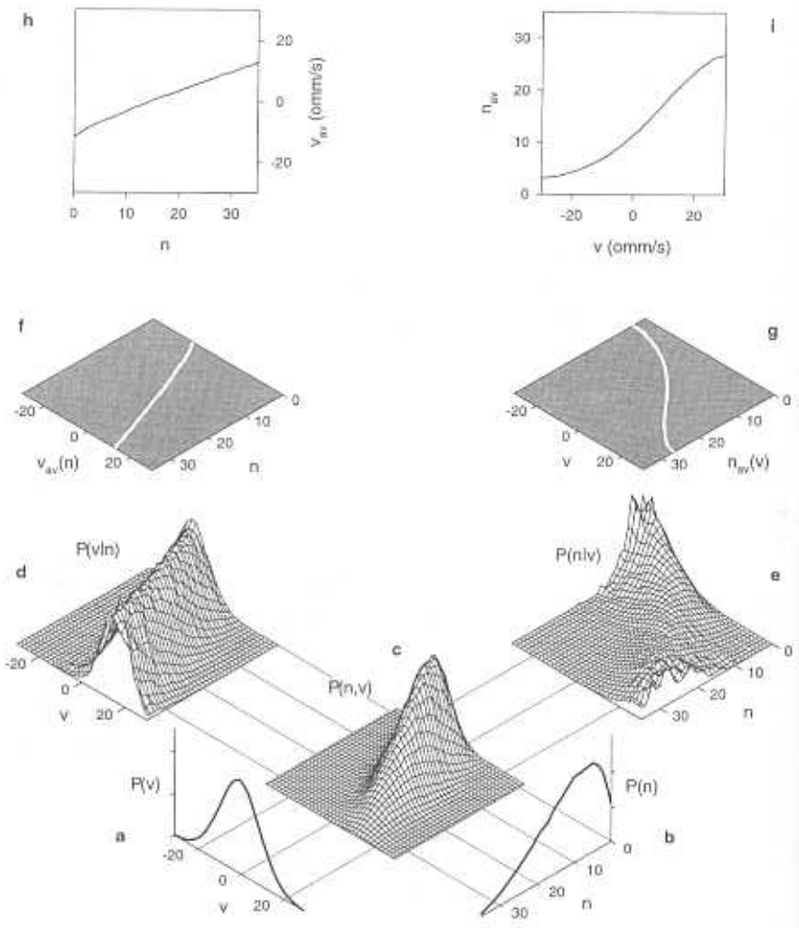
$$P[s(t)|\{t_i\}] \times P[\{t_i\}] = P[\{t_i\}|s(t)] \times P[s(t)] \tag{2.3}$$

$$\Rightarrow P[s(t)|\{t_i\}] = P[\{t_i\}|s(t)] \times \frac{P[s(t)]}{P[\{t_i\}]}. \tag{2.4}$$

This last relation is called Bayes' rule, and it will play a crucial role in our thinking about the neural code.[1]

In Fig. 2.2 we illustrate the decomposition of probability distributions represented by Eq. (2.4). Our example is taken from experiments on the motion sensitive neuron H1 in the fly's brain, a system to which we return frequently (see section 2.2.3 for a more systematic introduction). To simplify matters, we collapse the full spike train $\{t_i\}$ down to the spike count, $n$, in a 200 ms window, and we summarize the stimulus $s(t)$ by the average angular velocity of motion across the fly's visual field, $v$, in a corresponding window. The joint

1. Certain words and phrases inspire passionate responses (Carlin 1978), even from scientists. Bayes' rule provides an example of this phenomenon. In modern language, the mathematics of Bayes' rule is elementary—it follows from the definition of conditional probability distributions. Nonetheless, mention of Bayes' rule can still trigger heated discussion, with partisans displaying a nearly religious zeal. For a review of the history and current status of the controversies see Earman (1992). The problem (we think) lies with the claim that all prior expectations about the world can be encapsulated in a probability distribution. To give an example from the history of physics (Weinberg 1983), Pauli postulated the existence of an elementary particle that we now call the neutrino, a particle that would be very difficult to detect, as a way of explaining the apparent non–conservation of energy in radioactive decay processes. He viewed the neutrino as a distasteful hypothesis, but the idea that energy could be created or destroyed seemed even more objectionable. Clearly the "strength" of his belief in conservation of energy exceeded his concerns about nearly unobservable particles. But would he have been willing to state a probability that energy is not conserved? In the present discussion we try to be quite explicit in saying that sensory signals are drawn from a probability distribution, so there is, in principle, no ambiguity. In laboratory experiments this "in principle" is translated into practice, since we as experimenters choose the stimuli and we can design the experiment so that stimuli are in fact chosen at random from a distribution $P[s(t)]$ that we have constructed. In a natural setting there is a deeper question: Is there a well defined probability distribution from which natural signals are drawn? Some efforts at specifying this distribution are collected in section 5.2.




probability distribution is then $P(n, v)$, from which we can construct the conditional distribution of spike counts given the stimulus velocity, $P(n|v)$, or the conditional distribution of the stimulus velocity given that we have observed a particular spike count, $P(v|n)$. The remaining factors in Eq. (2.4) are the marginal distributions for the stimulus, $P(v)$, and for the spike count, $P(n)$. The distribution of stimuli is determined by the experimenter, while the distribution of spike counts is a property of the neuron and reflects the dynamic range that the cell can use to represent the stimulus.



**Figure 2.2**
Illustration of Bayes' rule applied to experimental data obtained from an experiment on a motion sensitive cell (H1) in the blowfly. The fly viewed a spatial pattern displayed on an oscilloscope screen, and this pattern moved randomly, diffusing across the screen. At the same time, spikes from H1 were recorded. The figure depicts the statistical relations between a stimulus variable, $v$, and a spike count, $n$. $v$ is the value of the stimulus velocity averaged over a 200 ms time window. It is measured in units of the fly's photoreceptor spacing (ommatidia) per second, or omm/s. One ommatidial distance is about 1.3° of visual angle. $n$ is the number of spikes counted in a 200 ms window, delayed 20 ms with respect to the stimulus-averaging window. The choice of these variables is made here for the purpose of illustration, not because we think that the fly uses these categories—flying at a speed of ~ 1 m/s, the fly would surely crash if it averaged for 200 ms before making a decision. (a) Probability density $P(v)$ for all the 200 ms windows in the experiment, taken in an overlapping way at 2 ms increments. (b) Probability $P(n)$ of finding $n$ spikes in a 200 ms window, computed in the same way. (c) Joint probability density $P(n, v)$ for $n$ and $v$; $P(v)$ and $P(n)$ are the two marginal distributions of $P(n, v)$. As can be seen, $P(n, v) \neq P(n) \cdot P(v)$, which means that there is indeed a correlation between stimulus and response. We can look at this correlation in two ways, either forward or reverse. The reverse description is summarized in $P(v|n)$ shown in (d). This is a family of distributions of $v$, parameterized by the observed response $n$. In other words, for each $n$ we have a different distribution of $v$, and if we know that the count in a certain window is $n = n_0$, the distribution of velocities that could have given rise to that count is given by the slice $P(v|n_0)$ out of the family of conditional distributions $P(v|n)$. In the forward description we ask what values of $n$ could be induced by a given value of $v$. This is described by the conditional distribution $P(n|v)$ shown in (e). The white lines in panels (f) and (g) show the average values of $v$ given $n$, and of $n$ given $v$ respectively. These data are replotted in a standard orientation in (h) and (i). The average value $v_{av}(n)$ in (f) and (h) gives the best estimate of the stimulus given that a response $n$ is observed (see section A.7); this is akin to the problem an observer of the spike train must solve. The average $n_{av}(v)$ in (g) and (i) gives the average response as a function of the stimulus, corresponding to the forward description. As explained in the text, the reverse estimator can be quite linear, even when the forward description is clearly nonlinear.

Figure 2.1 emphasizes that the response to a fixed stimulus has an element of randomness; by showing individual spike trains we are showing samples drawn from the conditional distribution $P[\{t_i\}|s(t)]$. To emphasize the bidirectionality inherent in Bayes' rule we would like to show a similar figure, but with a fixed spike train and different samples of the stimuli drawn from $P[s(t)|\{t_i\}]$. This is a bit complicated, because in principle we should fix a long list of spike arrival times $t_1, t_2, \ldots, t_N$. We will return to this problem later in the text, but for now suppose that we fix just one spike arrival time and look at the stimulus at times surrounding this arrival time, as in Fig. 2.3. We see that, with the one spike arrival time fixed, the stimulus fluctuates from

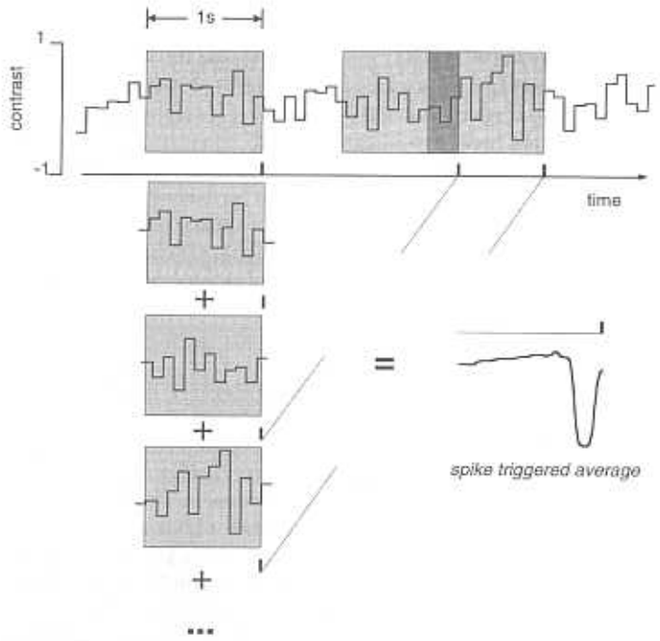



**Figure 2.3**
Construction of spike-triggered average. The top part of the figure shows a section of the stimulus, in this case a time sequence of light intensity incident on a salamander retina. Below the stimulus is the resulting spike response in a retinal ganglion cell. The spike-triggered average is constructed by averaging the stimulus waveform preceding each spike. The time course of the spike-triggered average is shown at the bottom right. On average, a spike from this cell is preceded by a transient decrease in the stimulus intensity—thus in the terminology introduced in Fig. 1.6 this is an "off" ganglion cell. From experiments by Warland and Meister (1995).

spike to spike, in complete analogy to the fluctuations of the spike train from presentation to presentation in Fig. 2.1. Nonetheless, the stimulus surrounding a spike has a nonzero average, and we shall see in section 2.1.3 that this average stimulus waveform provides a useful description of the cell's response properties. Figures 2.1 and 2.3 are just the beginning of a quantitative analysis, but we hope that they provide some intuition for the problem of translation and the importance of Bayes' rule. The bilingual dictionary of spikes and sensory signals must be written in a probabilistic format, and Bayes' rule tells us how the two halves of the dictionary are related.



An important point about bilingual dictionaries is that, armed with one half we can always produce the other half, if each is truly complete. Thus we could make a list of all the English words in the Dutch–to–English dictionary, and from the listed meanings of the Dutch words we ought (in principle) to be able to find corresponding Dutch words for each English word on the list, thereby constructing the English–to–Dutch dictionary. Bayes' rule tells us that this process works for probability distributions as well.

In addition to the abstract picture of bilingual dictionaries, Bayes' rule is telling us something very practical: We can characterize the neural code either by listing the rules for translating stimuli into spikes ($P[\{t_i\}|s(t)]$) or by listing the rules for translating spikes back into stimuli ($P[s(t)|\{t_i\}]$). If we can give a *complete* listing of either set of rules, then we can solve any translation problem. Roughly speaking, the traditional approach to the study of neural coding has been to fix the stimulus and examine the response of the neuron, then present another stimulus, and so on. In this way one works toward the characterization of the conditional distribution $P[\{t_i\}|s(t)]$, and we will make precise the way in which different methods of analysis tell us about different aspects of this distribution. If we could complete the task and really understand the full structure of the distribution $P[\{t_i\}|s(t)]$, then by Bayes' rule we could construct $P[s(t)|\{t_i\}]$ and hence give the rules our homunculus must follow as he attempts to interpret the spike train.

On the other hand, we can abandon the traditional approach and design experiments that characterize directly the distribution $P[s(t)|\{t_i\}]$, taking the point of view of the homunculus from the outset. Again, Bayes' rule tells us that if we can give a complete characterization then it doesn't matter which point of view we use. But the language example gives us the hint that a "complete dictionary" may be very difficult to produce. It is easy to imagine that for some neurons one point of view will be simpler than the other, and then we will be happy to take the simpler point of view. We will also want to know *why* one point of view is simpler than the other, and whether this fact is telling us something about the structure of the neural code. These ideas are developed in the rest of this chapter.

Before proceeding, let us look back at Fig. 2.2. We can see already that the two points of view we have described do indeed differ in complexity. The conventional point of view—holding the stimulus fixed and examining the probability distribution of responses—leads to the conditional distribution $P(n|v)$, which has a somewhat complex structure. At large negative velocities, for example, the most likely value of $n$ is $n = 0$, and it is only above some critical velocity that the distribution breaks away from $n = 0$ and forms a clear

peak at nonzero spike count. In contrast, the distribution of stimuli given the spike count, $P(v|n)$, seems to be a simple, nearly Gaussian, peak whose mean moves across the range of velocities as the number of spikes is varied. These differences are made clearer by actually computing the means in the two distributions. For the distribution $P(n|v)$, this mean is the average number of spikes produced as a function of the stimulus amplitude $[n_{\rm av}(v)]$, and this is the traditional measure of neural response. We see the familiar nonlinear, sigmoidal relationship observed by Adrian that has been reproduced in many systems. But when we ask for the mean of the distribution $P(v|n)$, which is the average stimulus velocity given that we observe a particular spike count, we see a very different, almost perfectly linear relation. The nonlinearity of the sigmoidal input/output relation, so ubiquitous in neurobiology, seems to have vanished. We are not yet ready to discuss the origins of this simplification, or even to understand why the average stimulus given the spike count is such an interesting quantity. But it should be clear that by changing our point of view—by translating from spikes back into stimuli rather than the conventional translation from stimulus to spikes—we have the chance of simplifying our description of the code.

### 2.1.2 Rates, intervals, and correlations

Given a particular time-dependent stimulus $s(t)$, a complete probabilistic description of the neural response is contained in the conditional distribution $P[\{t_i\}|s(t)]$, which measures the relative likelihood that spikes will arrive at the set of times $\{t_1, t_2, \cdots, t_N\}$. But, as we hinted above, the words *complete description* are a bit dangerous. No finite amount of data is ever sufficient to determine completely a probability distribution. In practice, experiments usually aim at characterizing the first few moments of a distribution—the mean, the variance, and so on. We need to see how these quantities can be defined and measured for the distribution of spike trains. The hope is that some structure in the moments will catch our attention.

The first step in trying to characterize a probability distribution is to measure its mean. In a sense we make precise in section A.1, the mean of the conditional distribution $P[\{t_i\}|s(t)]$ is a quantity called the *time dependent firing rate*, $r(t)$. The spike train itself is a very singular function of time, consisting of pulses at the times $t_i$. If we imagine sitting at one point in time, $t$, and then averaging the spike train over many presentations of the same stimulus, we arrive at the procedure shown in Fig. 2.1. When averaging, we should really count the number of times that a spike arrives exactly at the time $t$, but this never happens. Instead we count the spikes in a window of size $\Delta\tau$ centered



at $t$, and then divide by the number of presentations. In this way we measure the probability $p(t)$ that a spike occurs in our small window. If we make the window larger the probability will be larger, and we would like to characterize the response in a way that does not depend on our arbitrary choice of a window size. The way to do this is to take the smallest possible windows (if the windows are too small, a finite data set will not give reliable results) and recognize that as the window size becomes very small the probability of finding a spike must be proportional to the window size, so that $p(t) = r(t)\Delta\tau$. This defines the rate $r(t)$ as the probability per unit time that a spike will occur in a small window surrounding the time $t$. This function of time, illustrated in Fig. 2.1, is also called the post- or peristimulus time histogram.

The spike rate defined in Adrian's experiments was equal to the number of spikes in a rather large time window following the onset of the stimulus. This rate can be measured from just one example of the spike train, and it might be better to focus on the fact that it is a count of the number of spikes that occurred in that particular example. On the other hand, the time dependent rate $r(t)$ is a continuous function that determines the probability of spike occurrence at different times. Thus the time dependent rate is a property of an *ensemble* of spike trains—as is clear from its construction in Fig. 2.1—and is *not* knowable from observation of a single example of the neural response. This distinction between counting spikes and knowing the rate as a function of time will be crucial in the following sections.

The description of neurons as using a "rate code" presupposes that we all agree on the meaning of "rate." But we have seen that there are least two meanings—Adrian's original definition in terms of spike counts, and the time dependent quantity that measures the probability of spike occurrence. Implicitly, a description in terms of a "rate code" also assumes that we can state (and exclude) the alternative, usually called a "timing code." In the following paragraphs we allow ourselves to be drawn into a superficial paradox by losing track of the distinction between the two meanings of "rate," and we drag the poor reader along with us. We hope that this process, though a bit roundabout, brings us (and the reader!) one step closer to the crucial issues. At the end, the distinction between rate and timing codes will be less clear than one might have hoped, but we think that this is the right answer. Rather than trying to sharpen the rate/timing distinction, we will argue that the interesting question is whether sensory neurons produce large numbers of spikes or small numbers of spikes in the time windows relevant for behavior and decision making. This formulation of the problem, however, will remain in the background until section 2.2.1.

Perhaps the best studied examples of time dependent firing rates are in cases where the stimulus $s(t)$ is periodic in time. In the mechanical sensors of the inner ear, the analysis of time dependent firing rates is crucial to our current understanding of the system. If one counts the spikes produced by an auditory neuron in response to a "pure tone" (sinusoidal sound pressure variations at the eardrum), there appears to be a threshold sound intensity required to raise the spike count above its spontaneous value. But, at least for low frequency sound, this appearance is quite misleading. The time dependent firing rate in response to relatively quiet low frequency sine waves shows a nearly perfect sinusoidal modulation around the spontaneous rate, as illustrated in Fig. 2.4. Under these conditions, the time dependent rate $r(t)$ is

$$r(t) = r_0 + A\sin(\omega t + \phi), \tag{2.5}$$

where $r_0$ is the spontaneous rate, $A$ is the amplitude of the modulation, $\omega$ is the frequency of the sine wave stimulus, and $\phi$ is the preferred phase of firing relative to the stimulus. The modulation $A$ is proportional to the sound pressure. Thus the *probability* of spiking varies smoothly with the input signal—arbitrarily small signals produce proportionately small responses. The sensitivity of this rate modulation is astounding: In one species of frog, the vibration of the entire frog by one-tenth of an Ångstrom produces a modulation $A \sim 10$ spikes per second in neurons from the sacculus (Narins and Lewis 1984).

Periodic modulation of the firing rate by sinusoidal stimuli becomes even clearer at higher amplitudes. This behavior is often termed *phase locking* (Rose et al. 1967). Although the variations of the time dependent rate are locked to the cycles of the sine wave stimulus, the average firing rate can be much less than the frequency, so that spikes occur in only a small fraction of the cycles. It appears that the pattern of firing and skipping is random, but variations in the *probability* of firing are locked to the sine wave. Thus the timing of individual spikes carries information about the phase of the sine wave, and the times between spikes tend to cluster around integer multiples of the sine wave period.

At low frequencies, the wavelength of sound is long and our heads do not cast a significant acoustic shadow. Thus the intensity of sounds at our two ears is always the same for low frequency sounds, and the only clue about sound source location comes from timing—the sound waves arrive earlier at the closer ear. At high frequencies these time differences become ambiguous, but shadowing effects take over and intensity differences between the ears become significant. All of this was understood by Lord Rayleigh (Strutt 1877–78) in the late nineteenth century. In a series of experiments (with Lady



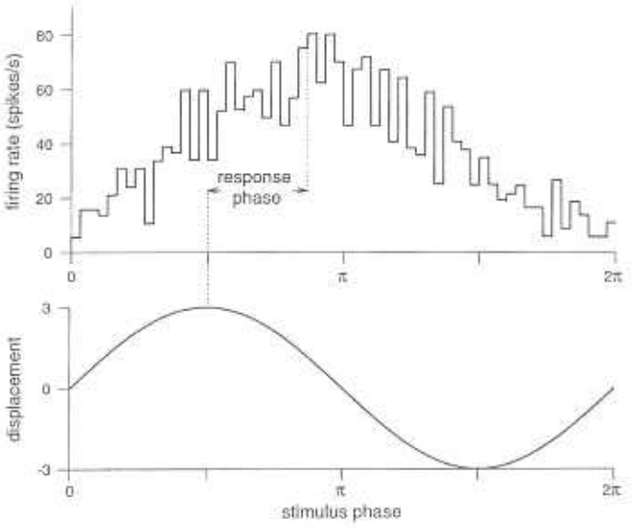



**Figure 2.4**
Phase locking in *Xenopus* lateral line receptors. Spike times are recorded in response to a maintained 5 Hz sinusoidal stimulus, in this case the vibration of a sphere in the water some distance away from the receptor (bottom panel). The arrival time of each spike can be registered as an absolute time from the onset of the experiment or, alternatively, as a phase relative to the stimulus sinusoid. One then constructs a phase histogram by analogy with the post-stimulus time histogram in Fig. 2.1, shown in the top panel. Redrawn from Kroese, van der Zalm, and van den Bercken (1978).

Rayleigh as the subject and the gazebo as the laboratory) he showed conclusively that we can hear phase differences between our ears, at least for low frequency sounds. These phase differences correspond to time differences of less than ten microseconds. Barn owls, which are especially good at localizing prey by acoustic cues alone, have time difference thresholds as low as one microsecond. There is no question that the temporal information essential to these discrimination tasks is carried in the phase locking of the auditory nerve, and in the case of the barn owl it has been possible to identify the neural circuits responsible for making the precise temporal comparison between phase locked spikes coming from the two ears. For a review of this work, see Carr and Konishi (1990).

Let us contrast the observation of phase locking with the classical notion of rate coding. In Adrian's original work, rate was defined by counting the spikes in a fixed time window following stimulus onset. The real content of the claim

that information is carried by the firing rate (as opposed to timing) must be that the precise temporal locations of the spikes within the window are *not* informative about the stimulus parameters. Clearly the spikes in primary auditory neurons do provide information about the stimulus by virtue of their precise temporal pattern. Sound localization at low frequency demonstrates that the brain can use this temporal information, ultimately down to the microsecond scale.

What we have just said is paradoxical (but we warned you!): We introduced phase locking as an illustration of time dependent firing rates, yet now we claim that this observation is inconsistent with the idea of rate coding. The origin of the paradox is in the dual definitions for "rate," and in particular the connection of these definitions to the time resolution with which we observe the spike train. When we construct a plot of rate vs. time from a post stimulus time histogram such as Fig. 2.1, the rate is a probability defined (in principle) in infinitesimally small time bins. In contrast, Adrian chose more macroscopic time windows, of order the total duration of the stimulus. In the case of phase locking to sine wave stimuli, we keep the essence of the timing cues so long as we use time bins that are significantly smaller than the period of the sine wave. Thus, if we are listening to tones at 100 Hz, counting spikes in bins of 2 ms is sufficient to reveal the periodicity of the firing, and even bins of 5 ms are enough to get a coarse measure of the phase variables that must be compared in sound localization, although the intrinsic precision of the neuron may be much better than this.

If we count spikes in 5 ms bins, are we measuring rates with small windows, or should we call this a timing code? This is another way of looking at the problem of defining firing rate, as mentioned earlier in this section: Do we use Adrian's method of counting spikes, or when we say *rate* do we mean the time dependent function described in the post-stimulus time histogram of Fig. 2.1? Lest we think that the auditory system is a special case, we shall see this problem come up in several other systems (section 2.2.1). Although our discussion here may seem a little confusing, we offer it as a first hint that the usual distinctions between rate and timing are not the essential distinctions required for understanding the neural code. For now, let us put aside the interpretation of neural responses in terms of rate or timing codes and return to the problem of characterizing the responses themselves.

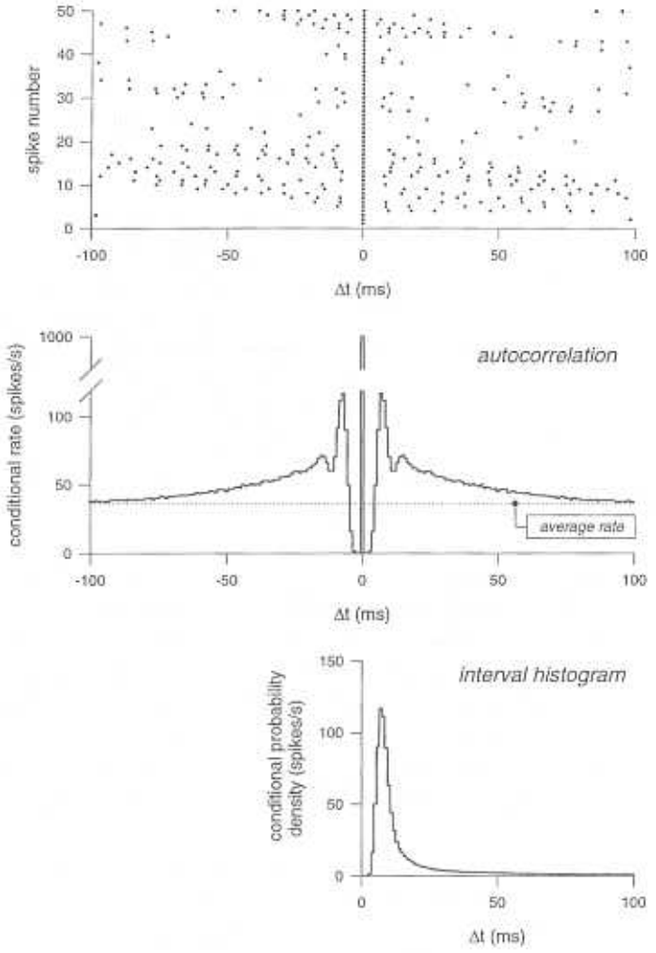
Having looked at the mean of the conditional distribution $P[\{t_i\}|s(\tau)]$, it is natural to ask about the analog of the variance. In the same way that the rate, which measures the probability of occurrence of one spike, is a first moment of the distribution, the natural formal definitions for the variances and covariances of the distribution $P[\{t_i\}|s(\tau)]$ are related to the joint probabilities of occurrence for two spikes (Fig. 2.5). There are many different ways of looking for two spikes. In one simple case, the two spikes occur in succession, with no spikes in between; in this case we are talking about the probability distribution for interspike intervals. In the other simple case, we ask about the joint probability for the occurrence of two spikes without regard to what happened in between; this is called the *correlation function*. The correlation function is often normalized by the firing rate so that it measures the probability of observing a spike at time $t + \tau$ given that a spike was observed at time $t$. We refer to this normalized correlation function as the *conditional rate*, as in Fig. 2.5. Notice that there are many other possibilities, such as the distribution of times between two spikes separated by exactly seven spikes, but these don't have much intuitive appeal. In several sensory systems one can find neurons with comparable firing rates but very different second-order statistics. In the vestibular system (Goldberg and Fernandez 1971), for example, such differences in statistics are evident even in the spontaneous activity of the primary sensory neurons.



The interspike interval distribution quantifies the probability that successive spikes will be separated by a particular interval in the same way that the time dependent rate quantifies the probability that a spike will occur at a given moment in time. Similarly, the correlation function quantifies the probability that two spikes will occur with a certain separation independent of what happens in between. All of these quantities are probabilities, so they are properties of an ensemble of spike trains and, again, are not accessible from observations on a single spike train. For more details about correlation functions, see section A.2.

One obvious question is whether cells in the central nervous system are sensitive to the higher order statistics of incoming spike trains. Thirty years ago Segundo et al. (1963) asked this specific question in a series of experiments on *Aplysia*. In an ideal experiment one would control all of the spike trains which provide synaptic input to a single cell, studying how changes in the statistics of these inputs influence the postsynaptic response. In practice this complete control is very difficult, and Segundo et al. emphasize that their experiment falls short of the ideal, but the results, exemplified by Fig. 2.6, seem clear: Different temporal patterns of spikes in the presynaptic nerve fibers result in very different postsynaptic responses.

Rather similar in spirit to the experiment of Segundo et al. is the recent work by Mainen and Sejnowski (1995). They study the behavior of cortical neurons in a slice preparation, where one can effectively eliminate synaptic couplings




using various pharmacological tricks. Mainen and Sejnowski then study the spike trains of an isolated cortical neuron in response to injected currents. The injected current is a controlled version of the myriad synaptic inputs that the cell would experience in vivo. By analogy with the *Aplysia* experiments, we would like to know if the cortical neuron can give reproducible responses to the temporal details of injected current waveforms, or if all that counts is the time average current. The result is clear: Cortical neurons can produce spikes in a rather deterministic relation to temporal features on the millisecond time



**Figure 2.5**
Construction of the autocorrelation function and the interval histogram. The top panel shows a spike sequence measured from the H1 cell in the fly visual system. The time of occurrence of each spike is represented by a single dot. The spike sequence represented by the lower horizontal row of dots is shifted repeatedly to the left to align each successive spike with $\Delta t = 0$. Thus the spike sequence represented by the second horizontal row of dots is simply the first shifted one spike to the left, the third row is shifted two spikes, and so on. In the middle panel the occurrence times of the spikes in each shifted sequence are averaged and normalized to a rate. This is the autocorrelation function, which can be thought of as the firing rate at time $t + \Delta t$ given the occurrence of a spike at time $t$. For large values of $\Delta t$ this conditional rate approaches the average rate, indicating that the memory that there was a spike at time $t$ is lost. A similar procedure can be followed to measure the probability of finding a particular interval between two successive spikes—the interspike interval distribution (bottom panel). In this case, rather than averaging all the spike occurrence times, we average only the times of the first spike after $\Delta t = 0$ in each shifted sequence. This interval distribution is normalized to a probability density.

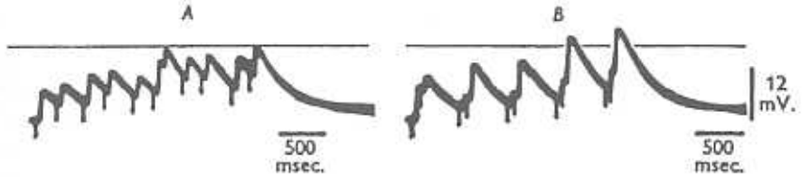



**Figure 2.6**
Dependence of postsynaptic response on temporal pattern of inputs redrawn from Segundo et al. (1963). Panels A and B show responses in the same cell to two different patterns of presynaptic pulses with the same mean frequency but different second order statistics. The difference in the threshold-crossings in the responses to these stimuli demonstrates that the temporal pattern of the presynaptic signal is important in determining the postsynaptic response.

scale, so that correlations in the synaptic input spike trains will change the cell's output. Indeed, most experiments on neurons in cortical slices rely on the fact that simple patterns of current injection, such as steps or pulses, will generate reproducible patterns of spikes.

It is worth remembering that the original analysis of action potential dynamics was a *deterministic* description—Hodgkin and Huxley (1952d) presented equations of motion for the voltage and conductances of the cell membrane, and these equations can be solved to predict the timing of action potentials in relation to the pattern of injected currents. Generations of experiments on

neurons in cortical slices have been fit to generalizations of the Hodgkin–Huxley equations, implicitly verifying the determinism of spike generation in response to simple patterns of current injection (Gutnick and Mody 1995). Reproducible spike generation is by no means limited to giant axons and cortical pyramidal cells, as can be seen from recent experiments on cells from the vestibular nuclei (du Lac and Lisberger 1995). Indeed, the now classic experiments on noise in the nerve cell membrane involved heroic efforts to reveal the tiny window of stimulus amplitudes in which spike generation is noticeably probabilistic (Verveen 1961; reviewed by DeFelice 1981).

Sensitivity to temporal patterns is determined both by the integration time and by the noise level of the cell. If we imagine that a neuron fires a spike whenever the total input reaches a threshold, we know from the classical mathematical literature (Rice 1944–45) that even the average rate of threshold crossings depends on the temporal correlation in the input signal, but if the cell's own electrical properties integrate the inputs over a very long time scale, this integration time (and not the correlations in the input) will determine the rate. Similarly, if the cell has an internal noise source which mixes with the input signal, precise relations between the temporal features of the input and the output spikes will be randomized. The Mainen–Sejnowski and related experiments suggest that the spike generating mechanism itself has a relatively short integration time and low noise level, at least under one set of conditions. Lass and Abeles (1975) showed that propagation of the action potential is also relatively noise free, with a ten centimeter length of myelinated axon introducing only a few microseconds of jitter in the arrival time of a spike. What remains as significant noise source is synaptic transmission (Katz 1966), and a number of recent experiments have emphasized that transmission across central synapses is surprisingly prone to failure (Allen and Stevens 1994; Bekkers and Stevens 1994), but when transmission occurs it seems to introduce very little temporal jitter. All these experiments suggest that the elements of neural signal transmission and computation are capable of preserving precise temporal relationships. In a long series of experiments, Abeles and coworkers have drawn attention to the behavior of cells in frontal cortex, where specific patterns of spikes separated by hundreds of milliseconds can recur with millisecond accuracy (see, for example, Abeles et al. 1993), and many investigators have noted that spikes produced in response to the onset of a sensory stimulus can be extremely reproducible, as illustrated by the data from bat auditory cortex in Fig. 2.10 (Dear, Simmons, and Fritz 1993). But we still do not have a complete, quantitative answer to the question posed by Segundo et al. more than thirty years ago—how precisely can a postsynaptic cell measure the arrival times of incoming spikes?



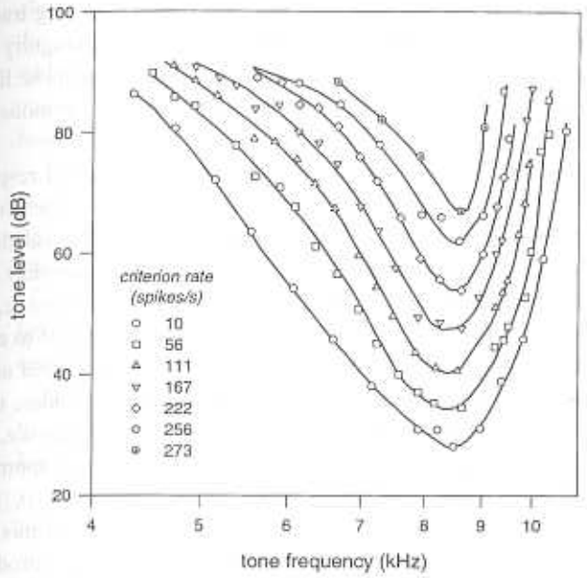



**Figure 2.7**
Iso-rate contours for a cell in the cat auditory nerve. Each curve describes the combinations of amplitude and frequency of a pure tone that give rise to a certain mean firing rate. Auditory neurons are "tuned" to a particular range of frequencies; as a result, the amplitude required to produce a given firing rate has a minimum, at about 8.5 kHz in this cell. The frequency tuning of auditory neurons means that changes in the amplitude or frequency of a tone can produce similar changes in the mean firing rate; thus we can move from one contour line of fixed firing rate to another contour either by changing the frequency and moving horizontally or by changing the amplitude and moving vertically. Redrawn from Evans (1982).

To illustrate the possibilities of coding with second-order statistics, consider again an auditory neuron stimulated at low frequencies. The number of spikes in response to a tone burst provides no information about phase, and there is a confusion between amplitude and frequency because loud sounds away from the peak of the cell's frequency sensitivity produce the same rate as quiet sounds at the best frequency, as illustrated in Fig. 2.7. But, at low frequencies, there is a tendency for interspike intervals to cluster around integer multiples of the stimulus period (Kiang et al. 1965), and this is related to the time dependent firing rates shown in Fig. 2.4. This clustering yields an independent estimate of the frequency, and the amplitude of the sound can then be

estimated unambiguously from the firing rate. Thus by keeping track of spike arrival times, one can resolve the amplitude/frequency ambiguity that arises from counting spikes in a large time window. Ideas along these lines can be traced back at least to Wever's (1949) early work on synchronous activity in the auditory nerve.

In this section we have seen the characterization of neural responses progress from the counting of spikes to the measurement of ensemble average, time dependent rates, and finally to the description of interval distributions and correlation functions. In each case, changes in the parameters of the sensory stimulus cause changes in our measure of the neural response, and hence these different measures of the response all have the potential to encode features of the sensory world. The conference organized by Perkel and Bullock (1968) produced a remarkable catalog of such candidate codes, drawing on experiments from a wide range of different systems. We can see, at least in outline, how these different codes reflect successively higher moments in the conditional distribution of spike trains given the stimulus, $P[\{t_i\}|s(\tau)]$. The problem is that characterization of neural responses in terms of this succession of moments does not seem to be converging—each time we introduce a new, higher moment some new coding strategy becomes possible.

### 2.1.3 Input/output analysis

One might hope that there exists a complete phenomenological characterization of a neuron. One could then predict the neural response to arbitrary input stimuli. In the engineering literature the search for such a characterization is sometimes called systems identification, and in the physics literature one considers a hierarchy of linear and nonlinear response functions (Pippard 1985). These methods are simplest for the case of linear systems, where the response to the sum of two stimuli is the sum of the responses to each of the stimuli in isolation. Sensory neurons are seldom linear in this sense, and a variety of nonlinear methods have been widely used, particularly the Volterra and Wiener approaches, which go under the rubric of white noise analysis or reverse correlation (see section A.3). For reviews of these methods see Sakai (1992) and Eggermont, Johannesma, and Aertsen (1983).

We want to emphasize that there are two distinct issues in the analysis of a neuron's input/output relation. First we need to construct a family of models that is rich enough to describe what neurons may do under reasonably natural conditions. Then we need techniques for measuring the parameters of these models in particular experiments. White noise analysis, or even the more



classical mapping of receptive fields in the visual system, addresses itself to the second problem—we have in mind a model of how the neuron responds, and we want to measure the parameters of this model. The parameters of our model might be collected into a first Wiener kernel, or we might list the location and dimensions of the receptive field.

This section may seem like a long digression from our main problem of characterizing the neural code. On the contrary, this section is about a set of methods for quantifying what neurons do under a wide range of stimulus conditions. When we read in the literature about the quantitative characterization of neural responses, some of the ideas discussed in this section are probably lurking at least in the background of the experiments. It is important for us to understand both the potentialities and the limitations of these approaches.

Wiener and Volterra methods are methods for measuring response functions. The idea that we can characterize the behavior of a system in terms of linear and nonlinear response functions is a generalization of the idea from calculus that we can expand a function in a power series. We emphasize once more that this *idea* of response functions is separate from the question of how we *measure* the response functions in particular experiments.

We recall that if we have a number $x$ and we take some function of this number $f(x)$, then if $f$ is sufficiently smooth we can approximate it in the neighborhood of some reference point $x_0$:

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2}f''(x_0)(x - x_0)^2 + \cdots . \tag{2.6}$$

This is the Taylor series, and $f'(x_0), f''(x_0), \cdots$ are the first, second, $\cdots$ derivatives of the function $f(x)$ evaluated at the point $x_0$. If we keep more and more terms in this series, then the series converges to the *exact* value of $f(x)$ for any $x$ in some range surrounding the reference point $x_0$, as illustrated in Fig. 2.8. Again this convergence is conditional on some notion of smoothness for the function $f(x)$, but for our purposes this is not a limitation. We can also make precise statements about how close we come to the true function if we keep only the first $N$ terms of the series. Convergence and bounding of errors means that as a model of the transformation $f(x)$, the Taylor series can be as accurate as we want. If we think about the set of all possible Taylor series, we know that somewhere in that set there is the true model for our function.

Knowing that with enough terms we can get as close as we like to the real function is perhaps not so useful. We would like to get away with keeping just the first few terms, possibly knowing that then we cannot extrapolate too far

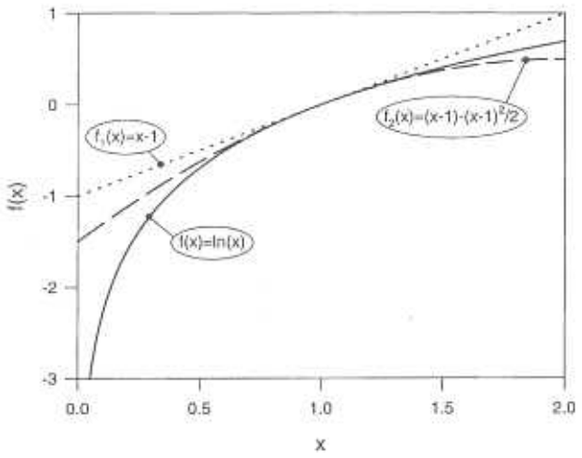



**Figure 2.8**
Taylor series approximation to $f(x) = \ln(x)$ (solid line). First order, $f_1(x) = x - 1$, and second order, $f_2(x) = (x - 1) - (x - 1)^2/2$, approximations to $\ln(x)$ in the neighborhood of $x = 1$ are shown.

from our reference point $x_0$. Thus, if we stop with the term $\sim (x - x_0)$ we are making a linear approximation, since we are approximating $f(x)$ as a straight line that passes through $x_0$ with the correct slope, as in Fig. 2.8. Notice that if we look *very* close to $x_0$ this linear approximation must be correct, again supposing that $f(x)$ is smooth. Thus we have the idea that for each function $f(x)$ there is some typical scale $\Delta x$ on which the nonlinear terms become important.

Volterra (1930) showed how one could generalize these ideas to the case where the input to our function is not just one number but a whole function itself, such as the stimulus waveform $s(t)$. These functions of functions are called *functionals*. Wiener (1958) showed how one could rearrange the Volterra series to measure more easily the coefficients in the expansion. But both from a mathematical and from a conceptual point of view, all of these ideas grow out of the more elementary notion of the Taylor series, as described by Eq. (2.6) and illustrated in Fig. 2.8. Mathematical details of the Wiener and Volterra formulations are summarized in section A.3.

To understand how these methods work, it is useful to think first about a system free from the complication of spiking. Imagine studying the dynamics of a neuron which does not generate spikes. There are still voltage dependent



conductances in the membrane, however, and we can write a set of coupled differential equations for the dynamics of the voltage and the populations of open and closed channels; these equations are of the same general form as the original Hodgkin–Huxley (1952d) equations, and they have a stable steady state solution at the resting potential $V_0$. If we inject a small amount of current, we expect that the resulting changes in voltage will also be small. In the limit that currents are very small, the voltage change should be proportional to the current, although the time dependence of the current may be different than the time dependence of the voltage. This regime of small currents is called *linear response*, and the function which characterizes the linear response is sometimes called the *transfer function*.

In the linear approximation we can write the voltage $V(t)$ in response to an injected current $I(t)$ as

$$V(t) = V_0 + \int_{-\infty}^{\infty} d\tau Z_1(\tau) I(t - \tau). \tag{2.7}$$

At this point $Z_1(t)$ is just some function which parameterizes the linear response of the voltage to current, but it will turn out that it is related to the Fourier transform of the electrical impedance (Horowitz and Hill 1980). Equation (2.7) predicts that if we inject a pulse of current at time $t = 0$, then the voltage changes are $\Delta V(t) \propto Z_1(t)$, and if we inject two pulses we just get the sum of the two voltage responses. This last statement is the definition of linear response. Equation (2.7) is somewhat complicated, because the voltage at one time is related to the currents injected at all (past) times. It will turn out that this description can be simplified by thinking about relations among Fourier components rather than relations among signals at specific times, and so we begin by taking the Fourier transform of both sides in Eq. (2.7). We define the Fourier transform of the voltage as

$$\tilde{V}(\omega) = \int_{-\infty}^{\infty} dt \exp(i\omega t)[V(t) - V_0], \tag{2.8}$$

and similarly for the current,

$$\tilde{I}(\omega) = \int_{-\infty}^{\infty} dt \exp(i\omega t) I(t). \tag{2.9}$$

Now we go through the steps of taking the Fourier transform of each side in Eq. (2.7):

$$\begin{aligned}
\tilde{V}(\omega) &= \int_{-\infty}^{\infty} dt \exp(i\omega t)[V(t) - V_0] \\
&= \int_{-\infty}^{\infty} dt \exp(i\omega t) \int_{-\infty}^{\infty} d\tau Z_1(\tau) I(t-\tau) \\
&= \int_{-\infty}^{\infty} d\tau Z_1(\tau) \int_{-\infty}^{\infty} dt \exp(i\omega t) I(t-\tau) \\
&= \int_{-\infty}^{\infty} d\tau Z_1(\tau) \int_{-\infty}^{\infty} dt \exp[i\omega(t-\tau)] I(t-\tau) \exp(i\omega\tau) \\
&= \int_{-\infty}^{\infty} d\tau Z_1(\tau) \exp(i\omega\tau) \int_{-\infty}^{\infty} dt \exp[i\omega(t-\tau)] I(t-\tau) \\
&= \left[ \int_{-\infty}^{\infty} d\tau Z_1(\tau) \exp(i\omega\tau) \right] \tilde{I}(\omega) \\
&= \tilde{Z}_1(\omega)\tilde{I}(\omega). \qquad (2.10)
\end{aligned}$$

Clearly the natural object is

$$\tilde{Z}_1(\omega) = \int_{-\infty}^{\infty} d\tau \exp(i\omega\tau) Z_1(\tau). \qquad (2.11)$$

The units of $\tilde{Z}_1(\omega)$ are voltage/current, or resistance. For a resistor, $\tilde{Z}_1(\omega) = R$, and more generally $\tilde{Z}_1(\omega)$ is called the frequency dependent impedance. For a purely passive cell membrane described by resistance $R$ and capacitance $C$, we have

$$\tilde{Z}_1(\omega) = \frac{R}{1 - i\omega\tau}, \qquad (2.12)$$

$$\tau = RC. \qquad (2.13)$$

We emphasize that the *concept* of impedance or linear response is not limited to circuits constructed out of the familiar resistors, capacitors, and inductors. Our thinking about linear response functions should be unencumbered by engineering analogies.

An exactly linear response is a bit unusual, so we expect more generally that the response of the cell voltage to small injected currents can be written as

$$\begin{aligned}
V(t) = V_0 &+ \int_{-\infty}^{\infty} d\tau Z_1(\tau) I(t-\tau) \\
&+ \frac{1}{2} \int_{-\infty}^{\infty} d\tau \int_{-\infty}^{\infty} d\tau' Z_2(\tau, \tau') I(t-\tau) I(t-\tau') + \cdots, \qquad (2.14)
\end{aligned}$$



Now $Z_2$ reflects the fact that if we inject two pulses of current they interact with each other in producing voltage changes that are not just the sum of the changes produced by each pulse alone. Similarly, if we inject sinusoidal currents at two frequencies $\omega_1$ and $\omega_2$, the term $Z_2$ measures the strength of the voltage response at the sum and difference frequencies $\omega = \omega_1 \pm \omega_2$. In our example of a cell with voltage dependent conductances, $Z_1, Z_2, \cdots$ can be related back to the activation parameters of the ion channels. But suppose that we do not have such a model, and instead use the series in Eq. (2.14) to describe the electrical dynamics of the cell from a phenomenological point of view. This is called a Volterra series, and there are theorems showing that, under certain conditions, this series provides a complete description of a system if we keep enough terms.

If Eq. (2.14) provides a complete description of the relation between (in our example) current and voltage, it is natural to ask how we could *measure* the response functions $Z_1, Z_2, \cdots$. One widely used technique is white noise analysis, which is inspired by Wiener's reformulation of the Volterra expansion (Wiener 1958). In the Wiener approach, the system is driven by random inputs and we choose an expansion of the system response so that the different terms in the expansion are statistically independent when averaged over these random inputs. This formulation has the advantage that one can measure the linear component of the system response even when the overall response is quite nonlinear. Furthermore, if we try to improve our description of the system by including more terms in the series expansion, we won't have to go back and revise our estimate of the lower order terms.

The essence of the Wiener method for analyzing nonlinear systems is in the cross correlation of input and output signals. We have seen that in the linear regime, if we inject current at one frequency we get voltage changes at that frequency. So Wiener proposes that we inject white noise currents, which contain all frequencies but with random amplitudes and phases, and look—frequency component by frequency component—for correlations between voltage and current. This correlation then measures the impedance $Z_1$.

The next step is to recognize that with all frequency components present, nonlinear terms like $Z_2$ in Eq. (2.14) produce frequency mixing, so that the voltage at frequency $\omega$ includes contributions generated by currents at *all* frequencies $\omega_1$ and $\omega_2$ such that $\omega_1 \pm \omega_2 = \omega$. So we look for correlations between $\tilde{V}(\omega)$ and the products $\tilde{I}(\omega_1)\tilde{I}(\omega_2)$, and these correlations measure the second order nonlinearities in the response. These ideas are made a bit more explicit in section A.3.

We need one more mathematical idea to use the Wiener method: *ergodicity*. The Wiener method is a reformulation of the Volterra expansion so that different terms are statistically independent when we average over the random input stimulus. Similarly, the different terms in the expansion are measured by computing a cross-correlation of input and output signals, and the cross-correlation is an average over the random inputs. How are we to understand these averages? In principle we should perform many identical experiments, so that we have an *ensemble* of inputs and outputs, and then compute the average over this ensemble. In practice, especially with random inputs, it is easier to let the experiment run continuously for a very long time. Since the stimulus is random, we can think of different time windows in this long run as being like different samples drawn from an ensemble of experiments, and when we need to average we can average over time. The statement that averages over time are equivalent to averages over the ensemble is the statement of ergodicity. For many problems we have an ergodic theorem, which tells us that the two averages are mathematically identical. Thus, when we want to cross-correlate inputs and outputs, we can do this by averaging over time in one long experiment.

How can these techniques be used to study spiking neurons? Now the input is again some continuous function of time $s(t)$ which we, as experimenters, can choose so that it looks like a sample of white noise. But the output of the neuron is not a continuous voltage; rather, it is a sequence of discrete spikes at times $\{t_i\}$. As described in section A.1, we can construct a function of time which describes these pulses,

$$\rho(t) = \sum_i \delta(t - t_i), \tag{2.15}$$

where the "delta function" $\delta(t - t_i)$ is zero unless $t = t_i$, but the peak at $t = t_i$ is infinitely high so that the integral of the function is one; then the integral of $\rho(t)$ over a time window counts the number of spikes in that window. But, unlike our idealized example of current and voltage, $\rho(t)$ is a random function not completely determined by the input $s(t)$. Suppose that we proceed naively, and cross-correlate the output $\rho(t)$ with the input $s(t)$. As explained in section A.3, this cross–correlation is equivalent to computing the average stimulus waveform surrounding a spike, which is the procedure shown in Fig. 2.3. This cross-correlation function has several different names in the literature—it is the "first Wiener kernel" or the "reverse correlation function," and it is also the "spike triggered average," the "mean effective stimulus," or the "triggered correlation function."



In their early work on white noise analysis of sensory neurons, de Boer and Kuyper (1968) emphasized that the first Wiener kernel is equivalent to the average stimulus that leads up to, or triggers, a spike. This suggests the interpretation that the neuron is "looking for" features of the random waveform that resemble the reverse correlation function, and that when such features are detected the cell fires an action potential. To see how this works, consider the simple case that the firing rate depends on a filtered version of the signal, so that

$$r(t) = r_0 g\left[\int_{-\infty}^{\infty} d\tau f(\tau) s(t - \tau)\right], \tag{2.16}$$

where $f(t)$ is the filter and $g[x]$ is some arbitrary memoryless nonlinear function. Then one can show that reverse correlation function, or first Wiener kernel, is proportional to the filter function $f(\tau)$. In this way reverse correlation allows linear filtering properties to be separated from the nonlinearities of spike generation, at least in the context of this simple model. The method is widely used for the measurement of tuning curves in the auditory system or receptive fields in the visual system; for examples see Eggermont, Johannesma, and Aertsen (1983), Reid and Shapley (1992), and DeAngelis, Ohzawa, and Freeman (1995).

If we deliver a signal whose shape is matched to the shape of the filter $f(t)$, specifically $s(t) = f(-t)$, then the rate $r(t)$ will be changed by an especially large amount. We can make this precise by saying that if we try all possible signals that have the same total power $\int dt s^2(t)$, then this particular signal will give the largest modulation of the firing rate. Thus we can think of the neuron as looking for this waveform, in accord with the triggering picture, but now the triggering is probabilistic.

If we continue along the lines of the Wiener method and cross-correlate the spike train $\rho(t)$ with higher powers of the stimulus $s(t)$, we end up with objects such as the average correlation function of the signal preceding a spike, and so on. These terms indicate that spikes may be triggered not just by particular features of the waveform itself, but by higher order features such as structure in the envelope of the waveform (Marmarelis and Marmarelis 1978). Auditory neurons tuned to high frequencies typically have a first Wiener kernel near zero but a nonzero second kernel, and the second kernel can be analyzed to show that the cell is sensitive to fluctuations in the envelope of the waveform as seen through a band-pass filter. Similarly, a pure visual motion sensor would have zero first kernel, so that flickering of the image would produce no modulation of the firing rate. But the second kernel would show that

the rate *is* modulated by the spatiotemporal correlation that corresponds to motion across the visual field. Although these methods can reveal sensitivity to more complex features in the stimulus, all of the cross-correlation functions in conventional white noise analysis of neurons involve only one spike arrival time. Because the averaging is always triggered by a single spike, this approach amounts to a Wiener expansion of the functional relation between the stimulus $s(t)$ and the time dependent firing rate $r(t)$.

White noise methods, as we have described them here, are a family of methods for measuring the terms in a series expansion of the input/output relation. As applied to spiking neurons, this input/output relation is the relation between the stimulus and the firing rate. Given that we are working within this class of models, the Wiener methods provide us with a very efficient way of measuring the response functions. In contrast, if we try to use sine wave stimuli, we have to do the experiment with each frequency in sequence, then use all possible frequency pairs to measure the second order nonlinearity, and so on. Even with the Wiener method, however, reliable estimation of higher order kernels requires very large amounts of data, and the analysis of neurons is seldom carried out to fourth order. Thus the mathematical problem of whether this general description converges to the right answer with enough terms is not so relevant to the design of real experiments. The more interesting question is whether we have some reason to believe in the validity of the linear or weakly nonlinear models obtained by keeping just the first few kernels.

Many of the phenomenological laws we are taught in physics courses are approximations, and they are the same kind of approximation we are discussing here. Thus, Ohm's law tells us that the current that flows in a wire is proportional to the voltage drop across the length of the wire. Similarly, Hooke's law tells us that the stretching of a spring is proportional to the force. If we pull hard enough or apply a large enough voltage, these linear relations break down, and nonlinear terms in the Wiener or Volterra expansion of the system response become important. The reason that the linear approximation works is that there is a dimensionless parameter $\alpha$ such that the $n^{\text{th}}$ term in the series is roughly proportional to $\alpha^n$, so that if $\alpha$ is small higher terms in the series become negligible.

When we stretch a perfect crystalline block, for example, the strain is shared equally among all the interatomic bonds along the direction of the stretch. Thus, if the whole crystal is lengthened by 5%, each bond is lengthened by 5%. We can translate this strain on the interatomic bonds into an energy, and (roughly speaking) we can take our parameter $\alpha$ as the ratio of this strain energy to the energy of the bond itself. Hooke's law works because the macroscopic strain energy, when shared among all the bonds, is small compared to the chemical bond energy. If we expect a series expansion to give a good description of neural responses, then we need to identify an analogous small parameter.



From dimensional analysis we can show that for neural responses to sensory stimuli, the parameter $\alpha$ must be like the typical size of the signal $s$ compared to some natural scale $s_0$. What sets the natural scale of the signal? One possibility is that the natural scale is set by the effective noise level in the system. Thus we might imagine that the natural scale of mechanical displacements in the ear is set by the level of Brownian motion, which is essentially the displacement at the threshold of hearing (for review see Bialek 1987). But in this case, rapid convergence of the series would require that signals are always on the order of the threshold signal. The auditory system exhibits all sorts of interesting nonlinearities in response to sounds that are barely audible, and the dependence of these nonlinear outputs on the intensity of input sounds is not at all what one predicts from the first few terms in a functional series of the Wiener type (Goldstein 1967). This is a clear example of how the first few terms of a functional series are not sufficient to describe perceptually important nonlinear responses in a sensory system, and we assume this failure arises from the lack of a natural small parameter. Analogous nonlinearities have been seen in the voltage responses of individual hair cells (Jaramillo, Markins, and Hudspeth 1993), and it will be interesting to see if the power series approach fails here as well—do the anomalous nonlinearities arise in single cells, or are they the result of collective interactions among many cells in the cochlea?

In the case of vision, one is often interested in the response of neurons to changes in light intensity (contrast) around some background level. Now the background level itself can set a natural scale, and the convergence of a series approximation will depend on the smallness of contrast as seen through the receptive field of the neuron. This has at least a chance of working, since the average contrast of the natural world isn't so large, of order 30% when seen through the photoreceptor array of the fovea and less when seen through larger apertures (Laughlin 1981; Ruderman and Bialek 1994). One must be careful, though, because the distribution of contrasts in natural images has a long tail (Ruderman and Bialek 1994). Photoreceptors and many retinal ganglion cells are quite linear up to 30% contrasts, and measurement of the first Wiener kernel provides an efficient method for extracting the form of this linear response in space and time. As an example of these ideas we discuss, in section 3.1.4, the linear response of photoreceptors and lamina cells in the fly's visual system.

Even for visual neurons that respond linearly to moderate contrasts, however, the form of the linear response depends very strongly on the background light intensity, as first noted for retinal ganglion cells many years ago (Barlow, FitzHugh and Kuffler 1957). If we want to give a completely general characterization of the input/output relations in the visual system, however, the separation of signals into a constant background and a small contrast is not allowed. Instead, changes of the background intensity should just be viewed as low frequency, large amplitude components of the input stimulus. But then no small number of terms in the Wiener series can describe the full response of a retinal ganglion cell. The system is adaptive, and adaptive nonlinearities tend to be strong and poorly described by power series. A simple idea such as a response time constant that depends on the background light level, as occurs already in the photoreceptors (Baylor and Hodgkin 1974), is very difficult to express in the language of Wiener kernels. It is not that the theoretically infinite series cannot describe the phenomena, but rather that practical low-order approximations to the series won't work. Thus some simple and robust features of the cell's response may be hidden from us if we try to force the data into a simple Wiener–Volterra description.

The problem of adaptation could be even more serious (and more interesting). In a white noise experiment on, for example, the visual system, one must choose not only the background light level but also the spectral density of contrast fluctuations that will be used as the probe signal $s(t)$. What if the visual system adapts not only to the mean light level but also to the variance of the light level, or contrast (de Ruyter van Steveninck et al. 1994; Smirnakis et al. 1995; de Ruyter van Steveninck et al. 1996; Smirnakis et al. 1996)? Again such adaptation is extremely difficult to describe in terms of a series expansion, and it seems more economical to describe the system by saying that it responds in different ways to signals drawn from different stimulus ensembles. But this means abandoning the idea that some systematic method will lead to a complete description of the response to arbitrary stimuli.

To summarize, we have seen that in some cases—such as the visual response to low contrast images—it is reasonable to expect that sensory neurons will give linear or nearly linear responses. In this limit, white noise methods provide a very efficient method for measuring the linear and nonlinear response functions of both spiking and nonspiking cells. It does not seem very efficient to use the Wiener or Volterra expansion to describe the profound nonlinearities associated with adaptation, but in a phenomenological approach one can use these methods to describe the changes in coding and computation that occur as the result of adaptation.



### 2.1.4 Models for firing statistics

In the preceding sections we have seen what can happen in the attempt to give a "complete" characterization of neural responses. Thus we start with firing rates as defined by Adrian, move on to time dependent rates and then intervals and correlation functions, and at each stage we see new phenomena that generate new candidate codes. Similarly, Wiener and Volterra gave us methods to quantify the notions of receptive fields and tuning curves in terms of linear and nonlinear response functions, but we also find that interesting phenomena are spread out over many different terms in these systematic expansions. It is not that exploration of higher order statistics or higher order Wiener kernels has not uncovered anything interesting. On the contrary, the problem is that each new order uncovers something new. We do not seem to be converging to a concise description of the neural code, so perhaps it is time to step back a bit.

In this section we explore simple, approximate descriptions of spike statistics. Such models can be useful, even if they are only approximate, because they give us some guidance about what to look for in the thicket of higher-order responses. In addition, we can make analytic statements about the implications of whole classes of such models for different interesting quantities such as the reliability of the code or the possibility of decoding the spike train. We will come back to these applications of the models, but here we want to understand how the models are defined and how one can check that they are good (or bad) approximate descriptions of a particular neuron. As one might hope, this exercise of checking the validity of simple models uncovers some interesting new phenomena.

Perhaps the simplest model of neural firing statistics is the Poisson model. The defining feature of a Poisson process is that the firing of one spike occurs with some probability per unit time—the rate—and this rate can depend on time but *not* explicitly on the occurrence times of the other spikes. How do we test the hypothesis that the spikes from a given neuron form a time dependent (or inhomogeneous) Poisson process, and what are the implications of such a model?

We start with the stimulus waveform $s(t)$, which determines the firing rate $r[t; s(\tau)]$. Because of the independence of spikes in the Poisson model it must be that this rate determines everything we might want to know about the spike statistics. The time dependent rate is, as in Fig. 2.1, the probability per unit time of observing a spike; more precisely, if we look in a bin of size $\Delta\tau$ surrounding the time $t$, the probability of observing a spike in this bin is $p(t) = r[t; s(\tau)]\Delta\tau$. If we want to know the probability of observing a

sequence of spikes at times $t_1, t_2, \ldots, t_N$, then we need to compute the probability of finding spikes in these bins, but also the probability that *no* spikes fall in any of the other bins. Because the spikes occur independently, the probability of spikes occurring in the $N$ chosen bins is

$$\begin{aligned} P(\text{spikes in bins}) &= r[t_1; s(\tau)](\Delta\tau) \times r[t_2; s(\tau)](\Delta\tau) \times \\ &\quad \cdots \times r[t_N; s(\tau)](\Delta\tau) \\ &= r[t_1; s(\tau)] r[t_2; s(\tau)] \cdots r[t_N; s(\tau)] (\Delta\tau)^N . \end{aligned}$$

As explained in section A.4, the probability of finding no spikes in any of the other bins is given by an exponential,

$$P(\text{no spikes in other bins}) = \exp\left\{ -\int_0^T dt\, r[t; s(\tau)] \right\},$$

where we are looking at the spike train over the interval $0 < t < T$. Putting the various factors together, we have the probability for observing the particular spike train $t_1, t_2, \ldots, t_N$, where we mark the spike occurrence times in bins of size $\Delta t$:

$$\begin{aligned} P[\{t_i\}|s(t)](\Delta\tau)^N &= P(\text{spikes in bins}) \\ &\quad \times P(\text{no spikes in other bins}) \qquad (2.17) \\ &= \frac{1}{N!} r[t_1; s(\tau)] r[t_2; s(\tau)] \cdots r[t_N; s(\tau)] \\ &\quad \times \exp\left\{ -\int_0^T dt\, r[t; s(\tau)] \right\} (\Delta\tau)^N . \qquad (2.18) \end{aligned}$$

We notice that the probability is proportional to the volume $(\Delta\tau)^N$ of the bins that contain spikes, as it must be—the probability of observing an event is larger if we are less precise in defining that event. We then divide out the bin sizes and work with the probability distribution $P[\{t_i\}|s(t)]$.

In working with probability distributions we should always be careful to keep track of units and of normalization. $P[\{t_i\}|s(t)]$ is the probability distribution for spike arrival times, so the term corresponding to the arrival of $N$ spikes must have units of $(\text{time})^{-N}$ or, equivalently, $(\text{rate})^{+N}$. If we integrate over the arrival times and sum over the spike counts, we have exhausted all possible spike trains, and so the total probability has to be one; this is the normalization condition. But when we integrate over all spike arrival times we have to be careful not to overcount—all the spikes are identical and we could always choose a different assignment of the $t_i$'s to the spikes, and to take care



of this fact we need the factor of $N!$ in Eq. (2.18). Checking for normalization is a good exercise to make sure we understand how to manipulate probability distributions, and the details of this calculation are given in section A.4.

How do we recognize a Poisson process? Perhaps the clearest test is given by the spike count, or pulse number distribution. If we look in an arbitrary time interval, which we will say runs from time 0 to time $T$, then from Eq. (2.18) we can calculate the probability of observing exactly $N$ spikes. The result is

$$P(N) = \frac{1}{N!} Q^N \exp(-Q), \qquad (2.19)$$

where $Q$ is the average number of spikes; that is,

$$\langle N \rangle = \sum_N N P(N) = Q, \qquad (2.20)$$

and $Q = \int_0^T dt\, r(t)$, as one might expect: The average number of spikes is the time integral of the firing rate. We can also calculate the variance of the spike count, and we find that this variance is equal to the mean:

$$\begin{aligned} \langle (\Delta N)^2 \rangle &= \langle (N - \langle N \rangle)^2 \rangle \\ &= \sum_{N=0}^{\infty} (N - Q)^2 P(N) = Q = \langle N \rangle. \qquad (2.21) \end{aligned}$$

We emphasize that these different statements about spike statistics are not separate models, but rather direct mathematical consequences of the Poisson model. Finally, these results show that the spike count distribution is *not* sensitive to the time dependence of the firing rate (which might be difficult to measure), but rather only to the one number $Q$ that is measurable as the mean spike count. For details of the relations among these different statements see section A.5.

In a series of papers, Teich, Khanna, and coworkers have measured pulse number distributions and the ratio of variance to mean in spike trains from primary auditory neurons in cats. Initial results were focused on the spike count distributions in time windows of $\sim 100$ ms, and were measured in response to stimulation with pure tones at the most sensitive frequency of the cell. The results (Teich and Khanna 1985) were in reasonable agreement with predictions from a Poisson model. Many authors have plotted the variance of spike counts versus the mean and obtained an approximately linear relationship. One must be careful, however, because the mean spike count can be changed in two very different ways. In most experiments one varies the stimulus parameters to change the mean firing rate, holding the counting window $T$ fixed. If one

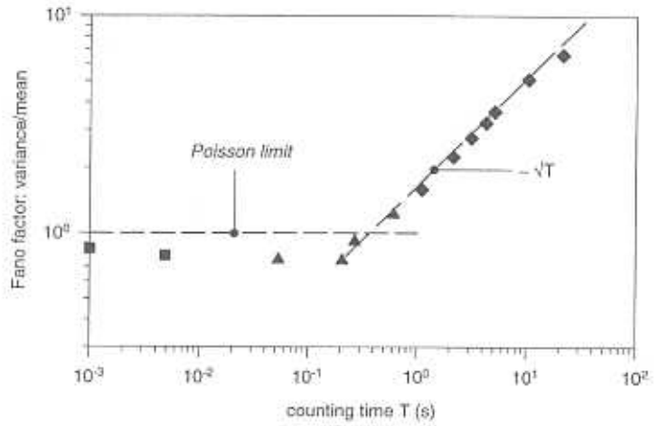



**Figure 2.9**
Evolution of the Fano factor with size of time window, $T$. The Fano factor is the variance of the number of spikes counted in a specified time window divided by the mean spike count, studied here for cells in the cat auditory nerve. In time windows smaller than 100–200 ms, the Fano factor is close to 1, consistent with Poisson fluctuations in the spike count. As the time window is increased beyond 500 ms, the Fano factor increases as $\sqrt{T}$—the variability of the spike counts exceeds that from Poisson fluctuations. Redrawn from Teich et al. (1990).

instead varies $T$ while holding the stimulus parameters fixed (Teich 1989; Teich et al. 1990), then the ratio of the variance to the mean—called the Fano factor—evolves as a function of integration time as shown in Fig. 2.9.

Figure 2.9 shows us that, while the Poisson model is unlikely to be an exact description of neural firing, it isn't a bad approximation over moderate time scales. On long time scales, however, something else is happening. One way to think about the increasing Fano factor is to imagine that, although the experimenters hold the auditory stimulus fixed, the firing rate of auditory neurons fluctuates, perhaps because of noise in the receptor cells or in the synapse between the receptor cell and the primary neuron. Ordinarily one would expect that this noise would be averaged away in larger time windows, so the Fano factor should plateau at large $T$; instead it seems to grow as $\sim \sqrt{T}$. This sort of excess noisiness at long times is what one would obtain, however, if the noise had a "$1/f$" type of spectrum. Most electronic devices have this sort of noise at low frequencies, and this noise limits the ability of experimenters to improve their measurements by increasing the integration time (Horowitz



and Hill 1980). The phenomenology of the Fano factor suggests that a similar problem may arise as the brain processes auditory information.

Returning to the more modest time windows, where a Poisson model has a chance of working, we can test this model in a very different way. While the spike count distribution is invariant to the time dependence of the rate, these time dependencies can produce correlations among spikes at different times. Johnson (1974) has studied the response of primary auditory neurons to pure tones and asked whether the correlation function can be predicted from observations on the time dependence of the rate using a Poisson model. This works quite well over a range of stimulus amplitudes, strongly supporting a Poisson model over reasonable time windows. To the extent that this procedure works, information carried in the second moments of $P[\{t_i\}|s(t)]$ is equivalent to the information carried by the time dependent firing rate $r(t)$, provided that this rate is defined in sufficiently small time bins.

The examples discussed above concern the response of auditory neurons to somewhat artificial stimuli, namely pure tones. More recently, Miller and Mark (1992) have studied responses to synthetic vowels, searching for departures from Poisson behavior. Quite dramatically they find that the variances of the Fourier components of the response are three times *smaller* than expected from a Poisson model. These data suggest that the neural response is more reliable when the system is confronted with more natural signals. This is obviously an important idea, to which we shall return several times.

Real spike trains can't be exactly Poisson processes. At the very least we know that spikes cannot come too close together in time, because the spike generating mechanism of all cells is *refractory* for some short time following the firing of an action potential. Thus the occurrence time of one spike cannot be completely independent of all the other occurrence times, as assumed in a Poisson model. When Poisson models give a good approximation to the data, it just means that the refractory time scale—or, more generally, any memory time scale in the spike generating mechanism—is short compared to the interesting time scales such as the mean interspike interval. This description is very general, and any attempt to understand near-Poisson behavior of neurons in terms of underlying molecular mechanisms must provide a basis for this separation of time scales.

If stimuli are constant, then the firing rate should be constant, and in a Poisson process the resulting interspike interval distribution is exponential. Furthermore, each interval is statistically independent of all the other intervals. As a first step toward a more realistic picture, one might try assuming that

intervals are still independent but that the interval distribution is nonexponential, incorporating (among other effects) refractoriness. Models of this type are called *renewal processes*.

Once again we can ask, how do we recognize a renewal process? For retinal ganglion cells in the cat, Troy and Robson (1992) have applied a test analogous to the test for Poisson behavior used by Johnson in the auditory nerve. Again the idea is that correlations among distant spikes must be built out of more elementary pieces, in this case the independent interspike intervals. If we measure the interval distribution we can calculate the correlation function (or its Fourier transform, the power spectrum) on the hypothesis that the intervals are independent. Under conditions of constant illumination, these calculations agree with experiment in impressive detail.

We see that models of neural firing statistics that might seem rather oversimplified actually work reasonably well if we don't push too hard. There are different attitudes toward such results. One attitude is that, because simple models are close to working, we should search exhaustively for the exact model that works perfectly. This means nailing down the corrections to the Poisson or renewal approximation, and then quantifying how rates and interval distributions are modulated by arbitrary stimuli—perhaps using the systems identification methods described above, caveats and all. A complementary attitude is that, if simple models come close to working, then we should exploit the simplicity and analyze these models thoroughly even though we know that they are not perfect. Much of the analysis of the simple models can be done with pen and paper rather than with computer simulations, so we have the chance of developing some intuition and generating some understandable and testable predictions. One shouldn't trust the details of such predictions, since they depend on the exact form of the model, but it might be possible to identify some robust qualitative conclusions that lead to the design of new experiments. It is this more intuitive use of the models that we shall emphasize in the following sections.

## 2.2 TAKING THE ORGANISM'S POINT OF VIEW

Firing rates, interval distributions, and so on, are average quantities, properties of an ensemble of spike trains rather than a single spike train. Claiming that "information is carried by the firing rate" doesn't really tell us how the neural code works because we haven't explained how the brain can measure "the firing rate" from a single example of the spike train, and this is the problem which generated our paradoxical discussion in section 2.1.2. As we start to see



just how serious this problem can be, we will also be led to a new point of view on the neural code.

### 2.2.1 Intervals in the signal and intervals between spikes

Consider again the example of interval coding in an auditory neuron. The interval distribution characterizes the response to a single tone of fixed amplitude and frequency. Real-world signals can be thought of as tones that are modulated, perhaps by large amounts, in both amplitude and frequency. If the modulations are very slow, many spikes are fired before the parameters of the tone change significantly. From these many spikes one can build up the distribution of intervals and thereby estimate the amplitude and frequency. But modulations in many biologically significant sounds (speech, bat echolocation, frog calls, cricket chirps) occur on time scales of 5–20 ms, during which time a cell firing 100 spikes per second can generate just one or two spikes. With such a small number of spikes we cannot accumulate a reasonable interspike interval distribution, and we cannot even get a good estimate of the rate before the parameters of the stimulus have changed.

The bat auditory system provides a clear example of the importance of small numbers of spikes. In recordings from the auditory cortex of *E. fuscus*, Dear, Simmons, and Fritz (1993; Dear et al. 1993) have studied the responses of cells to pairs of ultrasonic pulses that simulate the bat's own sonar call and a returning echo. A substantial fraction of neurons are selective for the echo delay, which measures the distance to the target under natural conditions. These delay-tuned cells respond weakly if at all to a single pulse or to simplified signals such as pure tones. If one chooses a delay that matches the tuning of the cell, then a single call–echo pair produces an average of just one spike, as shown in Fig. 2.10. This spike itself appears to occur at a precise time relative to the arrival of the echo.

In trying to characterize the response of a neuron, it is tempting to focus on the stimuli that generate the largest responses, and this complicates our efforts to estimate the "typical" number of spikes representing significant stimulus variations. In the cortex especially, neurons are often extremely selective for complex features of the stimulus, and hence one might worry that much higher spike counts could be observed in response to a properly chosen signal. For experiments in the bat auditory cortex, however, these concerns are answered by combining the neural recordings with our understanding of the animal's behavior. The cells studied in the experiments of Fig. 2.10 *are* sensitive to complex features of the stimulus, being selective both for delays and for the combination of harmonics that makes up a natural bat call. But we are

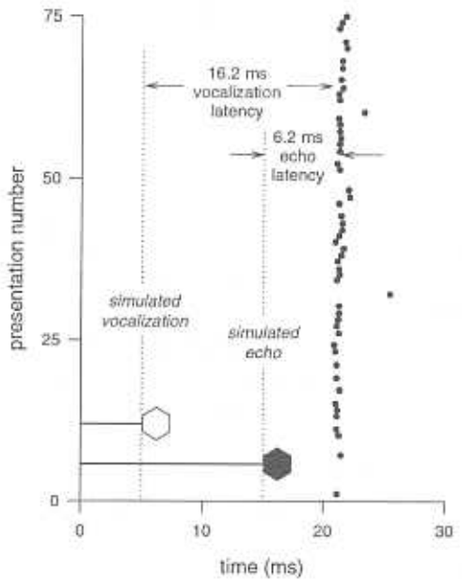



**Figure 2.10**
Response of a neuron in the bat auditory cortex. The response to a simulated echo consists of usually one and occasionally no spikes; spike responses are generated at a very constant latency. Redrawn from Dear, Simmons, and Fritz (1993).

not free to search the space of all possible stimuli, because we know that bats navigate using relatively stereotyped calls—behavior is driven by stimuli that are almost identical to those used in the physiological experiments. Finally, we know that single call–echo pairs are sufficient for the bat to make behavioral decisions (Griffin 1958; Simmons 1989). Thus it seems inescapable that significant variations in the bat's acoustic environment are represented by of order one spike or less from each cortical neuron.

The similarity of time scales in natural signals and typical interspike intervals is not confined to the auditory system. In the fly visual system, as we discuss in later sections, movements across the visual field can result in the generation of a compensating flight torque within 30 ms (Land and Collett 1974). During this time the handful of movement-sensitive neurons (Hausen 1984) can generate just a few spikes each. In certain species of moths, complex bat-evading flight paths are triggered by bat cries just loud enough to produce one or two spikes in each of the two most sensitive auditory neurons (Roeder, 1963).



In the primary visual cortex of monkeys, preattentively discriminable textures produce an average of 1 to 3 spikes per cell within the 50–100 ms behavioral decision time (Knierem and van Essen 1992). For the comparable cells in cats, optimally chosen moving gratings produce modulations of less than 3 spikes per 100 ms (see, for example, Reid, Soodak, and Shapley 1991). When we discuss the reliability of neural computation in chapter 4, we shall see that the discrimination ability of many neurons is dominated by short time windows in which roughly one spike is fired. Similarly, from an information-theoretic point of view (chapter 3), we shall see that a large fraction of the information available from the response to a transient stimulus is carried by the first spike or two.

Recently Gallant, Connor, and van Essen (1994) have studied the responses of primary visual cortical neurons under conditions where the monkey is allowed to move its eyes and scan a static image as it chooses. Under these conditions average firing rates are 10 to 50 spikes/s, so that the number of spikes produced during a single fixation period is roughly 1 to 5, as in the texture discrimination experiments. Although one can drive these cells to very high sustained firing rates, it appears that typical spike counts under more natural conditions are much smaller.

In the rat somatosensory cortex there are cells responsive to the displacements of individual facial whiskers, and the rat uses this tactile sense to explore its environment. In a typical behavioral experiment, the rat is asked to run around a track and then stop to make a tactile discrimination of the texture on some target. The rat is free to make and break contact between her whiskers and the target, so she controls the strength and duration of the tactile stimulus; after contact is broken the rat makes her decision and turns to the right or left as appropriate to the task. Recordings of the spike activity of neurons in the primary somatosensory cortex during such behaviors (see, for example, Fee and Kleinfeld 1994) reveals that the brief contact period required for decision making produces an average of order one spike in the most responsive cortical cells, as illustrated in Fig. 2.11. This result is very much the same as in the bat auditory cortex: The stimulus consists of a brief pulse, the dynamics and magnitude of the pulse are determined by the animal's own behavior, a single pulse is sufficient for a behavioral decision, and cortical neurons produce on the order of one spike per pulse.

It is likely that the importance of small numbers of spikes is not limited to the early stages of neural signal processing. In the farthest reaches of primate visual cortex, cells respond selectively to particular faces; for a review see Gross and Sergent (1992). Even a 20 ms presentation of a face evokes

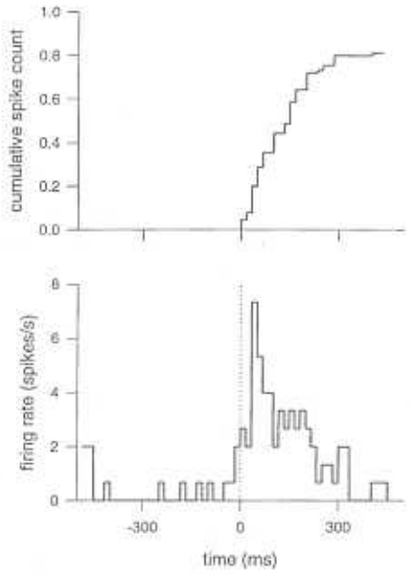



**Figure 2.11**
Dynamics and cumulative spike count for a cell in the rat somatosensory cortex. At time $t = 0$ the rat's whisker makes contact with an object it must identify, and at some point, which varies from trial to trial, contact is broken and a decision is made. But we see that, on average, the entire encounter with the object—the dynamics of which is controlled entirely by the rat herself—produces slightly less than one spike. Redrawn from experiments by Fee and Kleinfeld (1994), with our thanks to authors.

long-lasting ($\sim$ 400 ms) firing, but if the face is followed by a mask this continuing activity can be silenced and the *maximum* output of the cell is of order 5 spikes (Rolls and Tovee 1994). Nonetheless we can perform reliable recognition under these rapid-mask conditions; for further analyses of cortical responses on short time scales see Panzeri et al. (1996). More generally, Thorpe and coworkers have emphasized that surprisingly sophisticated visual tasks can be performed so rapidly that each layer of the visual system has the chance to fire only of order one spike before passing its "result" on to the next stage of processing; for a brief overview of these arguments, see Thorpe (1990; Thorpe, Fize, and Marlot 1996).

In the rat hippocampus, which receives input only from higher sensory cortices, cells have been found that are selective for the spatial location of the rat (O'Keefe and Nadel 1978). When the rat is exploring its environment



freely, these "place cells" fire at peak rates of $\sim$ 30 spikes/s (see, for example, Wilson and McNaughton 1993; O'Keefe and Recce 1993). If we imagine that the rat has knowledge of its own position with roughly centimeter accuracy, then since it can move at speeds of $\sim$ 20 cm/s, hippocampal signaling about position must be based on 1 or 2 spikes per cell, of the same order as in sensory cortex.

Although there are only a small number of spikes per neuron, there are many neurons. One could then imagine building up estimates of the firing rate, the interval distribution, and so on, by averaging over an ensemble of cells. It is certainly the case that many sensory signals are shared among large numbers of cells, and that the overall performance of the organism depends on its ability to integrate this large array of data. This does not excuse us from thinking about small numbers of spikes. To begin, many invertebrates do *not* have large numbers of cells with which to generate ensemble averages, and their brains certainly work. When we humans sit in a dark room, we are capable of seeing single photons, as discussed in section 4.1.2, and these perceptions must be based on a small *total* number of spikes from the entire array of $\sim 10^6$ neurons in the optic nerve. Recent experiments on hyperacuity (see section 4.2.2) suggest that these extremely precise spatial judgments are also based on small numbers of spikes. Finally, there are experiments suggesting that we can "feel" the individual action potentials produced by the mechanosensors of our skin (Valbo 1995).

The fact that our coherent perception of the world around us is based on large numbers of spikes from many neurons does not imply that the ingredients of these coherent percepts are similarly carried by large numbers of spikes or neurons. Experiments on the limits to our perception—photon counting, hyperacuity, the threshold of touch—suggest the opposite: Human observers can report reliably the arrival of small numbers of spikes from their sensory neurons. We will return to this issue many times.

There is a more fundamental problem with assuming that the nervous system has access to many spikes simply because it has many neurons. For example, if one wants to measure the rate of firing by pooling the spike trains of many cells, one must make the hypothesis that these many neurons carry the same (or nearly the same) signals, so that averaging their responses makes sense. Furthermore one must assume that these responses are statistically independent, so that averaging actually improves the reliability of the signal. If both these hypotheses are correct, averaging over neurons is equivalent to averaging over multiple presentations of the same stimulus, and the brain (remember our homunculus) can measure the conventional rates and interval

distributions even though each cell generates of order one spike. But this combination of redundancy and statistical independence is an extreme hypothesis, and we shall see that there is direct evidence against it.

To return to our discussion in the introduction, we are taking an empirical point of view, trying to understand how to build the homunculus who looks at just one cell. It may be that this particular homunculus can tell us very little, and that his only hope to make sense out of the spike train data is to confer with his fellow homunculi who monitor the outputs of other cells. Alternatively, the impoverished homunculus who looks at the spike train of just one neuron might be able to reach precise and unambiguous conclusions about a small part of the sensory world. But whatever the homunculus can say, and indeed whatever he brings to the discussion with his fellows, he must base on just a few spikes from the cell he observes.

In this section we have surveyed a selection of experiments from many different systems. Although we may be biased in our selection, it seems clear that—at least under some conditions—many neurons make use of *sparse coding in the time domain*. These cells fire of order one spike for each characteristic time of variation in the stimulus, where the "characteristic time" has to be defined with reference to some reasonably natural behavior. Under these conditions, individual spikes must carry significant information simply because there are no more spikes. This notion that single spikes can be important drives much of the discussion in the following chapters. We begin, however, with the obvious question.

### 2.2.2 What can small numbers of spikes tell the brain?

As we have outlined in the preceding sections, the traditional approach to studying the neural code has been to catalog the average behavior of neurons in response to changes in stimulus parameters. The rate versus timing debate has been framed as a question of whether all the information about the stimulus is carried by the firing rate—or, equivalently, by the spike count in some specified time window—or whether the timing of individual spikes within this window also correlates with the variations in the stimulus. But we have seen that, in several cases, the time windows of relevance to behavior contain of order one spike. In this limit, the colloquial distinction between rate and timing codes isn't very helpful. If a single spike is delayed by a few milliseconds, should we think of this as changing the spike count or rate measured in a small window, or is the spike timing per se the significant variable? Before discussing such subtleties, one might wonder how so few spikes can convey any information at all.



Whatever the ultimate characterization of the neural code, it seems unreasonable to imagine that the occurrence of a single spike can lead to complete certainty about the nature of the sensory stimulus. To talk about the information conveyed in small numbers of spikes, then, we need a language that quantifies our degree of certainty or uncertainty. This language is, again, probability theory. As described in section 2.1, conventional approaches to the neural code can be thought of as taking various slices through the distribution of spikes in response to a known stimulus, $P[\{t_i\}|s(t)]$. The organism, however, is not interested in predicting spike trains from known stimuli. On the contrary, the organism has access only to the spike train $\{t_i\}$ and must mediate behaviors that (hopefully) make sense in response to the unknown stimulus $s(t)$. From the point of view of the organism, then, we ask what one knows about the stimulus by virtue of observing the spike train. All of this knowledge is contained in the conditional distribution $P[s(t)|\{t_i\}]$, which measures the relative likelihood of different stimulus waveforms given the particular spike train $\{t_i\}$.

We have defined two different conditional probability distributions—the distribution of spike trains given the stimulus, and the distribution of stimuli given the spike train. As explained in section 2.1.1, these two distributions are related through Bayes' rule. Although this is a simple mathematical fact, Bayes' rule tells us some important things about the structure of the neural code.

We have emphasized that the probability distribution $P[\{t_i\}|s(\tau)]$ describes the *encoding* of stimuli into spike trains. Although it is impossible in practice, let us imagine that we have understood everything there is to know about this distribution—we understand the rules whereby sensory stimuli trigger neural spikes, including all the complexities of noise, adaptation, and nonlinearities. Then the distribution $P[\{t_i\}|s(\tau)]$ is fixed, once and for all. The fundamental consequence of Bayes' rule is that *this apparently complete knowledge of the neuron's encoding strategy is not, by itself, sufficient to tell us what a given spike train means or stands for in the outside world!*

When we observe a sequence of spikes at times $t_1, t_2, \cdots, t_N$ and ask what sensory stimulus caused these spikes, we need to look at the probability distribution $P[s(\tau)|\{t_i\}]$—the distribution that tells us the relative likelihood of different stimuli given our observations on the spike train. But from Bayes' Eq. (2.4), this distribution is a product of three terms:

$$P[s(\tau)|\{t_i\}] = P[\{t_i\}|s(\tau)] \times P[s(\tau)] \times \left(\frac{1}{P[\{t_i\}]}\right). \tag{2.22}$$

The first term is the encoding distribution described above, but the second term is the distribution of signals in the world. The third term, the probability of observing this particular spike sequence, serves to normalize the distribution. Even if we characterize completely the encoding of signals into spikes, the interpretation of these spikes as standing for signals in the outside world depends on the characteristics of the world itself. This is a familiar idea—the meaning of a statement depends on the context in which it occurs. Though we routinely use this idea to discuss the problem of communication between people, Bayes tells us that the notion of context is equally relevant to the interpretation of spike trains from a single neuron.

One of the most important aspects of the decoding approach, as captured in the metaphor of the homunculus, is that the stimulus is unknown to the animal. In the natural environment or in an experiment, stimuli are chosen at random from some probability distribution $P[s(t)]$ which defines the stimulus ensemble. Many experiments use simple ensembles (e.g., sine waves), but in these cases $s(t)$ can be predicted perfectly from knowledge of its past $s(t' < t)$. In this sense one can know the stimulus without looking at the spike train, and no new information is gained by observation of the spikes. To get started on the problem of what the spike train means about signals in the outside world, we need to choose these signals from an ensemble rich enough that we (or the organism) can really learn something about the world by continuous observation of the spike train. Obviously a completely natural stimulus—such as a tape recording of an evening at the frog pond—has this richness, but these natural signals are also difficult to characterize, as we discuss in section 5.2.

We can ask our question about the meaning of the spike train in any stimulus ensemble, and we know that the answer may be different in each case. Ideally, then, we would like to explore many different ensembles, working our way toward the signals that actually occur in nature (see section 3.3.3). This is also a major theme in the neuroethology literature, and it seems a rather basic "biological" point of view to say that signals acquire their meaning only in the context provided by the sensory environment as a whole. This view, however, seems opposed to some of the traditional quantitative analyses, which aim at a characterization of the nervous system or of a particular neuron as an isolated device. We believe that the approach developed in the following sections gives us a way of quantifying the ethologists' intuition, attaching numbers to the context dependent meaning of spike trains.

The characterization of the neural code from the point of view of the organism was advocated in early work by FitzHugh (1958), who emphasized the



need for the organism to perform a statistical analysis of the spike trains in its sensory neurons. Such an analysis was subsequently carried out by Barlow and Levick (1969) in experiments on the detection and discrimination of light flashes by ganglion cells in the cat retina, as will be described in section 4.1.2. These experiments, however, focused on forced-choice discrimination among a small number of possible signals. We have posed the more difficult problem of making continuous inferences about an unknown time varying signal, the "running commentary" discussed in section 1.2. As far as we know, the discussion closest to our own is that of Johannesma (1981), and we return to his ideas in section 2.3.1.

### 2.2.3 Response-conditional ensembles

It has been possible to give an experimental characterization of the conditional distribution $P[s(t)|\{t_i\}]$ from the responses of a motion sensitive neuron, H1, in the fly's visual system (de Ruyter van Steveninck and Bialek 1988). Because this system, and indeed this one identified neuron, provides examples for several of the ideas in subsequent sections, we take some time here to give a brief overview of fly vision and the role of visual movement estimation in fly behavior.

When we watch a fly buzzing around a room, we notice that its flight path consists of relatively straight segments interrupted by sharp turns, and this impression can be quantified (Wagner 1986a, 1986b, 1986c). If we turn out the lights, the fly lands. The ability of the fly to maintain a steady course depends on sensory, particularly visual, feedback. Careful analysis of the trajectories of flies during chasing behaviors indicates that a change in visual input can trigger a change in flight path with a latency of just 30 ms (Land and Collett 1974).

One can demonstrate the visual input to flight control by tethering the fly so that it hangs, wings flapping, from a torsion balance. If the visual environment of the fly rotates (on a drum surrounding the fly, or on a video monitor), the fly generates a torque. The sign of the torque is such that it tends to compensate the rotational motion. One can close the sensory–motor feedback loop artificially by giving the visual environment an added velocity proportional to the negative of the measured torque, as would happen if the fly were free to turn. Under these closed loop conditions, the fly will spontaneously fixate an object, creating (as best as possible under the circumstances) the image of flying straight toward that object. These basic facts about optomotor control were established in a series of experiments by Reichardt and collaborators, summarized by Reichardt and Poggio (1976).

More recently, Heisenberg and colleagues have emphasized that important aspects of this control loop are plastic, and can be modified by the fly in response to changes in the simulated flight mechanics (Heisenberg and Wolf 1984; Wolf and Heisenberg 1990). Physically, if the fly wants to turn by 10 degrees in 100 ms, the question of how much torque it should apply to its body is actually quite complex, because the fly is flying under conditions in which the airflow over its body is extremely unsteady (Dickinson and Götz 1993; Dickinson 1994). Furthermore, the exact answer to this question depends on the way in which the wings have hardened after metamorphosis, on whether the wings have been chipped or otherwise damaged, on the prevailing wind conditions, and even on whether the fly has eaten recently. Thus there can be no straightforward autopilot that converts sensory stimuli into motor programs through a constant, genetically determined rule; the fly must learn and constantly update the rule appropriate to its current flight dynamics.

Input to the fly's visual motion computations comes from a single class of photoreceptor cells arrayed beneath the lenses of the compound eye. Photoreceptor signals are processed by the cells of the lamina, medulla, and lobula before arriving at the lobula plate, where one finds a handful of large, identified movement sensitive neurons (Hausen, 1984). Some of these cells respond most strongly to "wide field" motion—coherent motion across the entire visual field, as would be induced by rigid rotation of the fly itself. Other cells are selective for narrow field motion, as occurs when a small object moves relative to the fly and the background (Borst and Egelhaaf 1989; Hausen and Egelhaaf 1989). Destruction of individual cells in the lobula plate produces specific deficits in optomotor behavior (Hausen and Wehrhahn 1983), strongly suggesting that these motion sensitive cells are an obligatory link in the path from visual input to motor output.

For the benefit of those readers more familiar with the vertebrate visual system, we emphasize several features of the fly's brain. First, the lobula plate is at least four synapses removed from the photoreceptors; in mammals the primary visual cortex is found at this stage. Second, processing is not simply feedforward through the layers of visual neuropil; instead, there are substantial lateral interactions at each stage, as well as projections back from medulla to lamina. Finally, fly vision is more than just motion detection: Insects have a memory for spatial patterns (Dill, Wolf, and Heisenberg 1993), a spontaneous preference for novel images (Dill and Heisenberg 1995), and individual neurons in the lobula are selective for the orientation of barlike stimuli in a manner rather analogous to cells in visual cortex (O'Carroll 1993).

We review here experiments on H1, which is a wide field, horizontal movement sensor. Under favorable conditions it is possible to record continuously



from H1 for periods of many days, using an immobilized fly, almost completely intact save for a small hole in the back of the head that allows access to the lobula plate. In these long experiments one must pause occasionally to feed the fly, but it should be clear that this very stable preparation makes it possible to address questions that require a very large statistical sample of neural responses.

In the experiments that probed the structure of $P[s(\tau)|\{t_i\}]$, the fly looked at a moving pattern presented on a display oscilloscope.[2] The signal waveform $s(t)$ is the time dependent angular velocity of the motion, with the spatial structure of the pattern held fixed. Most of the results discussed here are for a pattern that approximates spatial white noise along the horizontal axis and is uniform along the vertical. This is a very simple choice that distributes horizontal motion cues in a statistically uniform fashion across the entire visual field; we will also compare coding of these stimuli with the coding of more spatially restricted signals. Finally, the velocity waveform is chosen from a stimulus ensemble that approximates Gaussian white noise, so that the pattern diffuses across the visual field. Figure 2.12 shows the stimulus waveform from a segment of the experiment, together with the procedure for constructing the distributions $P[s(\tau)|\{t_i\}]$.

The distributions $P[s(\tau)|\{t_i\}]$ provide, in effect, a dictionary for the neural code in which we can look up the stimulus most likely to have generated a particular spike sequence. In addition, the width of the distribution $P[s(\tau)|\{t_i\}]$ measures our confidence in this most likely waveform as an estimate of the true stimulus. One can also use the distributions to quantify the information content of different spike sequences, demonstrating, for example, that short interspike intervals carry much more information per unit time than long intervals, and that the *absence* of spikes actually conveys a substantial amount of information per unit time. Perhaps the most important fact about the structure of the distribution $P[s(\tau)|\{t_i\}]$ measured in H1 is that it suggests the possibility of decoding the spike train, using the set of arrival times $\{t_i\}$ to generate a continuous estimate of the signal $s(t)$ in real time.

In the absence of any observations on the spike train, all we know is that the stimulus waveform was chosen from some a priori probability distribution

2. Flies, and many other insects, can respond to flicker (time variations of light intensities) at much higher frequencies than humans. Recordings in fly photoreceptors in bright background lights show clear responses above 100 Hz, while cells in the lamina that receive direct synaptic input from the receptors can have their *peak* response at nearly 100 Hz (see Fig. 3.12). This high temporal resolution precludes the use of ordinary video monitors for fly vision experiments; in the work reviewed here the display was refreshed 800 times per second (de Ruyter van Steveninck and Bialek 1988).

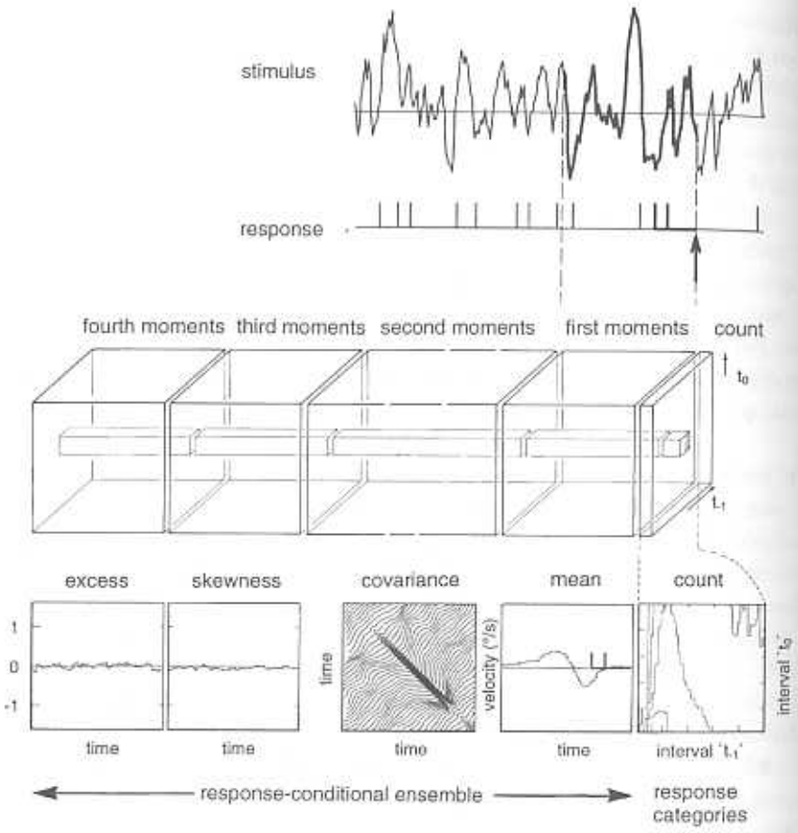



**Figure 2.12**
Procedure for constructing response-conditional ensembles. The two traces at the top show a sample of the stimulus (the velocity waveform $v(t)$ of a moving wide-field pattern), and a digitized representation of the response of the fly's H1 neuron to this movement. Occurrences of spike patterns, here consisting of an interspike interval $'t'_{-1}$ followed by an empty interval $'t_0{}^-$, are counted in the slice labeled "count" at the right hand side of the middle block; this procedure determines the joint distribution $P['t'_{-1}, 't_0{}^-]$. For each response category a 100 ms section, divided into 50 bins, of the preceding stimulus waveform is accumulated in the slot corresponding to $['t'_{-1}, 't_0{}^-]$ within the block labeled "first moments." Similarly, all the second moments (the products $v_i \cdot v_j$, $i, j = 1, \ldots, 50$ of all possible pairs of velocities in the 50 bins) and the third and fourth diagonal moments (all $v_i^3$ and $v_i^4$) are accumulated. After normalization we obtain the mean waveform, the covariance, and the diagonal elements of the skewness and the excess of the preceding stimulus ensemble, for each response category $['t'_{-1}, 't_0{}^-]$. An example is shown in the row of panels at the bottom. After de Ruyter van Steveninck and Bialek (1988).



determined by the experimental or environmental conditions. The observation of a certain response changes our statistical knowledge from the a priori ensemble to the conditional ensemble: We judge some classes of stimuli more probable and others less probable by virtue of observing this particular spike train $\{t_i\}$.

Let us see how the structure of $P[s(\tau)|\{t_i\}]$ emerges from experiment. Imagine taking a snapshot of the spike train at some observation time $t_{\text{obs}}$. At this instant, a time $t_0$ has elapsed since the last spike, while the second-to-last spike occurred a time $t_{-1}$ further in the past, and so on. We call this snapshot of the spike train $R$, since it is some particular response of the neuron. We would like to know what the observation of this snapshot $R$ tells us about the stimulus. Because the behavioral response times are short, it makes sense to look back at only a short history of the spike train in defining $R$.

Now let us suppose that we have performed a very long experiment, presenting the system with some randomly chosen, continuously varying $s(t)$. In such a long experiment, the response $R$ will have occurred many times—there will be many observation times $t_{\text{obs}}$ such that the last spike occurred at $t_{\text{obs}} - t_0$, and so on. Looking backward in time from each such $t_{\text{obs}}$, we will see some particular waveform $s(t_{\text{obs}} - \tau)$. As the experiment proceeds we keep a list of all the waveforms that precede the response $R$. In making this list we choose, out of the ensemble of all possible waveforms in our experiment, a particular subensemble, the *response-conditional ensemble*. This subensemble consists of waveforms which are chosen randomly out of the distribution $P[s(t_{\text{obs}} + \tau)|R]$, so that our experiment has given us a sort of Monte Carlo sampling of this distribution. In practice (de Ruyter van Steveninck and Bialek 1988), $P[s(t)|R]$ was approximated as a multidimensional Gaussian, characterized by the mean velocity vector $w_R(\tau)$ and the covariance matrix $C_R(\tau_1, \tau_2)$, with $R$ denoting the particular response that forms the condition. One can confirm that this is a good approximation by computing some of the third and fourth moments.

Representations of the response-conditional ensembles for a selection of different responses are provided by Fig. 2.13. Two simple conditions, a single spike $[0']$ and a 50 ms period of non-firing $[{}^-50^-]$, are depicted in Figs. 2.13a and b. The mean stimulus waveform conditional on a single spike is a smooth function of time, peaking 25 ms before the spike occurs. The line plot represents the covariance matrix, scaled by the a priori stimulus variance. The figure shows that the off-diagonal elements have a region of negative covariance, centered at about 35 ms before the spike occurs. The fact that the off-diagonal values are negative means that the waveforms that constitute the

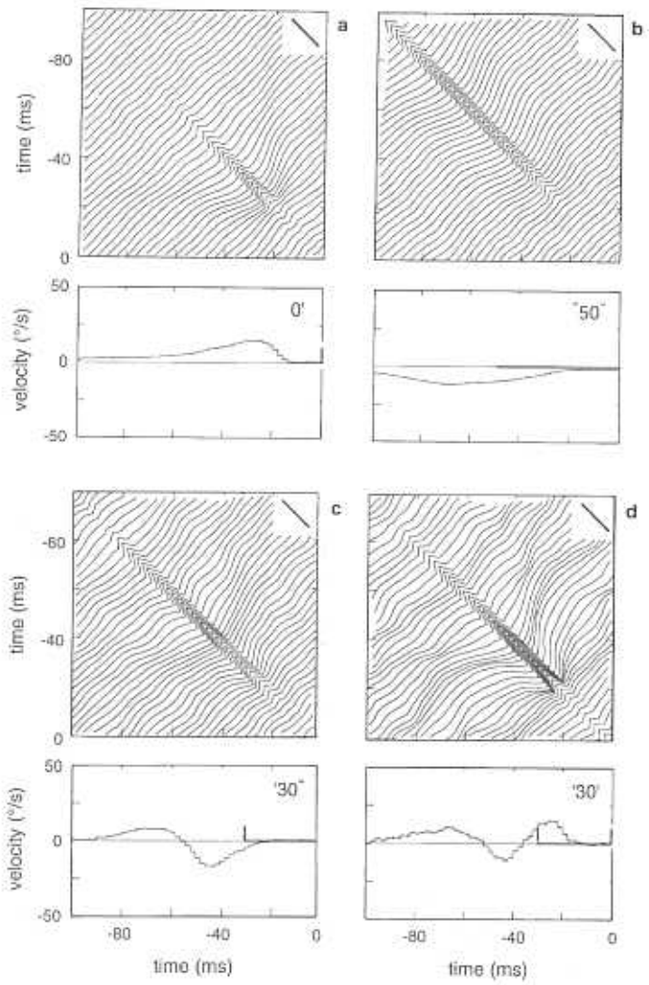



**Figure 2.13**
(a-h) Response-conditional ensembles for a selection of different response categories. For each category the response-conditional ensemble is represented by a conditional mean waveform (bottom), and a covariance (top). Response categories on which the ensemble are conditional are shown in the upper-right of the mean waveform boxes. The superscripts $^-$ and $'$ signify, respectively, the absence and the presence of spikes. For example, $'5'10^-$ stands for an interspike interval of 5 ms, followed by a 10 ms interval without spikes. The abscissae represent time with respect to $t_{obs}$ (the last point



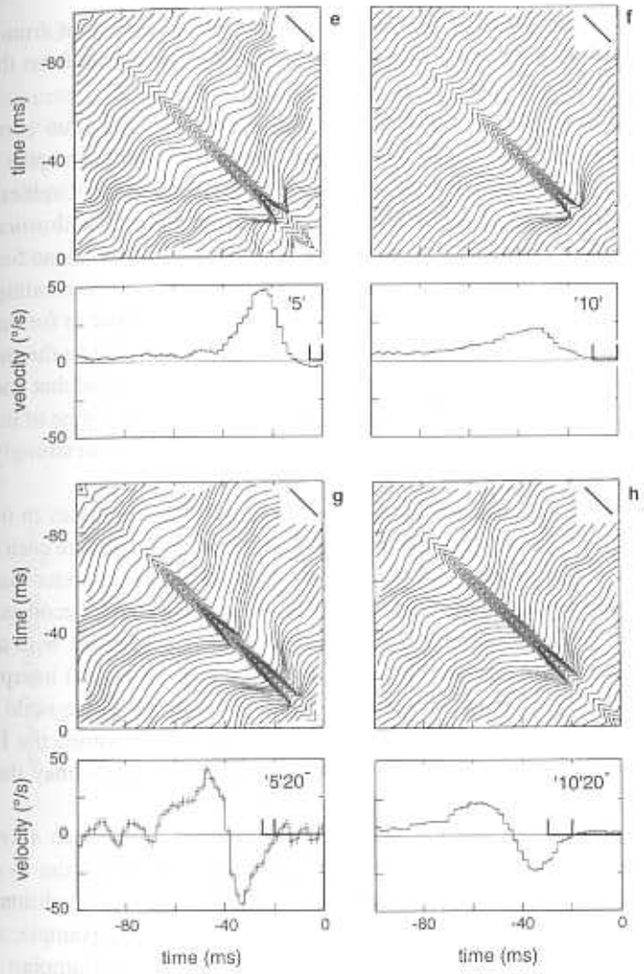



**Figure 2.13 (continued)**
of the response). Covariance matrices are represented by consecutive sections through the top-left bottom-right diagonal. The elements of this diagonal are set to zero to provide a reference. The calibration bars in the top-right corners represent a length of 0.05. Positive (negative) values are in the top-left (bottom-right) direction. Error bars on the mean waveforms are standard errors of the mean. After de Ruyter van Steveninck and Bialek (1988).

response-conditional ensemble are constrained to deviations from the mean that have less power at low frequencies than waveforms from the a priori ensemble.

It is clear that more complex conditions on the spike train correspond to more complex conditional mean waveforms—these complex spike sequences stand for more complex signals. For example, for a pair of spikes separated by a given interval the conditional mean waveform varies drastically as we change the interspike interval. For an interval of 10 ms, one can find the conditional mean waveform by adding up the waveforms corresponding to the individual spikes, but this is clearly not true for much shorter or for much longer intervals. It is interesting to note that 10 ms is very close to the most probable interval in this experiment, and it is only for this interval that one observes linear superposition of the single spike waveforms. The degree of nonlinearity in stimulus coding is larger with intervals that deviate more strongly from the most probable firing pattern.

For long intervals we can distinguish three different phases in the average velocity waveform: two positive peaks occurring 25 ms before each of the two spikes, and a trough in between. As the interval becomes shorter than 10 to 15 ms, these separate phases merge into a single peak, which becomes very high for the shortest intervals. This transition occurs for intervals with a length of the order of the photoreceptor integration time. A functional interpretation is that, loosely speaking, structure in stimulus events on a time scale below the photoreceptor integration time cannot be detected. However, the H1 neuron can generate intervals of shorter duration. These intervals may therefore be used to encode higher stimulus amplitudes.

One interesting question concerns the precision with which a hypothetical observer must measure the arrival times of the spikes in order to obtain the maximum possible information. Clearly, if two response conditional ensembles corresponding to different interspike intervals, for example, are essentially indistinguishable, the observer loses very little by lumping these two intervals in one bin. Since we approximate the distributions $P[s(\tau)|R]$ as multidimensional Gaussians, there is a natural measure of distinguishability: The length of the vector that points from the center of one distribution to the center of the other, appropriately normalized by the covariance matrix. This is the signal to noise ratio for discrimination between signals drawn from the two distributions, and it is also the discriminability parameter $d'$ used in the analysis of psychophysical experiments (Green and Swets, 1966); for more details, see the discussion of discrimination experiments in chapter 4.



Quantitatively, if we are presented with the event $R_1$ or the event $R_2$, $d'$ is related to probability that we can distinguish these events by looking at the stimuli that gave rise to them. If it is not possible to make this distinction, then $d'$ is near zero and we might as well consider $R_1$ and $R_2$ to be the same event. If the distinction is very easy, then $d'$ is large, and the crossover defining reliable distinction is conventionally taken as $d' = 1$. In Fig. 2.14 we show the results of analyzing discriminability among the response conditional ensembles that correspond to different intervals $t_0$, and these results are summarized

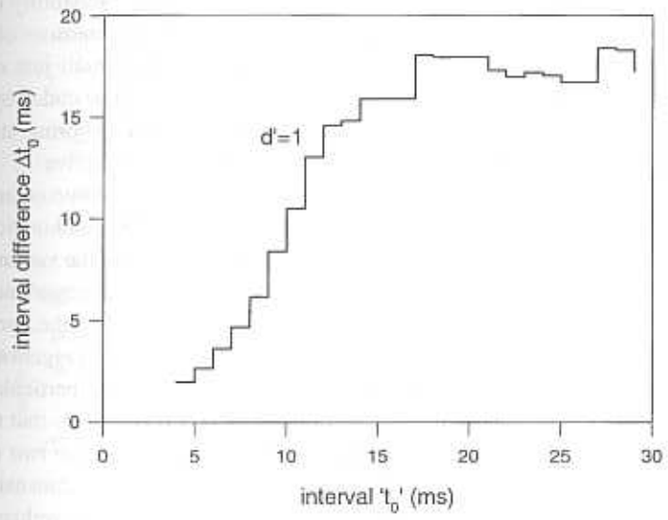



**Figure 2.14**
Discrimination between response-conditional ensembles as a function of the interval lengths. The ability to distinguish two signals drawn from different probability distributions is related to the overlap of these distributions. If the distributions are Gaussian, the discriminability can be quantified by a parameter $d'$ (Green and Swets 1966); see also Fig. 4.19. The contour line gives the value of $\Delta t_0$ as a function of '$t_0$', for which the response-conditional ensembles corresponding to spike intervals '$t_0$' and '$t_0 + \Delta t_0$' can be discriminated with $d' = 1$. As the interval $t_0$ is increased, the size of the increment $\Delta t_0$ at which the two intervals are distinguishable initially increases and then levels off at about 17 ms.

by the interpolated contour line of the values of $\Delta t_0 = t_0[R_2] - t_0[R_1]$ where $d' = 1$, as a function of $t_0[R_1]$.

The conclusion from Fig. 2.14 is that, once a spike is fired, the precision with which the observer must remember its position should be high shortly after the spike has occurred. As time goes on and no subsequent spike is fired, high time resolution is less crucial. After about 17 ms, the mutual timing of the two spikes is no longer important, and they can be considered as representing independent events. In this case the observer may forget that there was a spike, the only salient feature of the history being that no spike was fired during the past 17 ms. Because the precision required for optimal information extraction is not terribly great, we may say that the neural code is substantially robust to timing errors, even in a single neuron.

Observation of a particular spike train narrows the probability distribution in certain directions, as represented in the covariance matrices of Fig. 2.13. There must be other spike trains that stand for the stimuli just outside this narrowed distribution. These neighboring signals could be coded by any spike sequences, but a notion of smoothness such that neighboring stimuli were represented by similar sequences of spikes would be attractive.

We recall that covariance matrices can be decomposed into eigenvalues and associated eigenvectors. The eigenvectors determine the combinations of stimuli that vary independently, and the eigenvalues measure the variances in each of these independent directions. The eigenvalues of the covariance matrices in response conditional ensembles are almost all equal to the corresponding values in the a priori ensemble: There are just one or two eigenvalues which have been reduced significantly by virtue of observing a particular response $R_i$ (de Ruyter van Steveninck and Bialek 1988). This implies that the narrowing of the probability distribution is confined to just one or two dimensions in the space of all possible stimulus waveforms, and these dimensions are defined by the eigenvectors associated with the reduced eigenvalues. To see if coding obeys a notion of smoothness, we have to check that similar sequences of spikes code for signals that differ only along these one or two dimensions. This is in fact the case, as shown in Fig. 2.15.

The probabilistic methods described above provide precise data on the information conveyed by particular spike sequences $R$. For our homunculus this means that we have solved part of the problem: If asked to interpret a short sequence of spikes, the homunculus can refer to the response-conditional ensembles as a literal dictionary for translation from the language of spikes back into the language of sensory signals. But we require that the homunculus give a *running* commentary, which requires combining information from succes-

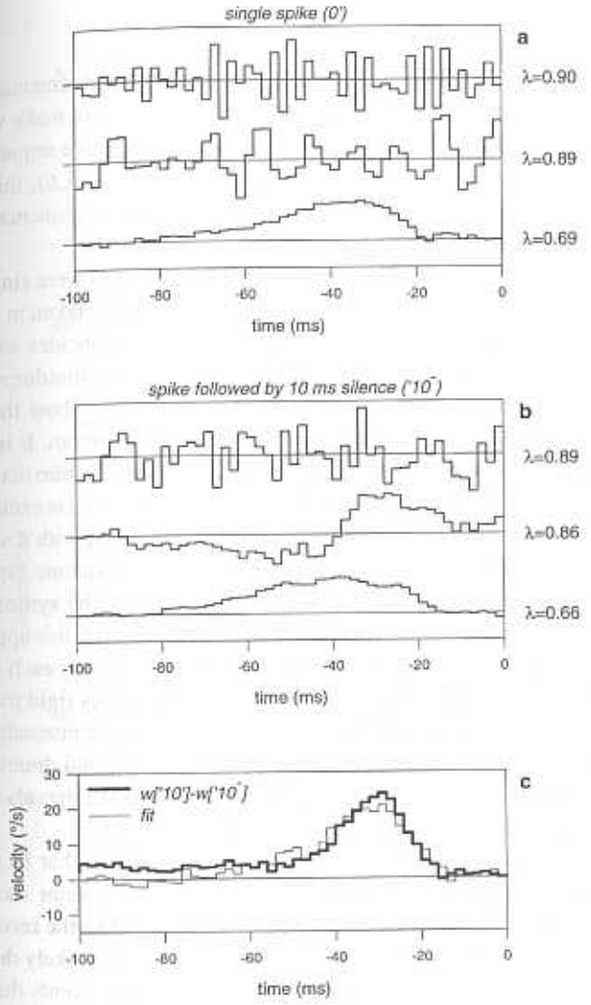



**Figure 2.15**
Eigenvectors and their eigenvalues for the three lowest eigenvalues of the covariance matrix for a single spike (a) and a single spike followed by a 10 ms empty interval (b). For the single spike there appears to be only one significant eigenvector: those with $\lambda = 0.89$ and $\lambda = 0.90$ are fluctuating too fast to be meaningful. When a single spike is followed by a 10 ms empty interval, a second smooth eigenvector develops with $\lambda = 0.86$. The fat line in (c) shows the difference in average waveforms for a 10 ms closed interval, $w['10']$, and a 10 ms open-ended interval, $w['10^-]$. The thin line is a fit of the first two eigenvectors in (b) to this difference waveform. Various similar cases corresponding to response categories "close" to $['10^-]$, such as $['5'5^-]$ and $['16^-]$, can be fitted with these two eigenvectors, which means that these eigenvectors serve as coordinates for measuring the changes in the conditional distributions for the various response categories.

sive sequences. The data presented here give no direct experimental guidance on how to perform this combination. As a first step, one can make what is admittedly a crude approximation, namely that successive spike sequences $R$ are generated independently. After some algebra (see section A.6), this assumption leads to a specific formula, Eq. (A.148), relating the sequences of $R_i$ to the best estimate of the stimulus waveform.

The response events $R_i$ analyzed in the H1 experiments were single spikes, closed intervals, and closed double intervals, with the latter taken in a nonoverlapping way so that the last spike of the previous event coincides with the first spike of the present event, and so on. Figure 2.16 presents stimulus reconstructions for two seconds of the experiment. The traces also show the stimulus waveform, and the spike sequences are depicted at the bottom. It is clear that the reconstruction based on single spikes has much less structure than the other two. This is related to the direction selectivity of H1: The cell is excited by motion in one direction and inhibited by motion in the other, with a small spike rate at zero velocity, so that the cell has much greater dynamic range for the encoding of positive velocities. The reconstructions can be symmetrized by considering an "anti-neuron" which sees the stimulus $-s(t)$; this approximates the situation in nature where the fly has two H1 cells, one on each side of the head, that are stimulated in antagonism as the fly undergoes rigid rotations (de Ruyter van Steveninck and Bialek 1988). It is perhaps more interesting that the reconstructions are much more symmetric with intervals and double intervals than with single spikes, in effect because the analysis of intervals allows the spaces between the spikes to represent negative velocities.

Another noteworthy property of the reconstructions is that in regions of high spike activity the reconstruction generally overestimates the stimulus. The overestimate becomes less pronounced, however, when the reconstruction depth increases to three spikes. The effect is therefore most likely due to serial correlation in the spike train, and the explanation is that events that occur at larger separation in time are less correlated. In other words, the assumption of statistical independence becomes a noticeably better approximation as the reconstruction depth increases. To be fair, it is not clear that a reconstruction depth of three spikes is sufficient to convincingly validate the independence hypothesis, but the quality of the reconstruction at this depth is already quite good. These results encourage us to think that a more systematic attempt at reading the neural code will be successful.

We should end this section with a note of caution. In each of the previous sections where we have explored some apparently systematic, quantitative approach to the neural code, we have faced the explosion of new structures at

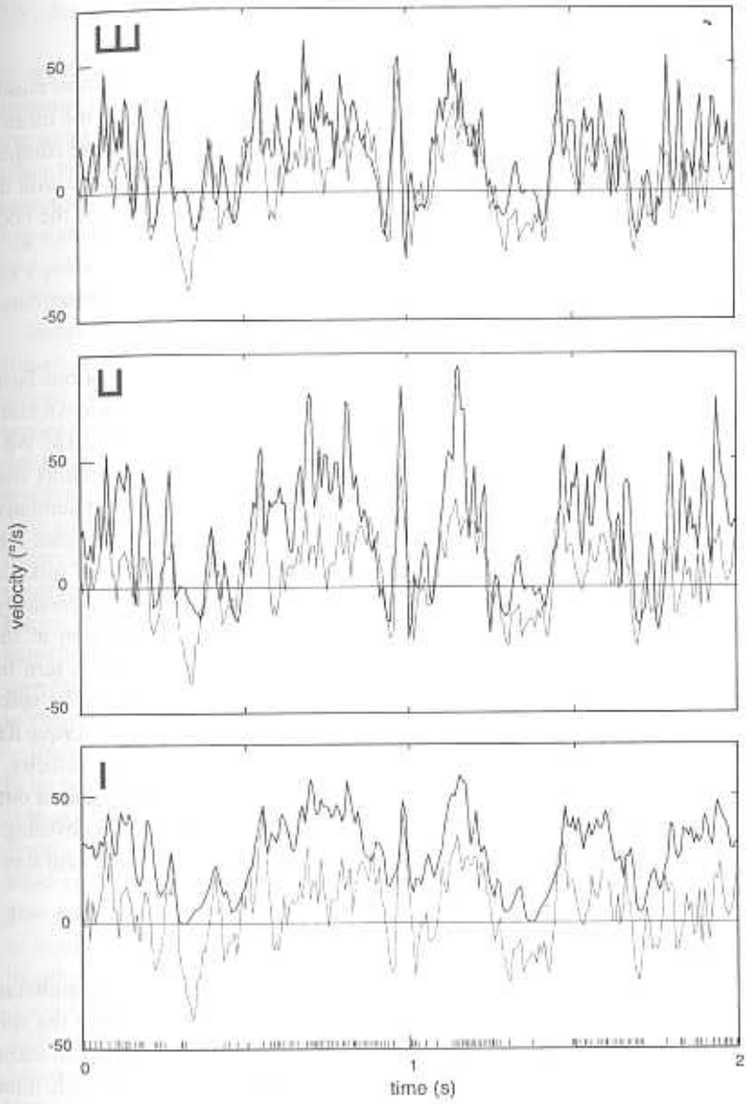



**Figure 2.16**
Reconstructions (dark lines) of angular velocity (thin line) using reconstruction depth of 1, 2, and 3 spike sequences (from bottom). Reconstructions using only single spike sequences (bottom) capture large fluctuations in the stimulus but miss many details. Including sequences of two spikes (middle) improves the reconstructions, but clearly the reconstructions systematically overestimate some aspects of the stimulus. These systematic errors are reduced in reconstructions based on triplets of spikes (top).

successive orders of approximation. The response-conditional ensembles seem to suffer the same difficulty: We see simple structures in the mean waveforms conditional on the occurrence of one spike, somewhat more complex structure conditional on two spikes, and still more complex structure with three spikes. To convince ourselves that we understand the structure of the code, we shall have to tame this complexity.

## 2.3 READING THE CODE

We have defined the problem of neural coding in terms of one basic question: Given the spike train $\{t_i\}$, what can we say about the unknown stimulus waveform $s(t)$? One possible answer is of the form in Fig. 2.16: We can "read" the spike train and translate back all the way to the stimulus itself. This is, if correct, a very simple answer, and it lends itself to quantitative analysis: How accurate are the reconstructed waveforms? How complex are the reconstruction algorithms? How do errors in the measurement of spike arrival times affect the reconstruction? Stimulus reconstruction is not necessarily a problem solved by the animal. It is, however, of the same character as the problems the animal must solve. For example, the fly can initiate a turn based on visual motion signals alone, which means that it translates the spike output of its motion sensitive visual neurons into a torque, and this torque has a component roughly proportional to the time dependent angular velocity. The torque signal is a continuous analog waveform that the fly synthesizes out of discrete spike sequences in its sensory neurons. The problem of recovering analog signals from the spike train is then a fundamental step in the neural processing of sensory data.

### 2.3.1 Why it might work

The major difficulties in decoding are the sparseness and randomness of each particular example of the spike train. In general, decoding the spike train requires learning to interpolate between the discrete spikes to estimate a continuous stimulus waveform, and it is not obvious that such interpolation is possible. Certainly the spike train does not determine the stimulus waveform uniquely under all conditions: Many different stimuli can produce the same spike train, and repeated presentations of identical stimuli do not produce identical spike trains, as in Fig. 2.1.

Formally, the question of whether decoding is possible concerns the structure of the conditional probability distribution $P[s(t)|\{t_i\}]$. If $P[s(t)|\{t_i\}]$ is sharply peaked at a particular stimulus waveform $\bar{s}(t;\{t_i\})$, then it makes sense



to say that the spike train $\{t_i\}$ "stands for" this stimulus, as in the discussion of response conditional ensembles. Furthermore, the width of this peak measures the accuracy with which the $\bar{s}(t;\{t_i\})$ approximates the true stimulus. On the other hand, if $P[s(t)|\{t_i\}]$ is a broad smear with no distinct peak, there is no reasonable sense in which one can decode the spike train to recover the analog signal $s(t)$. The intermediate case is where $P[s(t)|\{t_i\}]$ has several peaks or a ridge of maxima, so that estimates of $s(t)$ based on the spike train $\{t_i\}$ are ambiguous. These different possibilities are illustrated in Fig. 2.17.

Motivated in part by the experimental results on $P[s(t)|\{t_i\}]$ in the fly, Bialek and Zee (1990) formulated the problem of decoding in the context of simple models for the statistics of spike encoding, specifically the Poisson model described in section 2.1.4. In these models the problem of estimating the stimulus $s(t)$ given the spike train $\{t_i\}$ turns out to be equivalent to the problem of predicting the trajectory of a particle subject to an impulse each time a spike occurs. In the absence of spike train input, the particle undergoes Brownian motion, tracing out random trajectories drawn from $P[s(t)]$. These random trajectories are modified both by the spikes, which provide impulsive forces, and by a steady force related to the dependence of firing rate on the stimulus. If it is really possible to reconstruct the stimulus from the spike train, then this combination of steady and impulsive forces will cause the trajectories to cluster around the true stimulus waveform.

Because the connection between spikes and signals is probabilistic, we have to be more precise in defining the reconstruction problem. One natural choice is to ask for the most likely stimulus given the particular spike train, or, equivalently, the most likely trajectory given the force. This is maximum likelihood estimation; it generalizes the maximum likelihood decision rules which give the maximum percent correct performance in discrimination tasks such as those used in psychophysical experiments. For a discussion of maximum likelihood see sections 4.1.3 and A.16, as well as Green and Swets (1966). Another natural choice is to compute the average stimulus waveform given the spike train. This strategy is optimal in the sense that the mean square error ($\chi^2$) between the estimate and the true stimulus will be minimized. If the distribution of stimuli given the spike train is well-behaved, these different estimation strategies will all give very similar results. If, on the other hand, slight changes in our criterion for the best estimate produce large changes in the estimation algorithm, it is reasonable to say that robust stimulus estimation is not possible from the spike train alone. In the context of the model system, one can ask explicitly about the conditions for robust estimation, and these conditions are related once again to the shapes of the distribution $P[s(\tau)|\{t_i\}]$, as in Fig. 2.17.

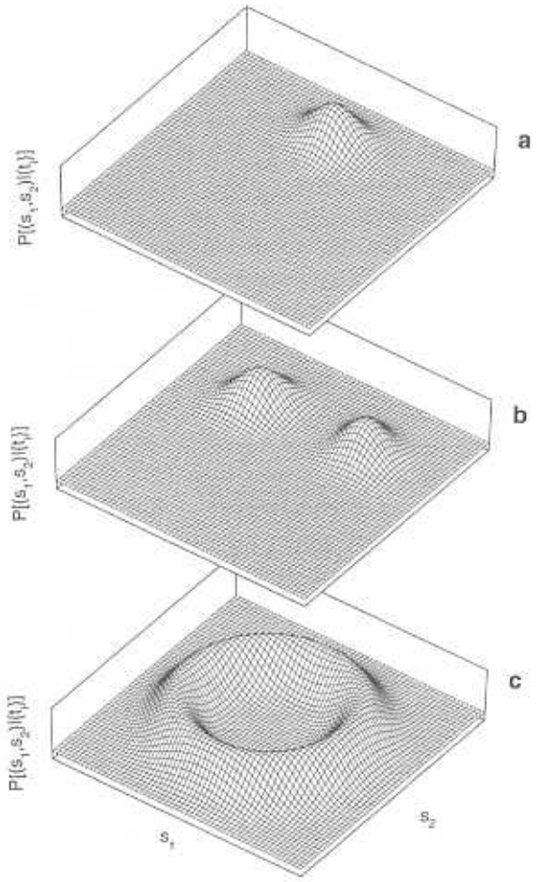



**Figure 2.17**
Structure of $P[s(\tau)|\{t_i\}]$. The success of direct decoding of the spike train depends critically on the structure of the conditional distribution $P[s(\tau)|\{t_i\}]$. If $P[s(\tau)|\{t_i\}]$ has a single well-resolved peak, as in (a), decoding should be possible; our estimate of the stimulus should be at the peak or near the peak of $P[s(\tau)|\{t_i\}]$, depending on the choice of metric. On the other hand, if $P[s(\tau)|\{t_i\}]$ has no discernible structure or does not have a single peak, as in (b) and (c), the best estimate is not well defined.



In a later section (see especially Fig.2.24 and the surrounding discussion) we will look at experiments that address this question more directly. For the moment we will focus on the best estimator in the $\chi^2$ sense, the conditional mean, and in section A.7 we review the connection between optimal estimation and the conditional mean.

Following our discussion of input/output analysis in section 2.1.3, we know that in many systems the average trajectory of a particle is linearly related to the applied forces, so that

$$\langle x(t)\rangle = \int d\tau K_1(\tau)F(t-\tau), \tag{2.23}$$

where $K(\tau)$ is the linear response function; for simplicity we assume that the average position in the absence of a force is zero, so we don't have to add a constant term to Eq. (2.23). In the analogy to spike encoding, the force consists of a series of pulses at the spike times $t_i$,

$$F(t) = \sum_{i=1}^{N} \delta(t - t_i), \tag{2.24}$$

and the trajectory $x(t)$ corresponds to the signal waveform $s(t)$; for a discussion of the delta functions see section A.1. Thus, linear response of the mechanical system corresponds to *linear reconstruction* of the signal from the spike train:

$$\begin{aligned} s_{\rm est}(t) &= \int d\tau K_1(\tau) \sum_{i=1}^{N} \delta(t - \tau - t_i) \\ &= \sum_{i=1}^{N} K_1(t - t_i). \end{aligned} \tag{2.25}$$

The meaning of this equation is shown schematically in Fig. 2.18. Rather than characterizing how the nervous system converts signals into spikes, Eq. (2.25) characterizes the process by which an observer of the spike train could estimate or reconstruct the stimulus. Thus we imagine building a "black box" that takes the spikes as input and returns the stimulus—or as close as we can get to the stimulus—as output.

Equation (2.25) says that the black box of Fig. 2.18 is approximately a linear device. More generally we might want to allow this box to have some nonlinearities, and this corresponds to the fact that the particle in the equivalent statistical mechanics problem has a nonlinear response to applied forces. Then the generalization of Eq. (2.25) is

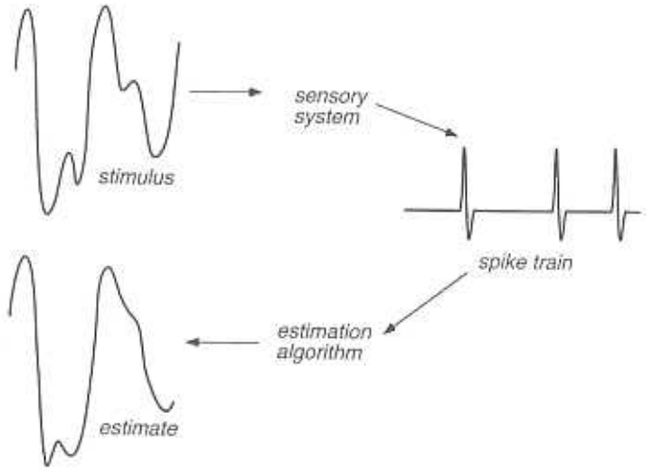



**Figure 2.18**
Schematic of stimulus estimation.

$$s_{\rm est}(t) = \langle s(t) \rangle = \sum_i K_1(t - t_i) + \frac{1}{2} \sum_{ij} K_2(t - t_i, t - t_j) + \cdots . \quad (2.26)$$

In the analysis of a model neuron it is possible to calculate all of the kernels $K_1, K_2, \ldots$, so that the procedure for decoding the spike trains of a model neuron is completely determined. We don't believe that these model neurons are exact descriptions of real neurons, so we need some more general understanding of what the expansion in Eq. (2.26) means.

To begin, the time constants that appear in the *decoding* filters $K_n$ are different from the time constants that characterize the *encoding* dynamics of the neuron, for example in the transformation from the stimulus waveform $s(t)$ to the time dependent firing rate $r(t)$. In fact, the structure of the kernels depends on the statistics of the input signals, so that the optimal strategies for reading the neural code depend on the nature of signals in the environment, even if the neuron does not adapt. These points emphasize the fact that the $K_n(t)$ are not properties of the neuron alone, but rather combined properties of the neuron and its sensory environment (Bialek and Zee 1990; Bialek 1990; Gabbiani and Koch 1996).

We have emphasized from the very beginning that, because of Bayes' rule, the meaning of a spike train must depend on the ensemble from which sensory



stimuli are drawn—we interpret what we "hear" from the neuron in light of what we expect. This is a mathematical fact that we can choose not to emphasize, but it will not go away. It is possible that the spike train provides an invariant representation of the stimulus waveform, or of certain features in the stimulus waveform, but this requires that the neuron adapt its computations and coding strategies to changes in the stimulus ensemble. Adaptation certainly occurs, but whether this process provides an invariant dictionary for the translation of spike sequences remains to determined. The reconstruction problem forces us to address explicitly the issues of context dependence and adaptation, because the reconstruction filters $K_n$ can be calculated independently in each new context.

Another way of thinking about the filter $K_1(\tau)$ is that it serves to separate the best estimate of the signal $s(t)$ from the randomness or noise inherent in the spike train. Spike trains contain power at high frequencies, corresponding to the time resolution with which we (as observers) can localize the spikes. But these high frequency data may just be noise, uncorrelated with the sensory input. On the other hand, the low frequency variations of the spike rate may be perfectly locked to the low frequency components of the signal. In this case we would like to attenuate the high frequency noise and enhance those components of the spike train that are strongly correlated with the signal. The filter $K_1(\tau)$ provides the best way of making this separation between signal and noise.

What would it mean to say that the reconstruction of the stimulus according to Eq. (2.26), or even more simply according to Eq. (2.25), is successful? It is important to realize that this procedure does not work automatically. For example, attempts to reconstruct the sound pressure waveform from the spike trains of an auditory afferent tuned to high frequencies will undoubtedly fail; high frequency auditory neurons are not sensitive to the absolute phase of the acoustic waveform, although they can code phase modulations. But if the absolute phase is irrelevant, then the cell gives the same response to the signal $s(t)$ and to the signal $-s(t)$, and the output of the cell cannot tell us which of these waveforms actually occurred. This is like the situations shown in Figs. 2.17b and c, where the spike train may be very informative about some aspects of the stimulus but nonetheless ambiguous, preventing the reconstruction of the waveform. In this case of the auditory neuron, it is usually thought that the cell encodes the envelope of the sound pressure waveform. Attempts to reconstruct this envelope were in fact successful, although the choice of a proper definition for the envelope poses interesting questions (Rieke et al. 1992).

The sort of qualitative failure encountered for auditory neurons is easy to identify. The more subtle question is whether the reconstruction algorithm in Eq. (2.26) is just another systematic expansion that fails to converge quickly enough to be useful. It turns out, for example, that so long as the first Wiener kernel of the neuron is nonzero, then the linear reconstruction filter $K_1$ will also be nonzero. This makes us worry that linear reconstruction is just linear response looked at from a different point of view. This is not correct, and in fact the two expansions are very different. The fact that both tools are related to ideas developed by Wiener only adds to the confusion.

If you are given observations of one signal $y(t)$ and try to estimate some other signal $x(t)$, there is a rigorous theory due to Kolmogoroff (1939; Kolmogorov 1941) and Wiener (1949), of how to choose a linear filter that results in the best possible estimate. In our case we observe the signal $\rho(t) = \sum_i \delta(t - t_i)$ and try to estimate $s(t)$, and the procedure in Eq. (2.25) is linear filtering, so perhaps we could dispense with all the discussion and say that we are trying to build the Kolmogoroff-Wiener filter that estimates the signal from the spike train.[3]

The Kolmogoroff-Wiener results (and their exegesis in subsequent literature) essentially solve any estimation problem that can be solved by linear filtering. But why should linear filtering work? More specifically, why should estimates based on linear filters be any good, and shouldn't we be able to do much better with a more complex nonlinear procedure? In the present context, neurons can be highly nonlinear devices, so that if we tried to expand the firing rate in powers of the stimulus, low order terms would *not* be sufficient under natural conditions. This is part of the problem in applying the Wiener or Volterra methods to real neurons. But our expansion is very different—rather than describing the input/output relation of the neuron, we are trying to describe a hypothetical black box that takes the spike train as input and returns an estimate of the stimulus waveform $s(t)$.

To get a feeling for the difference between the estimation problem and the more conventional input/output analysis, consider a simple detector whose average output $x$ is proportional to the input $s$. We can divide out the proportionality constant and write

3. The references cited here illustrate the ambiguities in translation between different symbol systems, as discussed for the translation between spike trains and sensory stimuli. We notice that Колмогоров can be mapped into Kolmogoroff *or* into Kolmogorov. This particular ambiguity is resolvable by reference to place cells (O'Keefe and Nadel 1978), or more specifically to nation cells.



$$x = s + \eta, \tag{2.27}$$

where $\eta$ is the noise, and we define the average of the noise to be zero ($\langle \eta \rangle = 0$). The essence of the problem can be understood without worrying about the time dependence, so in Eq. (2.27) the various quantities are just real numbers rather than functions of time. Characterizing the input/output relation of this system is trivial—on average, the output is equal to the input. Is the estimation problem equally trivial? More precisely, what does it mean to say that linear reconstruction of $s$ from $x$ will work?

Everything that we know about the signal $s$ by virtue of observing $x$ is contained in the conditional distribution $P(s|x)$. From Bayes' rule in Eq. (2.4) and Fig. 2.2, we can write

$$P(s|x) = \frac{P(x|s)P(s)}{P(x)}. \tag{2.28}$$

Given the signal $s$, the probability of observing a particular value for $x = s + \eta$ just depends on the distribution of the noise $\eta$, so we can write

$$P(x|s) = P_{\text{noise}}(\eta = x - s), \tag{2.29}$$

and hence

$$P(s|x) = \frac{1}{P(x)} P_{\text{noise}}(\eta = x - s) P(s). \tag{2.30}$$

We are interested in estimating $s$, and we know the mean square errors in our estimate will be smallest if we use as our estimate the conditional mean (see section A.7), that is,

$$s_{\text{est}} = \int ds\, P(s|x)\, s. \tag{2.31}$$

Our job is to evaluate this integral.

Suppose that, as is often the case, the noise $\eta$ is chosen from a Gaussian distribution. Then

$$P_{\text{noise}}(\eta) = \frac{1}{\sqrt{2\pi \langle \eta^2 \rangle}} \exp\left[ -\frac{\eta^2}{2\langle \eta^2 \rangle} \right], \tag{2.32}$$

and hence from Eq. (2.30) we have

$$P(s|x) = \frac{1}{P(x)} P(s) \frac{1}{\sqrt{2\pi \langle \eta^2 \rangle}} \exp\left[ -\frac{(s - x)^2}{2\langle \eta^2 \rangle} \right]. \tag{2.33}$$

In the simple case that the signal $s$ is also drawn from a Gaussian distribution,

$$P(s) = \frac{1}{\sqrt{2\pi\langle s^2\rangle}} \exp\left[-\frac{s^2}{2\langle s^2\rangle}\right], \tag{2.34}$$

we have

$$\begin{aligned} P(s|x) &= \frac{1}{P(x)} \frac{1}{2\pi\sqrt{\langle s^2\rangle\langle \eta^2\rangle}} \exp\left[-\frac{s^2}{2\langle s^2\rangle}\right] \exp\left[-\frac{(s-x)^2}{2\langle \eta^2\rangle}\right] \\ &= \frac{1}{Z(x)} \exp\left[-\frac{1}{2}s^2\left(\frac{1}{\langle s^2\rangle} + \frac{1}{\langle \eta^2\rangle}\right) + s\left(\frac{x}{\langle \eta^2\rangle}\right)\right], \end{aligned} \tag{2.35}$$

where $Z(x)$ is a normalization constant which is independent of the signal $s$.

Thus, if we observe Gaussian signals in a Gaussian noise background, the conditional distribution of signal given our data is also Gaussian. The conditional mean is the same as the most likely value, and it is easy to find by solving the equation $\partial P(s|x)/\partial s = 0$. The result is that our best estimate of the signal is $s_{\rm est} = K_1 x$. The kernel $K_1 = SNR/(SNR+1)$, where the signal-to-noise ratio is just the ratio of variances: $SNR = \langle s^2\rangle/\langle \eta^2\rangle$. This is a reasonable result, in that the optimal decoding of the output of a linear detector uses a linear system whose gain depends on the $SNR$. The dependence of the gain on the $SNR$ comes about because we have both prior knowledge about the input signals and specific knowledge from observation of the output of the detector. At high $SNR$ the detector output is reliable and the gain approaches unity. At low $SNR$ most of what we are seeing at the output of the detector is noise, and we scale down the detector output, relying more heavily on our prior knowledge. Thus the dependence of our decoding strategy on both the detector output and our prior knowledge causes us to underestimate the signal systematically.

The success of linear estimation in the case of Gaussian signals does not generalize. Thus, if we imagine that signals are chosen from an exponential distribution,

$$P(s) = (s_0/2)\exp(-|s|/s_0), \tag{2.36}$$

then the most likely value of the signal is actually a thresholded function of the detector output! The threshold appears at a point $x_0 = \langle \eta^2\rangle/s_0$ that depends on the typical values of the signal and noise, and if $x < x_0$ the most likely value of $s$ is $s = 0$, no matter what the precise value of $x$. This extreme nonlinearity is softened a bit if we ask for the estimator that minimizes $\chi^2$, but there are still



very different behaviors of the optimal estimator above and below the nominal threshold (Potters and Bialek 1994).

The example of the exponential distribution shows us that, even for a completely linear detector, the decoding problem can have a dramatically nonlinear solution. In this case linearity is destroyed by changing the distribution of input signals, but one can just as well change the statistics of the noise. Non-Gaussian noise alone is sufficient to destroy linear decodability. Of course, if the signal to noise ratio is very high, the distribution $P(x|s)$ approaches a delta function and hence its exact shape (the distribution of the noise) is irrelevant, but we know that biological detectors seldom operate in this limit.

This simple example shows us that the possibility of decoding the spike train with a linear filter such as $K_1$ in Eq. (2.30) really has nothing to do with the conventional notion of linearity in the input/output of the neuron, or the linearity of the relationship between the stimulus and the firing rate. It is even possible that the cell is linear by the conventional measures, but because the noise in the cell's response is non-Gaussian, linear decoding won't work. In this case one could *try* to reconstruct the stimulus waveform with a linear filter, and this procedure would produce some answer, but one could do much better by keeping (perhaps many) more nonlinear terms in an expansion like Eq. (2.26).

In the conventional input/output relation, we know how to test for linearity or nonlinearity. In particular, if we are not worried about time dependence we can make a plot of neural output versus sensory input, and ask if this is a straight line. This is what we did, for H1, in Fig. 2.2i, where the neural output is defined by the number of spikes in a 200 ms window and the sensory input is defined by the velocity of motion averaged over this window; clearly the input/output relation is *nonlinear* in the range of velocities used for this experiment. Can we make a similar plot for the decoding problem? We want to plot our best estimate of the stimulus versus the neural output, and we know that the best estimate—the estimate that will minimize our mean square error—is the mean value of the stimulus conditional on the observed neural output. This conditional mean is plotted in Fig. 2.2h, and we see that it is a *linear* function of the spike count throughout the observed range of spike counts. We emphasize that this plot is an explicit construction of the optimal decoder for the restricted problem of inputs and outputs averaged over a 200 ms window, and we see that nonlinear encoding of the velocity signal into firing rate coexists with linear decodability. As in our mathematical examples, the linearity of the decoding problem is simply a different question from the

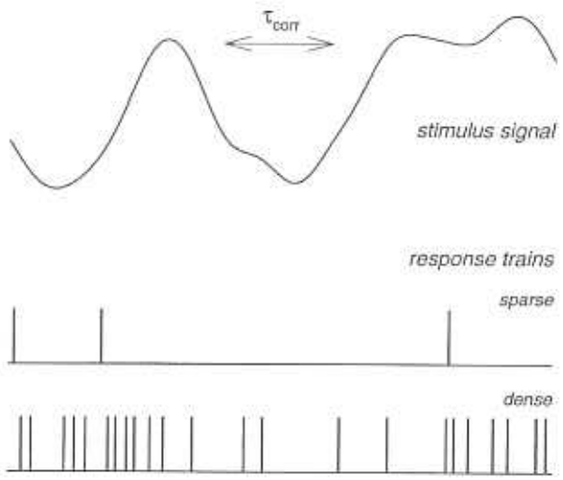



**Figure 2.19**
Estimation in sparse and dense spike trains. An important factor determining the success of the estimation process of Fig. 2.18 is the mean interval between spikes relative to the correlation time of the input signal. If the spikes are sparse, as in the top spike train, the stimulus correlation time divided by the mean interval between spikes provides a small parameter which can be used to construct a systematic approach to estimation, as described in more detail in the text. If, as in the lower response train, the number of spikes per correlation time becomes of order one, or bigger, this condition is not fulfilled and a perturbative approach to reconstruction is not feasible.

linearity of the encoding problem, and Fig. 2.2h is a direct demonstration that spike trains are linearly decodable.[4]

Decoding strategies depend on the correlation time $\tau_c$ of the signal in relation to the typical interspike intervals, as shown schematically in Fig. 2.19. The correlation time is the time over which one can predict the signal from knowledge of its past behavior. When signals are filtered by the nervous system, the correlation time can be changed, so we should think about the correlation time measured after neural filtering and just before spike generation. The occurrence of a single spike tells us something about the signal at the moment of spike generation—some voltage is equal to its threshold value. We

4. The linearity in Fig. 2.2h can be derived as a consequence of linearity in the reconstruction algorithm of Eq. 2.25 by averaging both sides of the equation over a time window larger than that spanned by the features of $K_1(t)$. Then the slope of the line in Fig. 2.2h is the average of $K_1(t)$.



can extrapolate this knowledge out to a time window of roughly $\pm\tau_c$ around the spike itself. If the next spike always occurs much later, with an interval $\tau \gg \tau_c$, then this next spike gives us independent information about the signal, and the contributions of the two spikes to our estimate of the stimulus waveform must just add. This suggests that our expansion of the optimal estimate in Eq. (2.26) is really an expansion in the average number of spikes per correlation time, or $\langle r \rangle \tau_c$. In the context of simple models, this identification of the expansion parameter can be verified by detailed calculations.

The description of neurons as having input/output relations as in Fig. 2.2i or 1.4 completely misses the fact that the spike train is a discrete sequence of events while the sensory stimulus is a continuously varying function of time. Under these conditions, estimating the stimulus cannot be a simple matter of inverting the input/output relation—what do we do in the spaces between the spikes? As we hinted in the discussion of the homunculus (section 1.2), giving meaning to the spike train involves generating a "running commentary," and clearly this requires extrapolation from one spike to the next. The structure of the best decoding algorithm is really controlled by the nature of this extrapolation, and not by the conventional input/output relation.

In our review of input/output relations, we emphasized that any series expansion approach, like the Wiener or Volterra expansion, will work (in practice) only if we can identify some small parameter that enforces the rapid convergence of the series. Unfortunately, for the study of encoding it is often difficult to identify such a small parameter. But, for decoding, we have a new possibility, namely that there are a small number of spikes per correlation time of the stimulus. Rather than trying to classify neurons as linear or nonlinear, we are led to ask which spike trains are linearly decodable.

Linear decodability defines a regime of neural dynamics in which each significant variation in the signal (on time scale $\tau_c$) triggers of order one spike or less. This is almost the opposite picture from that suggested in rate coding models, where information must be carried in windows of time that contain several spikes, enough to form a reasonable estimate of the firing rate over the window. Is there any evidence concerning the value of $\langle r \rangle \tau_c$? We have reviewed the evidence that at least some systems operate in a regime where, under natural conditions, roughly one spike is fired for each characteristic time of the signal, and we might thus expect that spike trains in such systems are linearly decodable.

In their early applications of white noise methods to the auditory system, de Boer and Kuyper (1968) emphasized the interpretation of the reverse correlation function—the mean stimulus that triggers a spike—as the feature of

the stimulus waveform that is signaled by the occurrence of a spike. This is much closer to the organism's point of view than in the more common use of white noise methods for system identification. Following these ideas, Gielen, Hesselmans, and Johannesma (1988) proposed that one could estimate the stimulus waveform as in Eq. (2.25), with the kernel $K_1(\tau)$ identified as the reverse correlation function. One can show that the reverse correlation function, which is the mean stimulus waveform given the occurrence of a single spike, is in fact the best estimate of the signal if all we know is that a single spike was fired (de Ruyter van Steveninck and Bialek 1988), and that the linear reconstruction filter converges to the reverse correlation function in the limit that firing rates go to zero, as emphasized by Gabbiani and Koch (1996). But, as the spikes come more frequently, the reverse correlation functions centered on successive spikes overlap and can provide conflicting estimates of the stimulus (Fig. 2.3). The correct resolution of these conflicts requires that we attach measures of confidence to the different estimates; this is accomplished by measuring the relevant probability distributions $P[s(t)|\{t_i\}]$, as described in section 2.2.3. It is a remarkable fact that, at least in the study of model neurons, one again arrives at an optimal estimator of the form of Eq. (2.25), but now the kernel $K_1(\tau)$ is determined not only by the filter characteristics of the neuron (the reverse correlation function) but also by the characteristics of the stimulus ensemble.

### 2.3.2 An experimental strategy

We have formulated the problem of reading the neural code as the construction of a (generally nonlinear) filter that takes the spikes as input and produces an estimate of the sensory stimulus, as in Eq. (2.26). In the context of models for the spike initiation process, this filter can be related to the parameters of the model, but we do not want to take such details of the models too seriously. Instead we want to take the general idea of decoding and use it as a tool for the design and analysis of experiments. Thus we would like to find an empirical approach to choosing the kernels $K_1(\tau)$, $K_2(\tau, \tau')$, $\cdots$, given the experimental stimulus $s(t)$ and the measured spike times $\{t_i\}$.

In the analysis of model neurons, the kernels were computed (for example) as those that minimize the mean square error between the estimate and the true stimulus waveform. More generally, we choose some error function $E[s(t), s_{\rm est}(t)]$ that determines how well our estimate describes the stimulus. We can then vary the kernels $\{K_n\}$ to minimize $E$. Ideally we want to minimize the *expectation value* of the error measure in the stimulus ensemble, but



in a long experiment we can try to minimize the time average error, using the idea of ergodicity discussed in section 2.1.3. Clearly we must be careful in this procedure, because there is a significant chance of overfitting the kernels to a limited data set.

What should we use as an error function? If we choose a quadratic error function, such as the mean square error, we can make considerable analytical progress in determining the best kernels. Thus we consider error functions of the form

$$E[s(t), s_{\rm est}(t)] = \langle |s(t) - s_{\rm est}(t)|^2 G[s(t)] \rangle, \tag{2.37}$$

where the average $\langle \cdots \rangle$ is over the repeated examples of the stimulus from the distribution $P[s]$, and $G[s]$ is a positive functional of $s$ (e.g., $G[s] = 1, |s|, s^2, \ldots$). We begin by choosing $G[s] = 1$, so that $E$ is the conventional mean square error, or $\chi^2$, between the stimulus and the estimate. Other choices of $G$ will impose heavier penalties for errors at large values of the stimulus; in section 2.3.3 we discuss how different metrics affect our decoding strategies.

A second issue in choosing the estimation kernels is causality. The motivation for this particular approach to studying the neural code came largely from thinking about what information is available to the organism, or to our homunculus, from a single example of the spike train. To extract this information in real time—not by recording the spike train on tape and coming back the next day to analyze it—our estimation strategy must be causal. Causality tells us that a spike cannot influence our estimate of the stimulus until the spike has occurred; this means that the kernels must be zero for negative times, e.g., $K_1(\tau < 0) = 0$. Causality *does not*, however, mean that a spike which occurred in the past cannot influence our present estimate of the stimulus; if the stimulus has a finite correlation time, knowledge of the recent history of the stimulus can contribute to the present estimate. There is also a strict causal relation between the stimulus and the spike times, in that a spike occurring at $t = 0$ is generated only by the preceding stimulus, $s(t < 0)$. But now we have a problem: A spike influences our estimate of the stimulus only *after* the spike occurs, but the stimulus influences the generation of the spike only *before* the spike occurs. The way out of this problem is to accept a delay in the estimation. Causal estimation necessarily introduces delays; the magnitude of the delay depends on the structure of the code, a point to which we will return.

We now have a well posed mathematical problem: find the kernels $K_n$ that minimize the error measure in Eq. (2.37), while obeying the constraint of

causality. This is rather like trying to make a least squares fit of some functional relation between two variables. Here the two variables are a bit more complex, since they represent the spike train and a continuous function of time, but the idea is the same. All of the mathematical details are given in section A.8. Some general concerns in problems of this type are that our fitting procedure might get "stuck" in local minima of the error measure, not finding the true best fit, or that the best fit itself may be too sensitive to noise in the observed data. In the experiments we review here, there are several circumstances working in our favor to help us avoid these problems. First, the error measure $E$ is quadratic in the kernels $K_n$, which means that the optimization of the filters is not a complex problem and there are no local minima. Second, the data sets for these experiments are quite large, consisting in some cases of $10^5$ spikes. Finally, as an added precaution, we can learn the filters on the bulk of the experimental data and then test the quality of the estimates on sections of data that did not contribute to the filter calculation, so we cannot "overfit" to the details of one data set and fool ourselves about the quality of the resulting reconstructions.

In the case of the H1 cell, we do have one more problem with any attempt to reconstruct the velocity waveform, as explained in connection with Fig. 2.16. Because H1 is selective for the direction of motion, it has a much larger dynamic range for the encoding of positive (back to front across the eye) motion than for negative velocities. We expect that velocity estimates based on the output of this one cell will be biased, or at least will have errors with a magnitude that depends strongly on the sign of the stimulus.

The fly solves the problem of asymmetric coding by having two H1 cells, one on each side of the head, that are stimulated in antagonism when the fly makes rigid rotational motions. It would appear that the two cells are identical except for the sign of their direction selectivity—both are excited by back to front motion, but on one side of the head this is left to right and on the other side it is right to left. One can create an experimental simulation of this effect by recording the response of one H1 cell to the velocity stimulus $s(t)$, and to the stimulus $-s(t)$ during a different part of the experiment. We then imagine that the responses to $s(t)$ and to $-s(t)$ correspond to the responses of two cells, one with positive and one with negative direction selectivity.

Information coded in the spike trains corresponding to the two polarities of the stimulus can be combined to give a reconstruction based on two "virtual" neurons. Since the two virtual cells are identical except for the sign of their direction selectivity, we insist that the equations for the reconstructed waveform be (anti)symmetric in the two spike trains:



$$s_{\rm est}(\tau) = \sum_i \left[K_1(t-t_i^+) - K_1(t-t_i^-)\right] + \sum_{i,j}\left[K_2(t-t_i^+, t-t_j^+) - K_2(t-t_i^-, t-t_j^-)\right] + \sum_{i,j} K_2'(t-t_i^+, t-t_j^-) + \cdots, \tag{2.38}$$

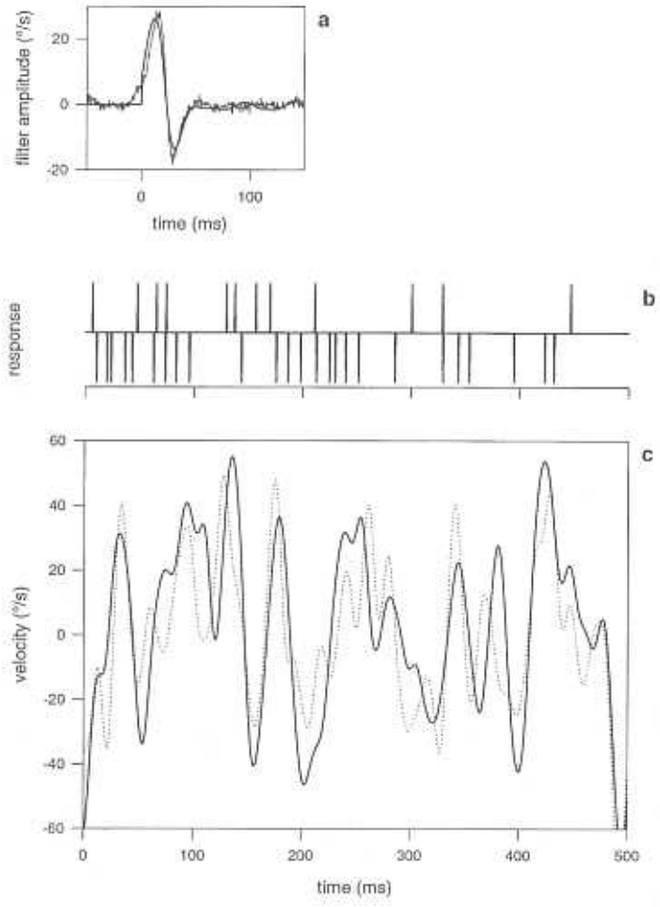
where $\{t_i^+\}$ and $\{t_i^-\}$ are the spike occurrence times in the response to $s(t)$ and $-s(t)$, respectively. $K_2'$ is a second kernel with contributions from one spike from each virtual neuron, which means that we allow (for example) coincident firing of the two cells to have special significance.

We emphasize that this picture of two virtual neurons is not essential to the idea of stimulus reconstruction. It is included here largely because this is how the original experiments (Bialek et al. 1990, 1991) were analyzed. Similar arguments were made in the analysis of motion discrimination by cells in monkey cortical area MT (Britten et al. 1992), as discussed in section 4.1.4. In each case it would be more satisfying to have an analysis based directly on simultaneous recordings from cells with opposite direction selectivity, but we think that none of the conclusions from either the fly or the monkey work will change significantly once this is done.

### 2.3.3 Qualitative features of a first test

Here we explore some of the qualitative features of the H1 experiments, which were the first test of the decoding strategies described in the previous section (Bialek et al. 1990, 1991). The linear filters $K_1(\tau)$ obtained in this experiment are shown in Fig. 2.20a, and the reconstructions using these filters are shown in Fig. 2.20c. We see that the optimal filters integrate over time intervals on the order of 30–40 ms. Since behavioral decision times in the fly are also on the order of 30 ms, the structure of the code appears to be well matched to the behavioral decision making process. Very few spikes contribute to the reconstruction at any given time—which we have emphasized *must* be true from the behavioral data.

Another way of measuring the width of the filter is to note that it significantly attenuates frequencies greater than 25 Hz. This has a number of consequences: First, we expect that the signal to noise ratio of the reconstructions will peak at frequencies below 25 Hz. Second, we may be systematically underestimating the stimulus at high frequencies. Finally, the code should be relatively robust to timing errors on the order of a few milliseconds, because




such errors will introduce only high frequency noise in the spike train, and this high frequency noise will be attenuated by the filter $K_1(\tau)$.

As discussed in section 2.3.1, there are theoretical reasons for believing that the reconstruction series, Eq. (2.3.2), should be dominated by the first term or the first few terms. Furthermore, this linearity of reconstruction is distinct from linearity of response. To test these ideas, we first determine that H1 is not simply responding linearly to the variations in angular velocity. In Fig. 2.21 we show the firing rate as a function of time (as in Fig. 2.1), and we compare this rate with the predictions from filtering the stimulus through



**Figure 2.20**
Estimation for H1 experiment. Estimation filters are shown in (a). The noisier trace is the estimation filter calculated directly from the stimulus and spike train. The filter has been shifted by 40 ms to produce a causal estimation procedure. The smooth trace is the estimation filter calculated by expansion in a basis set of causal functions with an estimation delay of 40 ms. Details of each calculation are described in the text and section A.8. Spike responses to a short section of the stimulus (dashed line in (c)) are shown in (b); upward spikes are responses to this stimulus. Downward spikes are responses to the same stimulus, but with a change of sign of the velocity. As there are two H1 cells (one for each eye) with mirror symmetric directional selectivities, the sign flipped stimulus induces a response in H1 typical for the contralateral H1 cell when stimulated with the original velocity waveform. We use both spike trains in the reconstruction to symmetrize the procedure, approximating the movement signal seen by the contralateral H1 cell. The estimate (solid line in (c)) is constructed by convolving the filter in figure (a) (in this case the acausal shifted filter) with the spike trains. Stimulus and estimate have been smoothed with a Gaussian filter with a standard deviation of 5 ms.

the first Wiener kernel. The true rate exhibits much larger and more rapid variations than expected from the linear model. This corresponds to the fact that individual spikes or small clusters of spikes are produced at rather precise times in relation to particular variations of the stimulus waveform, so that the Gaussian distribution of input signals is transformed into a very nonGaussian distribution of rates $r(t)$. The crucial point for our discussion, however, is that the experiment probes H1 with stimuli that drive it out of the linear response regime.

Given that H1 is responding nonlinearly, can we decode the spikes with a linear filter? Systematic errors in the reconstructions, especially at high velocities, would indicate that nonlinear terms might be important. One check for systematic errors is to plot the average stimulus given a particular value of the reconstruction against the reconstruction value itself. Saturation effects should show up in this plot as deviations from a straight line at high stimulus levels. In Fig. 2.22 we see that these effects do not occur.

Another way of checking the success of linear reconstruction is to add in the next terms of the expansion in Eq. (2.26), ostensibly to see if these additional terms improve the quality of the reconstruction (Rieke 1991). The short answer is that the inclusion of nonlinearities does not make a statistically significant change in $\chi^2$. But perhaps $\chi^2$ is too crude a measure of the quality of the reconstruction, and nonlinear terms may make a more subtle difference. In section 3.2.3 we discuss a more complete method of analyzing the quality of reconstructions, defining an effective noise level at each frequency. This

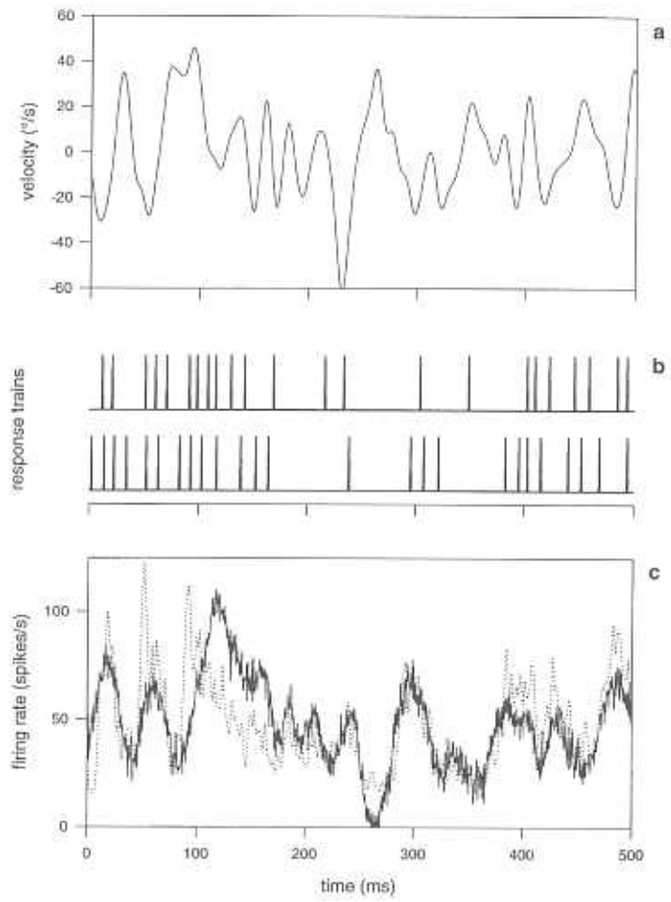



**Figure 2.21**
Stimulus, firing rate, and first order estimate of rate from H1 experiment. Panel (a) shows a short section of the angular velocity stimulus. This stimulus was repeated 100 times, producing 100 spike responses, two of which are shown in (b). From these spike responses we measure the time dependent firing rate (dashed line in (c)), as in Fig. 2.1. The solid line in (c) shows the predicted firing rate from the first term in the Wiener expansion. The first order estimate of the rate captures the slow modulations in firing rate, but fails to capture fast modulations in the rate.



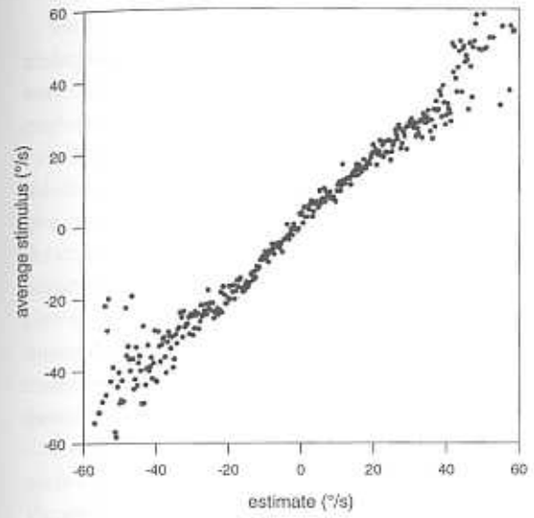



**Figure 2.22**
Average stimulus conditional upon the estimate plotted as a function of the value of the estimate, for the reconstruction in Fig. 2.20. Systematic errors in the estimate (especially saturation effects) should show up as deviations from a straight line. The lack of such deviations suggests that nonlinear terms do not contribute significantly to the estimation process.

effective noise level also does not seem to be improved by adding nonlinear terms to the reconstruction procedure, and we shall see that over a range of frequencies this noise level approaches the limit imposed by the signal and noise properties of the photoreceptor inputs to the motion computation (section 4.3). We will be able to make a similar argument in the analysis of primary sensory neurons, where it turns out that the information (in bits) provided by the linear reconstruction approaches the physical limits imposed by the statistics of the spike train itself (section 3.3). In both of these cases there is a regime in which the linear reconstructions are as good as possible, so they could not, even in principle, be improved significantly by the addition of nonlinear terms.

The fly is faced with an interesting dilemma, perhaps typical of sensory signal processing. Behavioral considerations push for short decision times, but short times mean that the system is more susceptible to noise at each stage of processing. To explore the relation between reliability and decision times, one can find the causal reconstruction at various delay times. To test the quality of these different reconstructions, we calculate the crosscorrelation of

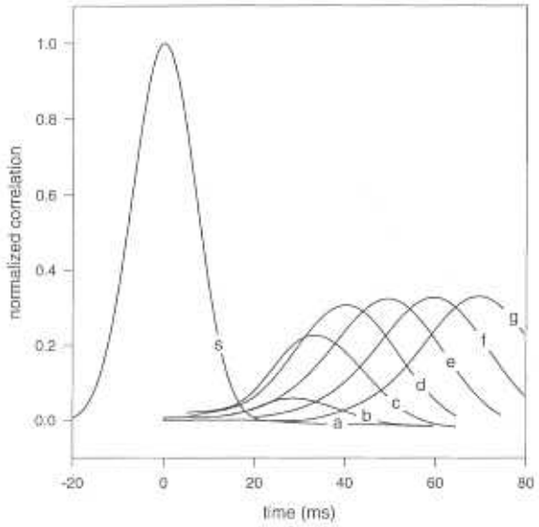



**Figure 2.23**
Normalized crosscorrelation between stimulus and reconstruction for various delays; also shown is the autocorrelation of the stimulus (curve S), which has been smoothed with a Gaussian filter with 5 ms standard deviation. The crosscorrelations shown are for stimulus estimates with delays of 10 ms (a), 20 ms (b), 30 ms (c), 40 ms (d), 50 ms (e), 60 ms (f), and 70 ms (g). The crosscorrelation increases for delays between 10 and 40 ms. Further increases in the delay have minimal effect on the crosscorrelation.

the reconstructions with the stimulus, as shown in Fig. 2.23. For a delay of 10 ms, close to the intrinsic delay for phototransduction, the reconstruction is almost uncorrelated with the stimulus. For delays in the range of 10–40 ms, the reconstruction improves with increasing delay. This improvement saturates for delays greater than 40 ms, close to the behavioral reaction time of 30 ms: The structure of the code and the behavioral decision times are quite well matched.

We have been describing the response of H1 in a simplified world where only the angular velocity varies in time. In fact, the firing rate in response to a given velocity waveform depends on the spatial structure of the patterns presented to the fly, so that there appears to be ambiguity among several stimulus variables—velocity, contrast, and spatial frequency in the case of grating stimuli. In addition, the response of H1 adapts to the velocity waveform itself, so that the encoding of velocities will depend on the ensemble from which the signals are chosen.



We have already seen that in a statistically stationary spatial environment, ambiguities in the response of H1 do not impede the estimation of the velocity waveform. What happens when we change the spatial environment? This was explored in the analysis of two preliminary data sets (Rieke 1991). In the first case the stationary random pattern was replaced by a single vertical stripe. In the second case the random character of the pattern is retained, but the effective contrast seen by the photoreceptors is increased by an order of magnitude. As expected for a wide field movement sensor, H1 does not code a single wide stripe as well as it does a whole field of stripes. We also expect that the precision of velocity estimation in a field of random stripes should be improved by increasing the effective contrast, and this is observed. The theoretical significance of these variations in precision is considered in section 4.3.3. Although quantitative differences among the reconstructions are clear, no qualitative differences are seen among the codes in these different stimulus ensembles. In each case it proves possible to reconstruct the stimulus by linear filtering of the spike train, the inclusion of nonlinear terms in the reconstruction does not have a significant effect, and the decoding filters themselves are remarkably similar. Although the issue remains to be explored more systematically, these results provide a hint that the strategy for reading the code in H1 may have a substantial degree of invariance with respect to changes in the parameters of the stimulus ensemble.

Another way in which decoding strategies might change is if we deliberately emphasize different aspects of the signal. In our discussion of coding thus far, we have chosen filters that minimize the mean square error, so that all aspects of the stimulus are weighted equally. What if we are especially interested in accurate estimates at large velocities? Once again, we need to know the structure of the probability distribution $P[s(\tau)|\{t_i\}]$: If the distribution has a single well defined peak, our choice of metric should not influence the decoding strategies dramatically. On the other hand, consider a cell that responds identically to two different stimuli. In this case the best decoding strategy will depend on the a priori probabilities for the two stimuli and on the cost of making mistakes in estimating these stimuli. Depending on the metric, we may form an estimate of the stimulus corresponding to one peak of the distribution $P[s(\tau)|\{t_i\}]$, or to the other peak, or to some compromise in between.

We return to our error measure in Eq. (2.37), and carry through the calculations of the linear kernel $K_1$ for different functionals $G[s]$. Figure 2.24 shows linear reconstruction filters calculated using $G[s] = s^2$ and $G[s] = |s|$ for the H1 experiments; both cases provide heavier penalties for errors at times when the signal is large. Aside from an overall scaling, the filters for these different

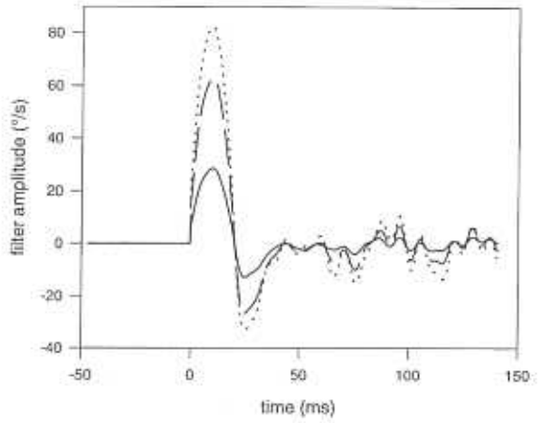



**Figure 2.24**
Estimation filters calculated for different metrics. Filters were calculated using the power series approach (see section A.8.2) with a delay of 40 ms. In each case the error function $E$ defined in Equation 2.37 was changed: the solid line is for $G[s] = 1$, in which case $E$ is the mean square error, the dashed line is for $G[s] = |s|$, and the dotted line for $G[s] = s^2$. The structure of the filters is similar, with the exception of a scale factor.

metrics have essentially the same shape. We do not see a dramatic shift in the decoding strategies for these different metrics.

These first experiments on H1 strongly suggest that the strategy of linear decoding works, and that at least for this system we are very close to giving the rules that must be followed by the homunculus. We would like to know if these rules have some generality, and we need to develop tools for quantifying the performance of the homunculus who follows these rules. As things stand, it might be that a more complex set of decoding rules would extract much more information from the spike train. It does seem true, however, that the problem of decoding the spike train is very different from that of describing the encoding, as first hinted in Fig. 2.2h, and that the decoding problem has the chance of being solvable even under conditions where the encoding is very nonlinear.

## 2.4 SUMMARY

We have proposed an approach to neural coding that centers on understanding what an animal can infer about the sensory environment from its own neural signals, and on how this information can be extracted from the neu-



ral responses. This is the problem that must be solved by the homunculus, but it is also the type of problem that must be solved by the organism. We have seen how many different experimental approaches to the neural code can be fit into a general probabilistic framework, and we have seen the crucial role of Bayes' rule in relating different points of view on the code and in establishing the context dependence of the code.

The problem of interpreting spike trains seems to have a simple solution, in that it is possible to estimate directly the waveform of unknown stimuli from observations of a single spike train. The somewhat vague "running commentary" we had hoped to receive from the homunculus is replaced by a quantitative reconstruction of time varying signals in the sensory environment. The estimation procedure itself is very simple, consisting, in essence, of an appropriately chosen linear filter. Although the details of this filter are probably not important, its structure allows us to see how the neural code is matched to behavior and also hints at the robustness of the code with respect to noise or variations in the importance of different stimulus features. Having seen that the organism's point of view can be pushed to its logical conclusion, we return to the more quantitative issues that motivated us in the introduction.

# Chapter 3
# Quantifying information transmission

In this chapter we try to quantify the information that sensory neurons convey about the outside world. The framework for this undertaking is provided by Shannon's information theory. Although there is a long history of information theoretic analyses in neurobiology, there is also an undercurrent of concern that the fundamental concepts of information theory are inappropriate for biology. Thus, our first task is to understand how information theory allows us to pose mathematically precise questions about the function of the nervous system. This discussion highlights the problem of understanding the ensemble from which sensory stimuli are drawn in the natural environment. Information theory also places limits on what is possible for any neural code, in the same way that the physics of diffraction places limits on what is possible for any imaging system. We then turn to experiments that aim at a direct measurement of information transmission by sensory neurons, exploring the conceptual and technical difficulties of such measurements. Finally we show how the reconstruction methods introduced in chapter 2 allow us to place a lower bound on the rate of neural information transmission in a complex sensory environment. This leads us to new experiments on information transmission in primary sensory neurons, to the demonstration that (at least in one case) more natural stimuli are coded more efficiently, and to the intriguing result that these neurons come close to the optimal performance allowed by information theory.

## 3.1 WHY INFORMATION THEORY?

When we observe the spike train of a sensory neuron we learn, in principle, about many different aspects of the stimulus. Colloquially, we say that we are "gaining information" about the sensory stimulus, or that the spike train is "transmitting information." We would like to make these intuitive notions more precise. Part of the difficulty is the multidimensional character of nat-

ural signals: A frog call, for example, can be described by its fundamental frequency, by the amplitudes and phases of the different harmonics, and by the shape of the envelope. One auditory neuron might tell us a great deal about one of these parameters, or perhaps a little bit about all of them. If we isolate each parameter, we can do discrimination experiments analogous to those in psychophysics, as will be described in chapter 4, but in the real world signals are varying continuously along all dimensions at once. Neurons are nonlinear and adaptive, so that the encoding of one stimulus dimension is not independent of the context provided by variations in all the other dimensions. How do we characterize the system's performance under natural conditions? Is there a quantitative measure of neural performance under these conditions analogous to the psychophysical discrimination threshold? Answers to these questions are provided, at least in part, by information theory. Information theory not only quantifies our intuitive notions of gaining information; it also puts information on an absolute scale, so that we can make meaningful statements about whether the information transmission rate in a particular neuron is large or small.

### 3.1.1 Entropy and available information

We can think about information transmission by sensory neurons in terms of the schematic in Fig. 3.1. Before we observe any spikes, we know that not all stimuli are equally likely, but rather that signals in the real world have structure and limitations; we indicate this by sketching a region in the stimulus space from which signals are likely to be chosen. When we observe the spike train, the range of possible stimulus waveforms is narrowed into a smaller region of the stimulus space, as described in the discussion of response-conditional ensembles in section 2.2.3. The information provided by the spikes about the stimulus measures this reduction on a logarithmic scale, so that a reduction by a factor of two in the range of possible stimuli is counted as one bit of information. For example, imagine that frogs call with fundamental frequencies scattered uniformly throughout a 50 Hz range, and that observation of the spike train of a single cell allows us to determine this frequency with a precision of 5 Hz. In this case we gain $\log_2(50/5) \sim 3.3$ bits of information.

To make these ideas more precise, we review some of the key points in Shannon's formulation of information theory (Shannon 1948, 1949). We are interested in cases where we observe some "output" $Y$ and are trying to gain information about the "input" $X$ (Fig. 3.2). In practice, $X$ is a sensory signal which is described as a function of time, and $Y$ is a set of spike arrival times. The device we are trying to characterize is a communication channel—we can



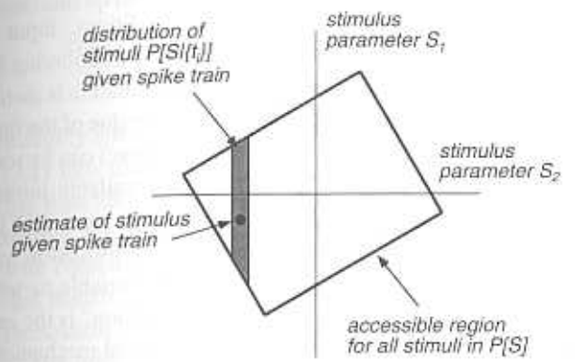



**Figure 3.1**
Schematic representation of stimulus space. Consider stimuli which are described by two parameters, $S_1$ and $S_2$. Prior to observation of the spike train we know the distribution of stimuli, $P[S]$, described by the box. Upon observation of a particular spike train $\{t_i\}$ the stimulus distribution is reduced from $P[S]$ (box) to $P[S|\{t_i\}]$ (gray region). Given $P[S|\{t_i\}]$ and a criterion for weighting errors in our estimate, e.g. maximum likelihood, we choose a particular estimate $S$ (dot).

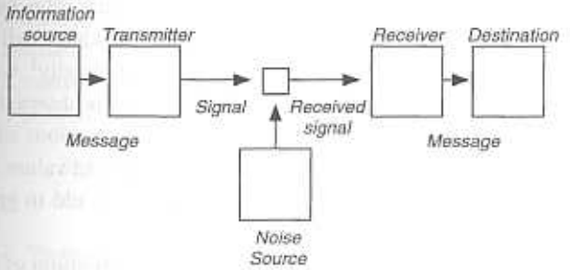



**Figure 3.2**
Shannon's communication system. The information source selects a particular message out of a set of possible messages. The transmitter changes this message into the signal which is sent over the communication channel to the receiver. The receiver converts the transmitted signal back into the message. In the process of being transmitted, unintended noise is added to the signal. Redrawn from Shannon (1948).

think of the outside world as sending us the message $X$, which we receive encoded as the output $Y$.

Under some set of natural or experimental conditions, input signals are chosen from a probability distribution $P[X]$. Before considering information transmission we need to characterize how much information is *available*. If the distribution $P[X]$ is so sharply peaked that only one value of the input variable is possible, then no information (in the colloquial sense) can be transmitted—the message is always the same. To quantify the available information, we need to measure the variability allowed by the distribution $P[X]$. Similarly, if the output $Y$ is always the same, there can be no information transmission, so we need a measure of variability for the output variable as well. The appropriate measure of variability or "available information" is the *entropy*—the same quantity defined in thermodynamics and statistical mechanics.

Any reasonable measure of variability must obey a significant constraint, namely additivity for independent variables. Suppose that the inputs $X$ are described by two statistically independent variables $X_1$ and $X_2$. Statistical independence means that the probability of observing a particular value of $X_2$ does not depend in any way on the value of $X_1$. Formally this implies that the probability of observing particular values for both $X_1$ and $X_2$ is just the product of the probabilities of observing $X_1$ and $X_2$ individually, that is, $P[X_1, X_2] = P_1[X_1]P_2[X_2]$. Our intuition says that we can express a certain amount of information about the world by giving a particular value for $X_1$, and similarly for $X_2$. This "amount of information," which we call the entropy $S$, depends on the probability distributions, so symbolically[1] we can write $S\{P[X_1]\}$ for the amount of information available from observations of $X_1$, and $S\{P[X_2]\}$ for the information available from observations of $X_2$. But if $X_1$ and $X_2$ are completely independent, then, if we are told values for *both* $X_1$ and $X_2$, the information from each variable should just add to give the total information available.

The additivity of information means that for any distribution of $N$ independent variables,

$$P[X_1, X_2, \ldots, X_N] = P_1[X_1]P_2[X_2]\cdots P_N[X_N], \tag{3.1}$$

1. We use this notation to emphasize that the entropy is a property of a probability distribution. Thus the "entropy of an image" is not defined; we can speak only of the entropy of a distribution, or ensemble, of images. Nonetheless, one often refers in thermodynamics to the "entropy of the gas"—presumably because saying the "entropy of the probability distribution from which the velocities of the molecules in this sample of gas have been drawn" is too cumbersome. In a similar vein we will later refer to the "entropy of $X$," and we will write this as $S[X]$. We trust the reader will forgive us for not writing $S\{P[X]\}$ every time.



we must be able to define the entropy of each individual distribution $P_i[X_i]$ and add up the results to give the entropy of the full distribution:

$$S\{P[X_1, X_2, \ldots, X_N]\} = S\{P_1[X_1]\} + S\{P_2[X_2]\} + \cdots + S\{P_N[X_N]\}. \tag{3.2}$$

Roughly speaking, if we want to convert a product of distributions as in Eq. (3.1) into the sum of entropies in Eq. (3.2), the entropy must behave like the logarithm of the distribution. Shannon (1948) gives a rigorous version of this argument and shows that the only measure of variability consistent with certain simple requirements (including additivity) is exactly the entropy that Boltzmann defined for statistical mechanics. This is a beautiful and quite remarkable confluence of ideas.[2]

The intuitive notion of entropy is that it is the logarithm of the number of possible states the system can occupy. Thus, if $X$ is a discrete variable, so that it can only have values $x_1, x_2, \ldots, x_K$, and each of the $K$ values occurs with equal probability, the entropy is $S \propto \log K$. If we still have $K$ possible values, but they occur with *unequal* probabilities (some values are more likely than others), then the entropy is

$$S = -k \sum_{i=1}^{K} p_i \log p_i, \tag{3.3}$$

where $p_i$ is the probability of observing the $i^{\text{th}}$ possible value and $k$ is a constant. This expression has a natural generalization to the case where $x$ has a continuous range of values, so we have a probability distribution function $P(x)$ rather than a discrete set of probabilities. In this continuous case,

$$S = -k \int dx\, P(x) \log P(x). \tag{3.4}$$

2. We alert the reader to some notational difficulties. In the engineering and information theory literatures it is common to use the symbol $H$ to denote entropy. But in thermodynamics $H$ is sometimes the enthalpy, or the expectation value of the energy, and the energy viewed as a function of the system coordinates is always the Hamiltonian $H$. Since the distinction between energy and entropy is crucial, we cannot bring ourselves to offend the ghosts of Boltzmann, Hamilton, and Helmholtz by using $H$ for the entropy. The enthalpy may also be written as $U$. $V$ is the volume, and $Z$ is the partition function. $P$ and $Q$ are the momentum and position of a particle, $R$ is the gas constant, and of course $T$ is temperature. $W$ is the number of ways of configuring our system (the number of states), and we have to keep $X$ and $Y$ in case we need more variables. Going back to the first half of the alphabet, $A$, $F$, and $G$ are all different kinds of free energies (the last named for Gibbs). $B$ is a virial coefficient or a magnetic field, $C$ is the specific heat, $D$ is the electric displacement in a dielectric, and $E$ is the electric field. $I$ will be used as a symbol for information; $J$ and $L$ are angular momenta. $K$ is Kelvin, which is the proper unit of $T$. $M$ is the magnetization, and $N$ is a number, possibly Avogadro's, and $O$ is too easily confused with 0. This leaves $S$, which must be the entropy.

There are also generalizations to the case where the variable $X$ is itself a function, such as the waveform of sound pressure versus time in the auditory system or light intensity as a function of position on the retina; this notion of random functions (as opposed to random variables) is discussed in section 3.1.4. We can combine all these cases into a shorthand notation,

$$S = -k \int [dX] P[X] \log P[X], \tag{3.5}$$

where $\int [dX]$ stands for a summation over any discrete variables, integration over all continuous variables, and functional integration in those cases where the variables $X$ define continuous functions in space or time.

In thermodynamics it is conventional to choose a constant $k$ (Boltzmann's constant) with units such that the product of the entropy and the absolute temperature is an energy, although this is not essential. In information theory all interesting measures are dimensionless, so $k$ is a pure number that can be eliminated by choosing the base of the logarithm.[3] The convention is to choose logarithms to the base two,

$$S = -\int [dX] P[X] \log_2 P[X], \tag{3.6}$$

and the resulting quantities of entropy or information are called *bits*. One bit is enough information to choose between two equally likely alternatives. Again, in the simple case of discrete variables this becomes [see Eq. (3.3)]

$$S = -\sum_{i=1}^{K} p_i \log_2 p_i \text{ bits}. \tag{3.7}$$

If all the $K$ different signals are equally likely, then $p_i = 1/K$, and

$$\begin{aligned} S &= -\sum_{i=1}^{K} p_i \log_2 p_i \\ &= -\sum_{i=1}^{K} (1/K) \log_2 (1/K) \\ &= \log_2 K. \end{aligned} \tag{3.8}$$

But this is just the number of digits that we need to write $K$ as a binary number, and this digital representation is illustrated in Fig. 3.3. Corresponding

3. We recall that logarithms in different bases are related to each other by constant factors. In particular, we make frequent use of the connection between the logarithm to the base two and the natural logarithm—for any number $x$, $\log_2 x = \ln x / \ln 2$.



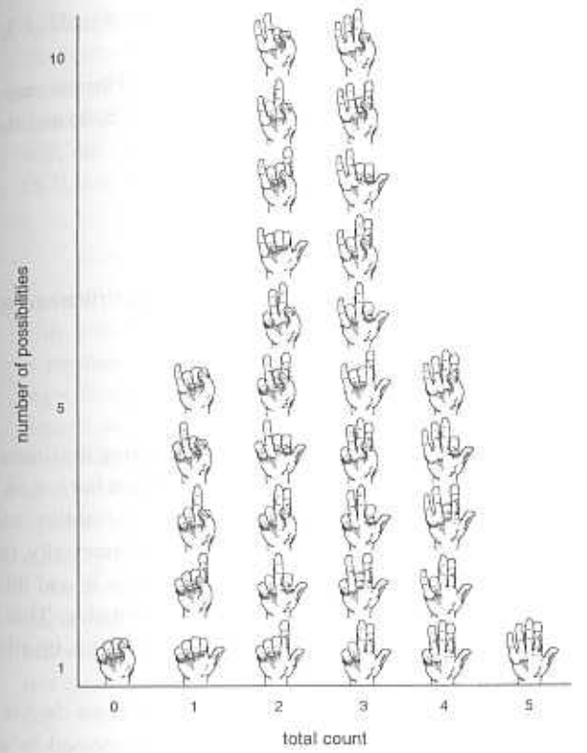



**Figure 3.3**
Encoding numbers in a digital code. Probably the most popular code used by people ordering drinks is one in which only the total number of raised digits carries the information. In this code, one hand can carry $\log_2(6) \approx 2.58$ bits of information (0 included, but not appreciated by any waiter). If we imagine the fingers to be time bins in a discretization of the spike train, with finger up (down) denoting the presence (absence) of a spike, then this conventional "bar code" is equivalent to a "rate code"—only the total number of spikes in the five bins, and not their temporal sequence, carries information. But, as the figure makes clear, if we keep track of "timing" and allow the position of each finger to carry information, then one hand can convey $2^5 = 32$ distinct messages, or 5 bits of information. This finger code has a greater capacity for carrying information, but the bar code is more robust as the message is, for example, invariant to being viewed in a mirror. This robustness derives from the redundancy of the code, since one number may be represented by several combinations of finger positions. One could also imagine neural codes in which particular patterns of spikes—represented here as particular finger configurations—are endowed with special significance.

to our intuitive ideas, the larger the number of possible signals ($K$), the larger the entropy, and the measure is logarithmic.

The Gaussian distribution provides a simple example for the case of continuous variables. Thus, if the average or mean value of $x$ is $M$ and the variance is $\sigma^2$, then the probability distribution is given by

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-M)^2}{2\sigma^2}\right]. \tag{3.9}$$

Substituting into Eq. (3.4), we find the entropy of the distribution (with details in section A.9):

$$S = \frac{1}{2}\log_2(2\pi e\sigma^2) \text{ bits}. \tag{3.10}$$

This expression for the entropy of a Gaussian distribution illustrates three important points. First, the entropy depends on the variance but not on the mean. This makes sense because the entropy is a measure of *variability*, and the variability clearly does not depend on the mean. More fundamentally, the value of the mean is dependent on how we choose the point $x = 0$, and this arbitrary choice should not change our notion of available information. This is the first hint that information theoretic quantities need to have certain invariances with respect to changes in parameterization of the signals.

The second important point about Eq. (3.10) is that if we double the width of the distribution, so that the standard deviation is increased by a factor of two ($\sigma \to 2\sigma$), then the entropy goes up by exactly one bit. Again this is because entropy is a logarithmic measure, so that differences in entropy count the *factor* by which the range of possible $x$ values has been increased or decreased, as in the example of the frog call given at the beginning of this section.

The final point about the entropy of a Gaussian distribution is that it looks a bit funny, because we are supposed to take the logarithm of the variance $\sigma^2$. But because the variable $x$ is a physical quantity, it has units—milliVolts, perhaps—so the variance has units (milliVolts)$^2$. We know what we mean by the logarithm of a number, but what do we mean by the logarithm of milliVolts? Worse, if we choose to measure in Volts instead of milliVolts, the numerical value of $\sigma$ will change by a factor of $10^3$, and it would seem that the entropy of the voltage distribution would change by $\log_2(10^3)$, or 10 bits. Of course no real physical quantity can depend on our choice of units, so something is wrong.



The problem with measuring the entropy of continuous variables is that the "number of possible states" is infinite. If we agree that voltage (to continue the example) is measured only to a resolution of $\Delta V$ milliVolts, then we can take the continuous voltage variable and place it in discrete bins of size $\Delta V$. But with discrete variables we can go back and recompute the entropy using Eq. (3.3), and as long as $\Delta V$ is very small (much smaller than $\sigma$), we find that

$$S = \frac{1}{2}\log_2\left[2\pi e\frac{\sigma^2}{(\Delta V)^2}\right] \text{ bits}. \tag{3.11}$$

This is almost the same answer as in Eq. (3.10), but now the voltage variance is normalized by the resolution of our measurements. The quantity inside the logarithm no longer has units (it is dimensionless), and so the entropy doesn't depend on our choice of units. But, if we think that measurements can become arbitrarily accurate, so that $\Delta V \to 0$, then the entropy becomes infinite. We should conclude that the entropy for continuous variables is not quite well defined, but it can be defined if we remember that measurements always have a finite precision.

Suppose that, instead of asking for the entropy of a distribution, we ask for the difference in entropy between two Gaussian distributions that have different variances, say $\sigma_1$ and $\sigma_2$. We assume that when we make measurements our resolution $\Delta V$ is the same in both cases. Then we can use Eq. (3.11) *or* Eq. (3.10) to compute the entropy of each distribution, and then take the difference, to find $\Delta S = \log_2(\sigma_1/\sigma_2)$ bits. Again this shows us that if the standard deviation changes by a factor of two, the entropy changes by exactly one bit. More importantly, we see that entropy *differences* are always well defined, even without an explicit limit on measurement precision. This is important because it will turn out that information transmission is measured by an entropy difference. We have loosely equated "well defined" with "independent of our system of units," and in fact this can be made more rigorous: Entropy differences are independent of *any* reparameterization as long as the reparameterization is invertible. We can shuffle the labels on the signals all we want—as long as the labels remain unique—and the entropy differences will not change.

The problems we encounter in defining entropy for continuous variables exist already in the original physics context for these concepts. In classical mechanics, particles can take on a continuous range of positions and velocities, and there is no natural scale for the precision of measurements. Correspondingly, the absolute entropy is ill defined, although again entropy differences

are perfectly sensible. This is fine, because one can observe only entropy differences (or the associated heat flows), not absolute entropies. In quantum mechanics Planck's constant provides a natural scale for the ranges of position and momentum, and absolute entropy becomes meaningful as a description of the degree of order or randomness in the system. As Planck's constant becomes small compared to typical motions of the system, all observable quantities, such as entropy differences, approach those calculated in the classical theory. The point of this digression is that the difficulties of defining entropy are not unique to Shannon's application of the concept and that these difficulties do not in any way impede the correct calculation of *observable* quantities such as heat flow or information transfer.

To return to our discussion of entropy as a measure of available information, we see that the entropy of an ensemble of signals depends on the number of different possible signals, not on the complexity of the individual signals themselves. As an extreme example, imagine that we call a friend on the phone and give a dramatic reading of either *Macbeth* or *Hamlet*. It is easy to imagine that large amounts of information are being conveyed. But if our friend knows in advance that we will be reading one of two plays, then all we need to tell her is which one. This is just one bit of information, nowhere near the naive or colloquial "information content" of the texts. As an alternative, imagine that all possible signals are stored on a large hard disk. If there are $K$ possible signals, then each signal has a unique address that is $\log_2 K$ bits long. To specify a signal we need to give its address, not the full contents of the corresponding sector on the disk. The entropy of the ensemble of signals is the length of the addresses, not the capacity of the disk.

These examples point out that information theory makes sense only in situations where the "receiver" of signals knows the full range of possibilities. In a simple case, this means that we are trying to transmit one of $K$ possible signals, and the receiver has a list of these signals. It doesn't make sense to analyze the information transmitted in Morse code unless we assume that the receiver *knows* that we are using Morse code, and hence that the elementary symbols are dots and dashes. At a more advanced level, successive symbols are usually not independent—only certain combinations of dots and dashes form letters, certain combinations of letters are much more likely to form words, and so on. Again, the "available information" is well defined only if we assume that the receiver of these signals knows about this statistical structure in the bit stream.

To make these ideas precise, we say that the entropy measures the information available by observing a particular signal chosen from a *known* distribution. Shannon developed information theory (and we will use it) primarily to deal with situations where information is being transmitted continuously over long times. In such a case we can imagine that anything we need to know about the distribution of signals is agreed upon at the outset, and that at long times this has no effect on the steady state *rate* of information transmission. Under these conditions, the formulation where the distribution of possible signals is assumed known seems to make perfect sense. Clearly the "transient" period where the receiver learns about the probability distributions of signals is of great interest, especially in a biological context, but this is distinct, at least conceptually, from an analysis of information transmission in the steady state.



It is often pointed out that Shannon's measure of information is blind to semantics or "meaning." Thus, a text written in a foreign language can have the same information content as a text written in a language we understand, and this seems wrong. In a similar vein, it is argued that Shannon's formulation is not relevant to biology because it does not take account of the organism's interest or lack of interest in different aspects of the world. The locations and trajectories of predators are deemed biologically more interesting than the detailed patterns of leaves on a tree, yet Shannon might assign these different signals similar information measures. In both of these examples, at least part of the problem lies in specifying the probability distribution from which signals are being drawn.

When we read a foreign language, we start out with essentially no knowledge of the correlations between successive symbols. We can interpret the incoming signals only as a string of letters and spaces, and if asked to assign an information content to this string we need to state an expected probability distribution for such strings. But what do we do about languages with unfamiliar symbols? Especially in this case, it seems difficult to give a probabilistic description of the nominal tabula rasa on which the new language is being recorded. The Shannon formulation, strictly interpreted, does *not* define the information content of texts in an unfamiliar language. If forced to give such a measure, the intrepid information theorist would have to interview the reader about his or her linguistic assumptions, and the resulting information content would be different for different individuals, in accord with our intuition.

There is a more subtle problem in the application of information theory to unfamiliar languages, namely that as we read more and more of the text we learn some of the structure of the new language. This leads us to revise our estimates of the probability distribution for strings of symbols, and hence changes the nominal information content of subsequent text. It is even possible that this information about the structure of the language or the writing style

always dominates the information about the particular text. As one analyzes longer and longer texts, this linguistic structure reveals itself gradually and the estimates of the entropy per letter become progressively smaller, so that in the limit of an infinitely long text the rate at which we gain information by reading is very small compared to the initial rate (Ebeling and Pöschel 1994).[4] This reminds us that the application of information theoretic concepts to language is subtle, and hence that one should be careful in using language as an example of the limitations on Shannon's formulation. We cannot resist pointing out that these issues are closely connected to the classic story about monkeys and typewriters (Kittel and Kroemer 1980).

Assigning an information content to, for example, a visual image requires knowledge of the distribution the image came from. If we assume—incorrectly—that images are drawn from a simple probability distribution such as spatial white noise, then many images of very different "significance" will be assigned the same information. But natural images come from highly structured probability distributions, and the correct information measure must be computed using the relevant distribution. Thus, the appearance of a threatening predator is (one hopes) a rare event. Once a predator appears, its pattern of motion, for example, is very unusual relative to the motions of other objects in the environment. The Shannon measure of information tells us that, *per event*, rare events convey more information than common ones. In the limit, the rarest events can convey arbitrarily large amounts of information per event. Thus the appearance and motion of a predator would be assigned a high information content, much higher than that of typical scenes, provided that the observer understands the statistics of natural scenes. One could even make the argument that rarity alone is sufficient to generate "biological interest" and, for example, attract our attention. The decision that a rare event is or is not threatening (or edible) might be a second and specialized step.

Another important application of information theory is to decompose the total available information into information carried by different features of the signal. In the visual world of a frog, for example, it is usually assumed that the tracking of flying bugs is a biological specialization (Lettvin et al. 1959). But since a static background provides zero information per unit time (no matter how complex it might appear!), it may be that by tracking just a few bugs the

4. Having reached this point in the book, the reader may have collected enough data to test this claim.



frog in fact captures a large fraction of the available visual information. This serves to emphasize the general point that "information" is not an absolute quantity, but rather a measure of how much one can learn *relative* to what one knows a priori. To say that certain neural or receptor cell signals "provide information" about different aspects of the world, we therefore must state our assumptions about the world itself.

Part of the difficulty in applying information theory to signals in a biological context is that the probability distributions for natural signals are very subtle objects, and we know very little about their structure; some of what we do know is described in chapter 5. Roughly speaking, the world is a very ordered place, and so the sensory signals that derive from the world have strong internal correlations. This lowers the entropy of the signals and limits the information available to our senses. But rare events still stand out, conveying a disproportionately large number of bits: Shannon tells us that a bear in the woods is interesting even if we don't know that bears can eat us.

### 3.1.2 Entropy of spike trains

The entropy plays a key role in our thinking about the neural code. On the one hand, the amount of information available is limited by the entropy of the input sensory signals, and we must be careful in the design of experiments to insure that we are not creating an artificial world of anomalously low entropy. On the other hand, the entropy of the spike trains themselves limits how much information these spikes could, even in principle, provide about the sensory input.

The entropy of spike trains was estimated by MacKay and McCulloch (1952), in what was probably the first application of information theory to the nervous system, a scant four years after Shannon's original work. They envisioned the spike train as being observed with some limited time resolution $\Delta\tau$, so that in each time slice (bin) a spike is either present or absent, as shown in Fig. 3.4. If we think of a spike as representing a '1' and no spike as representing a '0,' then, if we look at some time interval of length $T$, each possible spike train is equivalent to a $T/\Delta\tau$ digit binary number. But not all of these numbers occur with equal probability. In fact we know that some of these numbers never occur at all—for example, if $T$ is very large and $\Delta\tau$ is reasonably small, the string $11111\cdots11111$ never occurs because real neurons don't maintain very high firing rates over long periods of time.

Suppose that spikes occur at some mean rate $\bar{r}$, so that the probability of a '1' is just $p = \bar{r}\Delta\tau$. If we choose very small bins, so that $p$ is small, a long

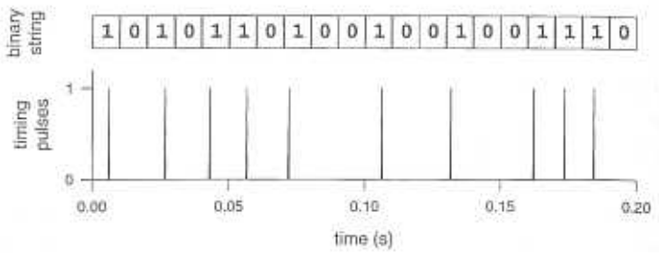



**Figure 3.4**
Binary representation of spike trains. By dividing the time axis into discrete bins, a spike train, represented here as a series of impulses, can be represented as a series of binary digits, where a 1 denotes a spike in the time bin, and a 0 denotes no spike.

segment of the spike train will map to a binary string with very few 1's. To estimate the entropy of the spike train we need to know how many such biased strings are possible. For simplicity let us assume that the spikes in different bins are uncorrelated, so that the possible strings are not constrained except by the bias toward having more 0's; we shall see later that this leads us to overestimate the entropy.

We can calculate the entropy of these strings by appealing to Shannon's definitions, but Brillouin (1962) explains nicely how we can stick with our intuition about numbers of possible signals and in some sense recover the formal definition of entropy. The idea is to imagine a very long string, of total duration $T$, so that there are $N = T/\Delta\tau$ bins. With a very long string there will be $N_1 = pN$ 1's (spikes) and $N_0 = (1-p)N$ 0's. The number of possible strings is just the number of different ways of arranging $N_1$ 1's and $N_0$ 0's; this a standard counting problem. The answer is that the number of possible strings is

$$N_{\text{strings}} = \frac{N!}{N_1!N_0!}, \tag{3.12}$$

where we recall that $N! = N \times (N-1) \times (N-2) \times \cdots 3 \times 2 \times 1$. The entropy is the logarithm of this number,

$$\begin{aligned} S &= \log_2\left[\frac{N!}{N_1!N_0!}\right] \\ &= \frac{1}{\ln 2}\left[\ln N! - \ln N_1! - \ln N_0!\right]. \end{aligned} \tag{3.13}$$



But we are dealing with long strings, so all the numbers $N, N_1, N_0$ are large. Then we can then use Stirling's approximation for the $N!$'s,

$$\ln x! = x(\ln x - 1) + \cdots, \tag{3.14}$$

where the corrections are negligible as $x \to \infty$. Substituting Stirling's approximation into Eq. (3.13), we find

$$\begin{aligned} S &= \frac{1}{\ln 2}\left[\ln N! - \ln N_1! - \ln N_0!\right] \\ &= \frac{1}{\ln 2}\left[N(\ln N - 1) - N_1(\ln N_1 - 1) - N_0(\ln N_0 - 1) + \cdots\right] \quad (3.15) \\ &= \frac{1}{\ln 2}\left[N\ln N - N_1\ln N_1 - N_0\ln N_0 - (N - N_1 - N_0)\right], \quad (3.16) \end{aligned}$$

which we can simplify because $N = N_0 + N_1$. Thus we have

$$\begin{aligned} S &= \frac{1}{\ln 2}\left[N\ln N - N_1\ln N_1 - N_0\ln N_0 - (N - N_1 - N_0)\right] \\ &= \frac{1}{\ln 2}\left[(N_1 + N_0)\ln N - N_1\ln N_1 - N_0\ln N_0\right] \quad (3.17) \\ &= \frac{1}{\ln 2}\left[N_1(\ln N - \ln N_1) + N_0(\ln N - \ln N_0)\right] \quad (3.18) \\ &= -\frac{1}{\ln 2}N\left[\left(\frac{N_1}{N}\right)\ln\left(\frac{N_1}{N}\right) + \left(\frac{N_0}{N}\right)\ln\left(\frac{N_0}{N}\right)\right] \quad (3.19) \\ &= -\frac{N}{\ln 2}\left[p\ln p + (1-p)\ln(1-p)\right], \quad (3.20) \end{aligned}$$

where in the last step we recall that $p = N_1/N$ and $1 - p = N_0/N$. Finally, if our spike train is of duration $T$, then the total number of bins $N = T/\Delta\tau$, and, by definition, $p = \bar{r}\Delta\tau$, so that

$$S = -\frac{T}{\Delta\tau\ln 2}\left[(\bar{r}\Delta\tau)\ln(\bar{r}\Delta\tau) + (1 - \bar{r}\Delta\tau)\ln(1 - \bar{r}\Delta\tau)\right]. \tag{3.21}$$

It is important to note that, from Eq. (3.21), the entropy of spike trains is proportional to their length, so that $S \propto T$. This is the same as in physics, where we have the notion that entropy is an extensive quantity. Thus, if we have a gas or liquid with fixed density, then if we increase the volume of the system the entropy increases in proportion. The spikes are like the molecules of a one dimensional gas, so fixing the density is like fixing the average firing rate, and once we do this the entropy is proportional to the length of the spike

train. It thus makes sense to think about the "entropy rate" $S/T$, which has the units of bits per second. This rate is, as we will see, the maximum possible rate at which the spike train can convey information about signals in the sensory environment.

To use our digital picture of the spike train as consisting of 1's and 0's, it must be the case that the probability of observing a spike in one bin is very small, that is, $p = \bar{r}\Delta\tau << 1$. As explained in section A.10, this allows us to approximate the entropy rate in Eq. (3.21) as

$$S/T \approx \bar{r} \log_2 \left( \frac{e}{\bar{r}\Delta\tau} \right). \tag{3.22}$$

The entropy of the spike train is shown in Fig. 3.5a as a function of the timing precision $\Delta\tau$. Perhaps the most important point is that the entropy of the spike train can be larger than one bit per spike. This is because interspike intervals are distributed over a range of approximately $1/\bar{r}$ seconds, and these times are measured with an accuracy of $\sim \Delta\tau$, so that each interval is chosen from $\sim (\bar{r}\Delta\tau)^{-1}$ possibilities. Each interval (and hence each spike) is thus associated with $\sim \log_2(1/\bar{r}\Delta\tau)$ bits of entropy. With $\bar{r} \sim 50\,\mathrm{s}^{-1}$ and $\Delta\tau \sim 1\,\mathrm{ms}$, this corresponds to 5.76 bits per spike, or 288 bits/sec.

The calculation of spike train entropy also provides a method for assessing the capabilities of different candidate coding schemes. As an example, let us contrast the result in Eq. (3.22)—which is the entropy of the spikes assuming that we track each spike to a temporal precision $\Delta\tau$—with a scheme in which we count spikes in some large window of duration $T$, or equivalently measure the rate of spiking in this window. In this case information can be carried only by the spike count $n$, and so we need to know the entropy of the spike count distribution, that is,

$$S(\text{spike count}) = -\sum_n p(n) \log_2 p(n), \tag{3.23}$$

where $p(n)$ is the probability of observing $n$ spikes in the window of width $T$. What should we choose for $p(n)$? We know that probability distributions must be normalized, so that $\sum_n p(n) = 1$. We also know that the average spike count in a window of duration $T$ should be $\langle n \rangle = \bar{r}T$, where $\bar{r}$ is the mean spike rate, as before. These two constraints do not uniquely determine the spike count distribution. We can ask, however, what spike count distribution consistent with these constraints will *maximize* the spike count entropy? This is a general strategy in the application of information theory—given some knowledge, we try to find the distribution that has the maximum entropy con-



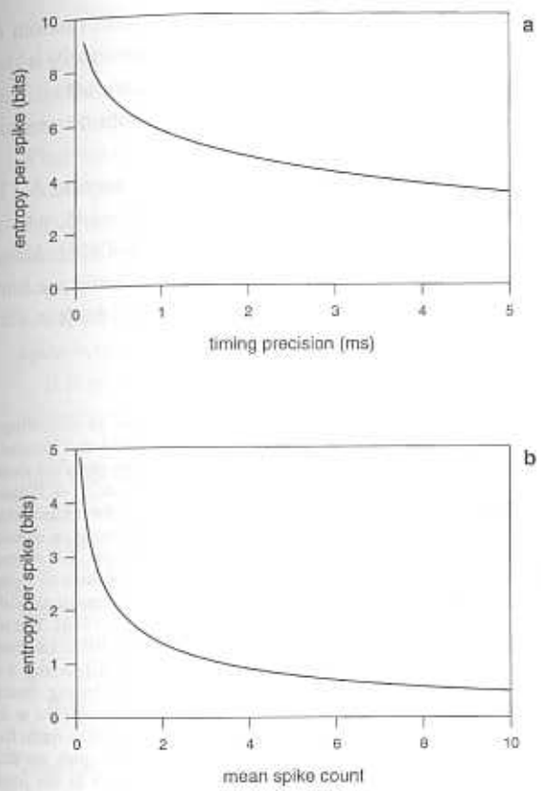



**Figure 3.5**
Spike train entropy as a function of timing precision. (a) The entropy rate $S/T$ is calculated from Eq. (3.21) assuming a firing rate of 50 spikes/s, and we divide by the mean spike rate $\bar{r}$ to give the entropy per spike. (b) The entropy is calculated from the upper bound in Eq. (3.24), and we divide by the mean count $\langle n \rangle$ to give, once again, the entropy per spike. Note that in (b) the timing precision is expressed as the mean number of spikes in one time bin; with the rate of 50 spikes/s, a mean spike count of 1 corresponds to a timing precision of 20 ms.

sistent with our knowledge. This distribution is the most random description that is not eliminated by what we already know; conversely it describes no structure beyond that which is established by the given facts.[5] This formulation allows us to make firm statements about information transmission even from limited data.

The answer to our maximum entropy problem (see section A.11) turns out to be an exponential distribution, that is, $p(n) \propto \exp(-\lambda n)$, where the coefficient $\lambda$ is set by the mean firing rate, $\lambda = \ln[1 + (\bar{r}T)^{-1}]$. Using this distribution we can calculate the entropy of spike counts, and we know that the entropy of spike counts in any real neuron will always be less than what we calculate. The result is that

5. Entropy provides a measure of variability or available information. In comparing two systems, the one with the larger entropy is the more variable or the more "random." We can also say that the larger entropy means that the system is "less ordered." If we have observed some, but not all, properties of a system—say the mean and variance of a random variable $x$—it is natural to ask for the probability distribution that is consistent with our data but does not introduce any unnecessary structure; the intuition is that this should be the "most random" distribution consistent with the data, and this is identified with the distribution that has the largest possible entropy. Statistical mechanics is based on the idea that, in thermal equilibrium, configurations of a physical system are drawn from the probability distribution that has maximum entropy consistent with the (known) average energy, and this point of view has been emphasized by Jaynes (1983). The solution to this maximum entropy problem is the Boltzmann distribution (Brillouin 1962; Landau and Lifshitz 1969). The fact that real noise sources are often Gaussian, as in Fig. 3.7, is also a manifestation of maximum entropy behavior, since the Gaussian is the distribution having maximum entropy at fixed variance (section A.13). In the text we use the maximum entropy idea to determine (for example) the limits on the information available to an observer of the spike train. But "maximum entropy methods" often refer to a different set of ideas, in which entropies are defined not just for probability distributions but also for other positive quantities such as the intensities in the pixels of an image. These notions of "image entropy" attracted considerable attention when Gull and Daniell (1978) found maximum entropy images that were consistent with data from radio telescope observations, and the resulting maps of the sky revealed remarkable levels of detail that could be confirmed using other astronomical techniques; recent books include Bevensee (1993), Buck and Macaulay (1991), and Skilling (1989). In fact neither Shannon nor Boltzmann told us about the entropy of an image, and indeed (as emphasized in section 3.1.1) information theory does not assign an entropy to a single image, only to the ensemble from which images are drawn. Consider then the following remarks from Gull and Daniell (1978): "Let $m_j$ denote the intensity at point j of a test map . . . There is some dispute about the correct definition of the entropy. We believe that the arguments of Freiden are correct for our problem. He concludes that the most probable map is that which maximizes $-\sum_j m_j \log m_j$. This formula expresses the configurational entropy of a map; this is not the same as the thermodynamic entropy of a beam of photons, nor is it the same as the information-theoretical entropy . . . ." Although progress has been made on these foundational issues, as described in the essays collected by Buck and Macaulay (1991), it is certainly not clear how "maximum entropy methods" as advocated by Gull, Daniell, Skilling and others are connected to the ideas of entropy in physics and information theory. This is one reason why mention of "maximum entropy" may trigger concerns in some readers' minds. We hope it is clear that our use of "maximum entropy" is quite literal—we find the probability distribution that has maximum entropy, as defined by Shannon and Boltzmann, subject to some stated constraints. We calculate the maximum entropy for the simple reason that we want to know the maximum possible value of the entropy.



$$S(\text{spike count}) \leq \log_2(1 + \langle n \rangle) + \langle n \rangle \log_2(1 + 1/\langle n \rangle) \text{ bits.} \quad (3.24)$$

In Eq. (3.24) we have written the entropy of spike counts in a window of size $T$ in terms of the mean spike count $\langle n \rangle$. To obtain the entropy per unit time we divide by $T$, and to obtain the entropy per spike we divide by $\langle n \rangle$. What we see in Fig. 3.5b is that the entropy per spike always decreases as the counting window gets larger, corresponding to larger mean spike count. Thus the capacity of the spike train to carry information declines as our "rate code" becomes more and more coarse in its time resolution. Indeed, if the windows contain very large numbers of spikes, then the available information capacity per spike or per unit time goes to zero! To achieve a capacity of one bit per spike requires $\langle n \rangle \leq 3.4$.

It is perhaps amusing to note that if we choose to count spikes in windows whose size is equal to the mean interspike interval, so that $\langle n \rangle = 1$, then the entropy is precisely two bits per spike. Thus, even a "rate code" can carry more than one bit per spike if we measure the rate on a time scale comparable to the interspike intervals themselves. If we imagine trying to count the numbers of spikes in very small windows, so that $\langle n \rangle \ll 1$, then the distinction between rate and timing codes becomes blurred—we are counting the presence or absence of a single spike in a small bin as a change in rate—and the entropy of the spike count distribution in Eq. (3.24) approaches the entropy of the full spike train given by Eq. (3.22).[6]

If the signals in the natural sensory world are varying sufficiently slowly, then it makes sense to divide the spike train into time windows, each of which contains many spikes. The spikes in each window are presumably responsible for telling us about the (approximately) static parameters of the signal averaged over that window, and the distinction between rate and timing codes is clear—do we gain more information about the static stimulus parameters by looking at the detailed timing of the spikes within the window, or is all the information contained in the spike count? If stimuli are varying on time scales comparable to the interspike intervals, the natural time windows contain at most a few spikes, and the rate code will transform smoothly into a timing

6. Notice that we have found the *maximum* entropy given the mean spike count, and this maximum goes smoothly into the MacKay–McCulloch entropy as we let the time resolution $\Delta\tau$ become small. But this means that the MacKay–McCulloch result itself is the maximum possible entropy of spike trains in the small $\Delta\tau$ limit where the spike train can be viewed as a binary string. This makes sense because MacKay and McCulloch ignored any correlations among successive spikes, fixing only the mean spike rate, and correlations can only reduce the entropy.

code, as we see from the entropy calculation. Thus, from an information theoretic point of view, the first question is *not* rate versus timing, but rather the number of spikes per interesting time window, as we have emphasized in chapter 2.

In computing the entropy of the spike train we are just counting the number of different spike trains that can be distinguished given our time resolution. In an ideal code, each of these distinguishable spike trains would stand for a unique signal or class of signals in the outside world. Then, if there are $K$ distinguishable spike trains, corresponding to an entropy $\log_2 K$, we have distinguished $K$ possible signals in the world. When we observe a particular spike train we know which of these possibilities actually occurred, so looking at the spike train is equivalent to observing one of $K$ sensory signals. Under these conditions the information available about the world is exactly equal to the entropy of the spike trains, whereas if we move away from this ideal code we will distinguish fewer signals in the world and hence gain less information. Thus, the entropy of the spike train sets a physical limit to the information a neuron can convey about external signals.

There are two important points about the entropy calculations. First, the results dispel the surprisingly persistent notion that each spike conveys at most one bit of information. Even though the signaling events are all–or–none, they are sparse, which means that the information per event can be large. The information per bin can never be bigger than one bit, but this is an uninteresting constraint as long as bins are much smaller than typical interspike intervals. Second, the entropy gives us a standard against which to measure the performance of real neural codes: How close do real animals come to using all of the spike train entropy for information transmission?

We want to emphasize this notion of putting the performance of the neural code on an absolute scale. When we think about building an imaging system in the laboratory, or analyze the performance of a visual system, we know that our intensity resolution is limited by the random arrival of photons. Similarly, our angular resolution is limited by diffraction. These basic physical limits help organize our thinking, as will be described in chapter 4. In the lab, photon shot noise tells us how bright our light sources must be to resolve small changes in image brightness, and in biology there are clear experimental questions to be asked about response of photoreceptors to single photons and the processing of these signals by subsequent layers of the visual system. In the absence of these ideas we have no way of knowing whether an imaging system which is sensitive to 1% contrast is good or bad. Similarly, to find out whether the French cave beetle *Speophyes lucidulus* that detects temperature changes of 1/1000 of a degree is really doing well,[7] we must compare this performance with the physical limits set by thermodynamic temperature fluctuations (Bialek 1987). In the absence of such absolute standards we don't really know if, for example, a change in firing rate of so many spikes per second in response to a given stimulus represents a large or a small change. In the same way that quantum and thermal fluctuations set the standard for the detectability of small signals, the entropy of the spike train sets the standard for the transmission of information.



### 3.1.3 Mutual information and the Gaussian channel

Now we return to our original problem, characterizing the amount of information that some outputs $Y$ carry about input signals $X$. We have seen that, to give the problem a clear formulation, we must assume that $X$ is chosen from some known probability distribution $P[X]$. The variability of $X$ is then measured by the entropy of $P[X]$, roughly the logarithm of the number of possible signals. Once we observe $Y$, however, the range of possible inputs is restricted, as schematized in Fig. 3.1. This reduced variability is described by the conditional distribution $P[X|Y]$, which measures the relative likelihood of different input signals $X$ given that we have observed a particular output value $Y$. Presumably, only a limited set of signals $X$ are consistent with our observations on $Y$, so the *conditional entropy*,

$$S[X|Y] = -\int [dX]P[X|Y]\log_2 P[X|Y], \tag{3.25}$$

7. This remarkable sensitivity is described in a series of papers by Loftus and Corbière-Tichané (1981, 1987; Corbière-Tichané and Loftus 1983). In addition to demonstrating millidegree thermometry, this work provides a clear example of how the full performance of a biological system may be revealed only by attention to the natural stimulus ensemble. Initial experiments (Loftus and Corbière-Tichané 1981), in which the antennae were stimulated with puffs of air differing from the ambient temperature by as much as several degrees (Celsius), suggested neural sensitivities on the order of $\sim 10$ (spikes/s)/degree with integration times of order one second. For the second set of experiments, Corbière-Tichané and Loftus (1983) measured the temperature in the caves where they collected the beetles and found that it was stable to a precision of 0.01°C over periods of half an hour. This suggested stimulation with small, slow temperature fluctuations, generated essentially as a residual after attempting to stabilize the temperature of the preparation. These experiments indicated that the receptor neurons respond to slow temperature drifts with sensitivities as high as $\sim 10^3$ spikes/degree. Finally, improvements in the stimulation apparatus allowed controlled application of still slower temperature drifts, comparable to those observed in the cave, and this resulted in sensitivities averaging more than $3 \times 10^3$ spikes/degree (Loftus and Corbière-Tichané 1987). With spontaneous firing rates of roughly 10 spikes/s, any reasonable integration time will be sufficient to detect reliably a few extra spikes, corresponding to a sensitivity of better than a millidegree, and this is extremely close to the limits set by thermodynamic temperature fluctuations (Bialek 1987).

is smaller than the total entropy $S[X]$. Intuitively, $S[X|Y]$ counts the logarithm of the number of $X$'s consistent with $Y$, and this number is smaller than the overall number of possible $X$'s, which is counted by the total entropy. This reduction in entropy is defined to be the information gained by observing $Y$. If we average over all values of $Y$, we obtain the mean information gained,

$$I = \int [dY] P[Y] \left( S[X] - S[X|Y] \right)$$

$$= \int [dY] P[Y] \int [dX] P[X|Y] \log_2 \left( \frac{P[X|Y]}{P[X]} \right). \tag{3.26}$$

We emphasize that this is the *average* information gained by observing $Y$. The information gained by observing a particular $Y$, say $Y_0$, is the entropy difference $S[X] - S[X|Y_0]$, and this can be much larger or smaller than the average. If $X$ and $Y$ are continuous variables, one typically finds that some particular observations of $Y$ provide arbitrarily large amounts of information—but these data are very unlikely to occur, so the average information is still finite. But it is important to remember that, very rarely, one expects the occurrence of unusually informative events.

We recall from our discussion of conditional distributions in chapter 2 that the probability $P[X|Y]$ of the input $X$ given the observed output $Y$ can be written as

$$P[X|Y] = \frac{P[X,Y]}{P[Y]}. \tag{3.27}$$

Then we can write the information in a more symmetric form,

$$I = \int [dY] P[Y] \int [dX] P[X|Y] \log_2 \left( \frac{P[X|Y]}{P[X]} \right)$$

$$= I = \int [dY] \int [dX] P[X,Y] \log_2 \left( \frac{P[X,Y]}{P[X]P[Y]} \right). \tag{3.28}$$

This quantity, the average information that observations of $Y$ provide about the signal $X$, is also called the *mutual information* of $X$ and $Y$. Note that it is symmetric under interchange of the two variables. This symmetry means that we can think of observations on the spike train as telling us what is happening in the outside world, or we can think of observations on the state of the world as predicting the spike train, and in either view the average information transfer is the same. This reminds us once more of Bayes' rule.

Another way of thinking about the mutual information is to imagine that $X$ and $Y$ are chosen independently from the marginal distributions $P[X]$ and



$P[Y]$. In this world, the entropy of the whole system, $S[X,Y]$, is just the sum of the entropies $S[X]$ and $S[Y]$ associated with the individual coordinates. But in the real world $X$ and $Y$ are correlated, so that observing $Y$ tells us about $X$. So the entropy of the whole system is *less* then the entropy of the two coordinates added together, and the amount by which it is less is exactly the mutual information:

$$I_{\text{mutual}} = S[X] + S[Y] - S[X,Y]. \tag{3.29}$$

One important example of mutual information is the Gaussian channel. To discuss this example we change notation a bit. We imagine that out in the world there is some signal $s$ (instead of the "input" $X$). We have an "$s$-detector" that provides a readout $y$ (instead of the "output" $Y$). On average, $y$ is proportional to $s$ with some gain $g$, but there is added noise:

$$y = gs + \eta. \tag{3.30}$$

We assume that the noise $\eta$ has a Gaussian distribution, which is often a very good approximation, and we assume that the signal $s$ also is chosen from a Gaussian distribution. These assumptions are formalized by the probability distributions $P(s)$ and $P(\eta)$:

$$P(s) = \frac{1}{\sqrt{2\pi \langle s^2 \rangle}} \exp\left[ -\frac{s^2}{2\langle s^2 \rangle} \right], \tag{3.31}$$

$$P(\eta) = \frac{1}{\sqrt{2\pi \langle \eta^2 \rangle}} \exp\left[ -\frac{\eta^2}{2\langle \eta^2 \rangle} \right]. \tag{3.32}$$

Because the output $y$ is completely determined by the input and the noise, we can write the conditional distribution $P(y|s)$ by finding the probability that the noise has the value required to satisfy Eq. (3.30):

$$P(y|s) = P(\eta = y - gs)$$

$$= \frac{1}{\sqrt{2\pi \langle \eta^2 \rangle}} \exp\left[ -\frac{(y - gs)^2}{2\langle \eta^2 \rangle} \right]. \tag{3.33}$$

Then we can calculate the distribution of readouts $y$ averaged over the inputs,

$$P(y) = \int ds\, P(y|s) P(s) = \frac{1}{\sqrt{2\pi \langle y^2 \rangle}} \exp\left[ -\frac{y^2}{2\langle y^2 \rangle} \right], \tag{3.34}$$

where the output variance $\langle y^2 \rangle = g^2 \langle s^2 \rangle + \langle \eta^2 \rangle$. Finally, we recall that the joint distribution can be written in terms of the conditional distribution and

the prior, $P(y,s) = P(y|s)P(s)$. These are all the ingredients we need to calculate the mutual information by substituting into Eq. (3.28). The steps of the calculation are outlined in section A.12, and the result is

$$I = \frac{1}{2}\log_2\left[1 + \frac{\langle s^2\rangle}{\langle \eta^2\rangle/g^2}\right] \text{ bits.} \tag{3.35}$$

We can express this result in a more intuitive form by thinking about our basic description of the system in Eq. (3.30). Rather than viewing our detector as a device that transduces the signal (multiplying by the gain $g$) and then adds noise, we can imagine that the noise is added to the signal itself, and then transduction is noiseless:

$$y = g(s + n_{\text{eff}}). \tag{3.36}$$

This procedure, which in this simple case defines the effective noise to be $n_{\text{eff}} = \eta/g$, is called "referring noise to the input" (Horowitz and Hill 1980). In general, the signal and the readout cannot be directly compared—for example, the signal could be mechanical (displacement in microns) and the readout could be electrical (Volts). This is especially true in the nervous system, where the inputs are continuous analog signals and the outputs are discrete spikes. Thus, to characterize the noise of the system we cannot simply measure a voltage noise or the variance of neural firing rate in some window; rather, we need to put this noise back into the same units as the input signals. It is this effective noise level that determines the detectability and discriminability of small signals, and it is this effective noise level that, when compared to the signal level, determines the mutual information. To see this we note that the variance of the effective noise is $\langle n_{\text{eff}}^2\rangle = \langle \eta^2\rangle/g^2$, so that in Eq. (3.35) the mutual information becomes

$$I = \frac{1}{2}\log_2\left[1 + \frac{\langle s^2\rangle}{\langle n_{\text{eff}}^2\rangle}\right] = \frac{1}{2}\log_2[1 + SNR], \tag{3.37}$$

where we define the signal to noise ratio ($SNR$) to be the ratio of the signal variance to the effective noise variance. This signal to noise ratio is dimensionless, so it is independent of the units we use for measuring the input and output.

Why is the Gaussian channel an important example? First, Gaussian distributions appear naturally when the quantity we observe is the sum of a large number of independent random variables; this is the content of the central limit theorem. The noise due to random arrival of photons at the retina becomes Gaussian at higher light intensities because large numbers of independent events are being summed by the photoreceptors. Similarly, if we observe single ion channels we see discrete transitions among states with different currents flowing across the membrane, but for a whole cell with many channels we can describe the noise as a Gaussian random current. Note that in each of these examples the elementary events have finite variance and they are independent of one another. If either of these assumptions is violated, the predictions of the central limit theorem can also be violated, and we should check for non-Gaussian distributions.



In information theory, Gaussian distributions are even more privileged. Imagine that we are trying to transmit information electronically, by driving current into a transmission line. The power we dissipate is proportional to the square of the current, assuming for the moment that the transmission line acts as a simple resistor. Efficient information transmission presumably means that we transmit as much information as we can at fixed average power cost. This means that we need to choose current signals from a distribution with a fixed mean square current—that is, with fixed variance. To transmit as much information as possible we need to maximize the entropy of these signals, so we need to know the distribution that has maximal entropy at fixed variance. The answer is the Gaussian distribution, as explained in section A.13. This result generalizes: If we have multiple variables and we fix the covariance matrix, the maximum entropy distribution is again a Gaussian distribution with this covariance matrix.

Gaussian distributions thus play a key role in information theory because they solve the problem of maximizing the entropy at fixed variance. To see how we use this fact, let us go back to our simple example of a detector that responds in proportion to the signal with added noise, Eq. (3.30). Suppose that the noise $\eta$ is Gaussian because it arises from lots of elementary random events, like ion channel noise, but that we don't really know the distribution of the signal $s$. The mutual information between the input $s$ and the output $y$ is, from Eq. (3.29), the difference in entropies,

$$I = S(s) + S(y) - S(s,y). \tag{3.38}$$

But the output $y$ is completely determined once we know the input $s$ and the noise $\eta$, and these two variables are independent of one another. Thus we can think of the combined entropy of $s$ and $y$ as really being the combined entropy of $s$ and $\eta$:

$$S(s,y) = S(s) + S(\eta). \tag{3.39}$$

Substituting, we see that the mutual information is just the entropy of the *output* minus the entropy of the noise,

$$I = S(y) - S(\eta). \tag{3.40}$$

This makes sense: In the absence of noise, every possible input signal corresponds to a unique output signal, and the information we gain by observing the output is the entropy of the outputs. In the presence of noise, the entropy of the outputs, which counts the number of distinguishable output states, is an overestimate of the information transmission because different outputs may correspond to fluctuations of the noise. The true information transmission is obtained by subtracting the entropy of this noise from the entropy of the outputs.

Now we can use the maximum entropy property of Gaussians to obtain an upper bound on the mutual information. We know that $S(y)$ must be *less* than the entropy of the Gaussian distribution with the same variance. But the variance of $y$ has two pieces, one from the variance of the noise and one from the variance of the signal. Since $y$ is just the sum of signal and noise, and by hypothesis the noise is Gaussian, assuming that $y$ is Gaussian is the same as assuming that the signal is Gaussian. Thus $S(y)$ is less than what we would calculate assuming that the signal $s$ comes from a Gaussian distribution whose variance is fixed at the true value of $\langle s^2 \rangle$. But if $S(y)$ is less than calculated from this Gaussian approximation, Eq. (3.40) shows us that the mutual information is less than what we calculate in the Gaussian approximation. We conclude that, if the noise is Gaussian, the mutual information is always *less* than what we calculated in Eq. (3.37), that is,

$$I \leq \frac{1}{2}\log_2\left[1 + \frac{\langle s^2 \rangle}{\langle n_{\rm eff}^2 \rangle}\right] = \frac{1}{2}\log_2[1 + SNR]. \tag{3.41}$$

Furthermore, we know that this maximum mutual information is reached when signals are chosen from a Gaussian distribution. The maximum mutual information is sometimes called the information capacity, although this is a slight abuse of the terminology (Shannon 1948).

Finally, the idea expressed in Eq. (3.40) is quite general. The information that a system transmits is equal to the entropy of its outputs minus the entropy of the 'noise' seen at the output. Since entropy must be positive—it is the logarithm of the number of alternatives, and the number of alternatives is at least one—this means that the transmitted information is always less than or equal to the output entropy. For us the most important application of this idea is to spike trains, where it implies that the information that a cell transmits



about signals in the outside world must be less than or equal to the entropy of the spike train. Thus, as emphasized in the previous section, the result of MacKay and McCulloch, Eq. (3.22) above, gives us a physical limit on the amount of information that neurons can transmit given the mean spike rate and some limited time resolution.

### 3.1.4 Time dependent signals

To use these ideas in a biological context we need to generalize a bit and consider signals that vary in time. From some sufficiently general point of view, nothing new happens when we think about time dependence, but in practice we need some new mathematical tools. Until this point we have talked about a signal $X$, and imagined that one could write down some set of numbers that defines this signal. When we think about time dependence, *one* example of a signal is already a function of time, and in principle one needs an infinite number of parameters to describe a function. The same problem arises in vision, where even at one instant of time the stimulus is a function of two spatial variables.

Intuitively we know that this infinity of parameters can't really be a problem. All the signals we are discussing can be filtered and digitized if we are careful enough, which means that if we look in a time interval $T$ at a function digitized with temporal precision $\Delta\tau$, there are only $T/\Delta\tau$ parameters, no matter how complex the function. Appealing to digitization in this way may settle any problems we have in talking about signals generated in the laboratory, but it is unsatisfying because it doesn't address the notion of "careful enough." One might worry that any digitization misses a little bit of the signal at each time step, and that these little bits add up to give a substantial contribution to the entropy or the mutual information. The alternative is that we learn to deal with the infinite numbers of parameters that describe continuous (not digitized) functions, which is what we do here.

In building up the tools to think about random functions of time, we take what may seem a rather long digression from our main themes. But we hope to communicate some understanding of the objects, like power spectra and correlation functions, that form the basis of the subsequent quantitative discussion. Our description of these ideas makes no pretense to mathematical rigor; we suspect that readers inclined toward rigor are already familiar with the ideas. Standard references on these matters include Papoulis (1965) and the delightfully brief text by Lighthill (1958). Translation of the mathematical ideas into computer programs for the analysis of real data is discussed by Press et al. (1992).

We are interested in describing a function of time, which we call $f(t)$, and we confine our attention to a time interval of size $T$, $0 < t < T$. All such functions can be written as Fourier series, that is, as a sum of sine and cosine functions with different frequencies, as schematized in Fig. 3.6:

$$f(t) = f_0 + \sum_{n=1}^{\infty} a_n \cos(\omega_n t) + \sum_{n=1}^{\infty} b_n \sin(\omega_n t). \tag{3.42}$$

To make this work, one has to choose the frequencies $\omega_n$ to "fit" into the time interval $T$, that is, $\omega_n T = 2\pi n$, where $n$ is an integer. Because of this fitting condition, all the sine and cosine terms have zero time average, so any time average component of $f(t)$ is carried by the constant term $f_0$; for the most part we ignore this term, since it can always be subtracted away, just as we define zero voltage when we make an electrical measurement. Notice also that, at least in principle, one needs an infinite number of Fourier coefficients, $a_1, a_2, \cdots, b_1, b_2, \cdots$, to describe the function. If we know the function, we can find these coefficients by doing the Fourier integrals:

$$a_n = \frac{2}{T} \int_0^T dt f(t) \cos(\omega_n t), \tag{3.43}$$

$$b_n = \frac{2}{T} \int_0^T dt f(t) \sin(\omega_n t), \tag{3.44}$$

$$f_0 = \frac{1}{T} \int_0^T dt f(t). \tag{3.45}$$

It is often convenient to combine the sine and cosine terms into the complex Fourier functions $\exp(-i\omega t) = \cos(\omega t) - i \sin(\omega t)$. In terms of these functions the Fourier series are written as

$$f(t) = \sum_{n=-\infty}^{\infty} f_n \exp(-i\omega_n t), \tag{3.46}$$

$$f_n = \frac{1}{T} \int_0^T dt f(t) \exp(+i\omega_n t). \tag{3.47}$$

Notice that each Fourier coefficient $f_n$ is a complex number, and it is often convenient to refer to the amplitude and phase of the $n^{\text{th}}$ component; these are real numbers $A_n$ and $\phi_n$, respectively, defined by

$$f_n = A_n \exp(+i\phi_n). \tag{3.48}$$

If the original function $f(t)$ is real then the coefficient with index $n$ has to be the complex conjugate of the coefficient with index $-n$, that is $f_n = [f_{-n}]^*$.

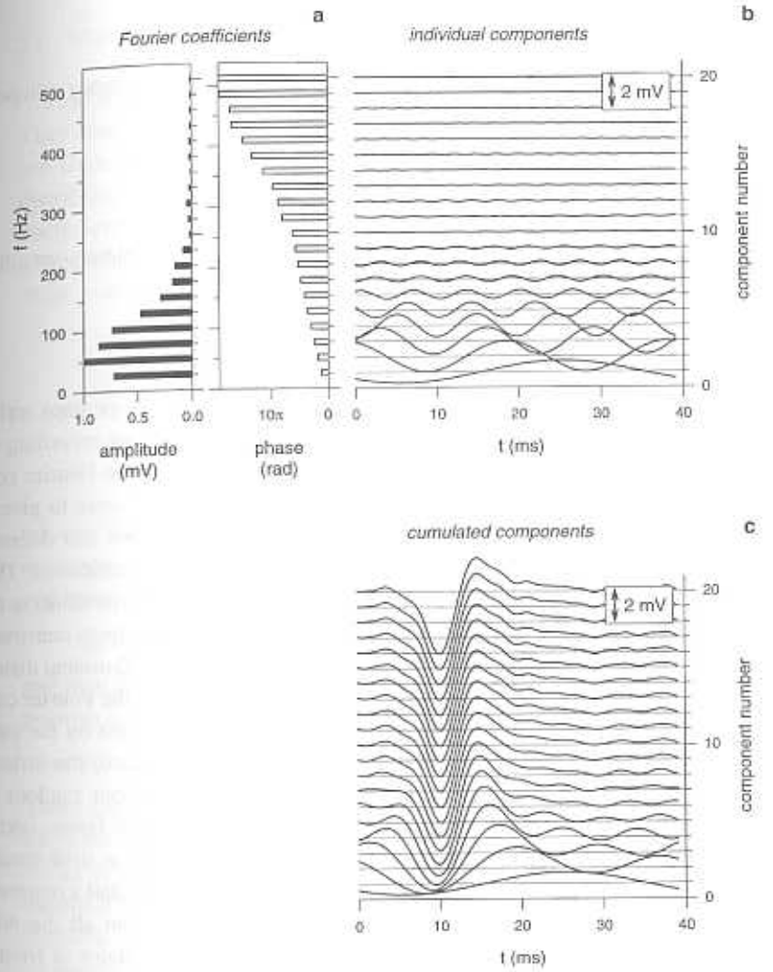



**Figure 3.6**
Construction of Fourier series. We consider a function which is defined in a window of 40 ms, so the frequency components are $\omega_n = 2\pi n(25\,\text{Hz})$. (a) Amplitudes and phases of the first 20 components. (b) Contributions of each component to the Fourier sum in Eq. (3.51). (c) The function $f(t)$ obtained by successive approximations in which we carry out the Fourier sum only with the first $N$ components, that is, with $1 \le n \le N$ in Eq. (3.51). Note that as we include more terms, more details of the function become visible, but this process stabilizes so that there are only very small changes after we have included ten components; in fact we can see from (a) and (b) that only the first nine components, corresponding to frequencies below 250 Hz, make significant contributions to the function. The function we construct here by Fourier synthesis is the graded voltage response of a blowfly large monopolar cell to a brief flash of light, as measured in the experiments of de Ruyter van Steveninck and Laughlin (1996a).

Then the amplitudes of the positive and negative frequency components are equal, and the phases are opposite, that is

$$A_n = A_{-n} \tag{3.49}$$

$$\phi_n = -\phi_{-n}. \tag{3.50}$$

Because of these symmetries, real functions are completely specified by the amplitude and phase of the positive components alone:

$$f(t) = f_0 + \sum_{n=1}^{\infty} A_n \cos(\omega_n t - \phi_n). \tag{3.51}$$

If we imagine choosing a function $f(t)$ at random—perhaps scribbling on a piece of graph paper without looking, or making a tape recording of sounds in the woods starting from some random time—then the Fourier coefficients $f_n$ of this function are chosen at random. Thus, if we want to give a precise definition to the "distribution of random functions," we can define this distribution $P[f(t)]$ by the distribution of the Fourier coefficients. The natural generalization of the Gaussian distribution for random variables is the notion of Gaussian random functions, by which we mean functions constructed from Fourier coefficients that are themselves chosen from a Gaussian distribution.

One might imagine that the Gaussian distribution of the Fourier coefficients is very complicated. But there is an important constraint on the structure of this distribution, stemming from the fact that our choice of the instant of time where $t = 0$ was arbitrary. That is, when we talk about random functions of time we usually assume that there is no clock that favors certain times over others—this is a notion of invariance with respect to time translation, or "stationarity." Thus, if one of the Fourier coefficients $f_n$ had a nonzero average value, then whatever random stuff was happening from all the other terms we could always find buried under it a nice clean oscillator of frequency $\omega_n$, and this would give a clock with which to distinguish, for example, different phases in the cycle of that sine wave. Therefore, the averages of the Fourier coefficients must all be zero.

More subtly, if two different Fourier coefficients, say $f_{37}$ and $f_{55}$, have zero average but nonzero correlations, then these two sine waves will beat against each other, and this beat will have an average amplitude, making a clock with frequency $\omega = |\omega_{37} - \omega_{55}|$. Again this clock destroys the notion that, on average, all instants of time are equivalent. Similar arguments tell us that the coefficients of sine and cosine have to be independent of each other, so that the phase $\phi_n$ at any given frequency is distributed uniformly from 0 to $2\pi$, again because the definition of phase is linked to our choice of $t = 0$.



Thus we see that Gaussian random functions are simpler than we thought: They consist of functions in which the Fourier coefficients $\{a_n, b_n\}$ are chosen from *independent* Gaussian distributions. A complete description of the distribution for the random function is given by listing the variances of these coefficients. In the complex Fourier representation of Eq. (3.46) we have to be a little more careful, in that the coefficient $f_n$ is the complex conjugate of $f_{-n}$. Thus, in this case, we write the variances as

$$\begin{aligned} \langle f_n f_{-m} \rangle &= \langle f_n [f_m]^* \rangle = 0 \qquad n \neq m, \\ \langle f_n f_{-n} \rangle &= \langle f_n [f_n]^* \rangle \\ &= \langle |f_n|^2 \rangle \\ &= \sigma^2(\omega_n), \end{aligned} \tag{3.52}$$

where $\sigma^2(\omega_n)$ is the variance or "power" in the Fourier component with frequency $\omega_n$. Note that we can write the complex number $f_n$ in terms of its real and imaginary parts,

$$f_n = \mathrm{Re}\, f_n + i\, \mathrm{Im}\, f_n, \tag{3.53}$$

and then $\mathrm{Re}\, f_n$ and $\mathrm{Im}\, f_n$ are independent Gaussian random variables for each positive component $n$, and they have equal variances

$$\langle (\mathrm{Re}\, f_n)^2 \rangle = \langle (\mathrm{Im}\, f_n)^2 \rangle = \sigma^2(\omega_n)/2. \tag{3.54}$$

These ideas about the probability distribution of the coefficients $\mathrm{Re}\, f_n$ and $\mathrm{Im}\, f_n$ are illustrated in Fig. 3.7.

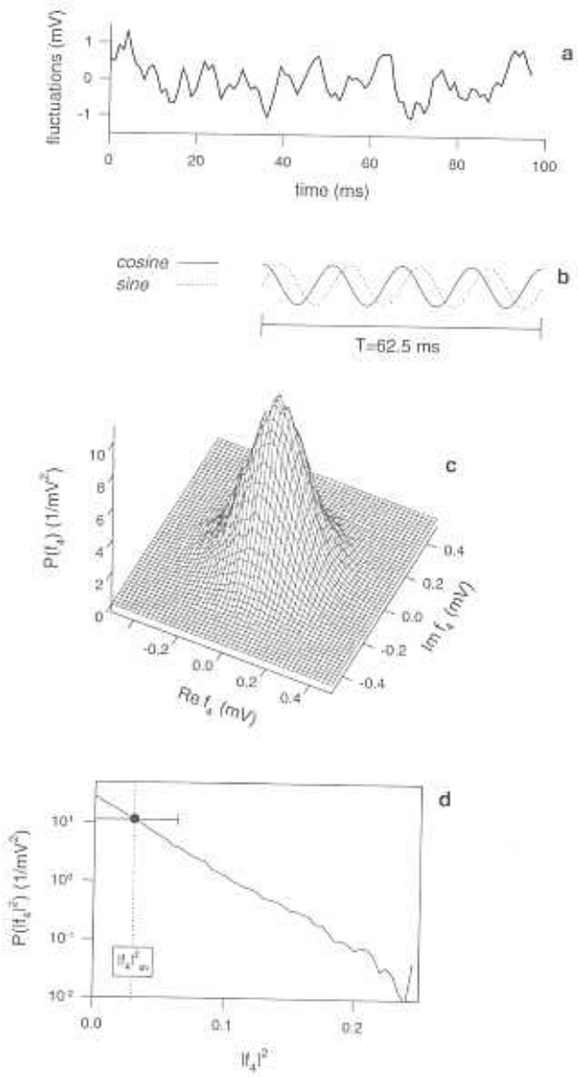
Let us now use the expressions for the variances of Fourier coefficients to calculate something we can observe more directly, namely the variance of the function $f(t)$ itself. The variance of $f(t)$ is given by substituting its Fourier expansion from Eq. (3.46):

$$\langle [f(t)]^2 \rangle = \left\langle \left[ \sum_{n=-\infty}^{\infty} f_n \exp(-i\omega_n t) \right]^2 \right\rangle \tag{3.55}$$

$$= \left\langle \sum_{n=-\infty}^{\infty} f_n \exp(-i\omega_n t) \sum_{m=-\infty}^{\infty} f_m \exp(-i\omega_m t) \right\rangle \tag{3.56}$$

$$= \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \langle f_n f_m \rangle \exp(-i\omega_n t) \exp(-i\omega_m t). \tag{3.57}$$

We notice from Eq. (3.52) that $f_n$ is correlated *only* with $f_{-n}$. This means that when we sum over all the values of $m$, only one term survives, the one with






$m = -n$, and all the other terms vanish. This gives us

$$\langle [f(t)]^2 \rangle = \sum_{n=-\infty}^{\infty} \langle f_n f_{-n} \rangle \exp(-i\omega_n t) \exp(-i\omega_{-n} t). \tag{3.58}$$

But the frequencies $\omega_n$ are just $n$ times the basic frequency $2\pi/T$, and $\omega_n = -\omega_{-n}$. This means that the exponential terms cancel, and all the $t$-dependence disappears, as it must: The variance of the random function must be the same for all values of $t$, because our definition of $t = 0$ was arbitrary (stationarity!). Finally, we use Eq. (3.52) to express the variance,

$$\langle [f(t)]^2 \rangle = \sum_{n=-\infty}^{\infty} \sigma^2(\omega_n). \tag{3.59}$$

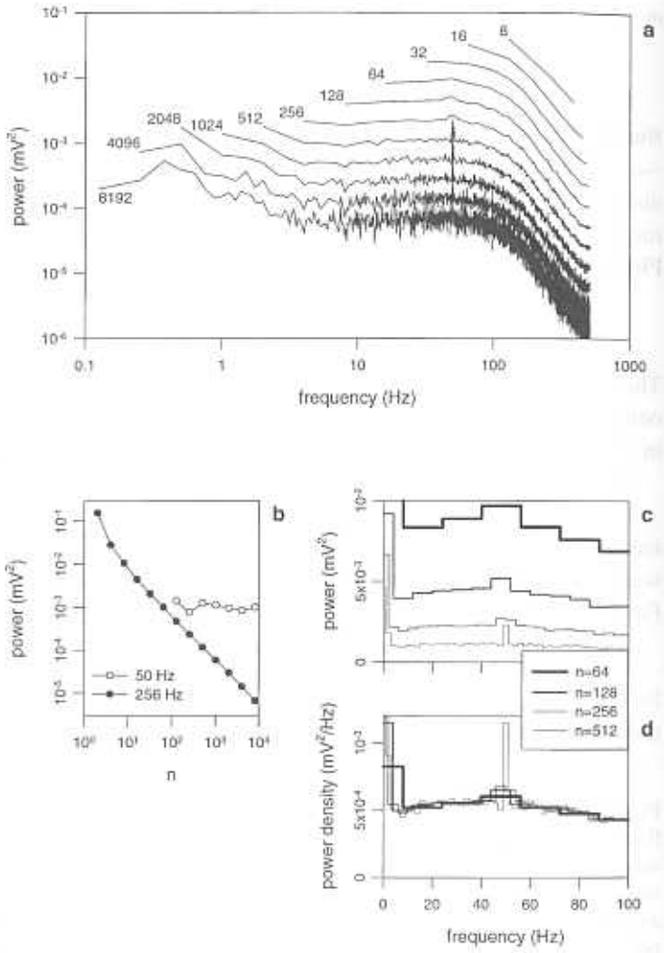
This result is very simple—the random function is made of many Fourier components, and the total variance of the function is the sum of the variances in the different components.

We should remember at this point that our characterization of the function $f(t)$ is tied to our choice of a fixed time window $0 < t < T$. This in turn means that our basic unit of frequency is $\omega_1 = 2\pi/T$. But if we let the time window become large, then this basic unit of frequency becomes very small. For example, if we look in a window of length one minute, we have

$$\omega_1 = 2\pi/T = 2\pi/(60\,\text{sec}) = 2\pi(0.016666\ldots\,\text{Hz}). \tag{3.60}$$

For many systems it makes sense to assume that if most of the "action" is in the neighborhood of, say, 10 Hz, which is $\omega_{600}$, the power at $\omega_{601}$—that is

**Figure 3.7**
Example of the probability distributions of Fourier components. Panel (a) shows a short segment from a 152 s recording of fluctuations in a blowfly photoreceptor cell in 1/1024 s bins. Panel (b) shows four periods of a cosine and a sine wave, each with a period of 16 bins, corresponding to 64 Hz. At every point of the recording we compute the inner product of the fluctuation waveform with the cosine component (real part) and with the sine component (imaginary part), and normalize this to the time window $T = 64/1024$ s. The joint probability distribution of the Fourier components at 64 Hz is the normalized 2-dimensional histogram of these two random variables shown in (c). This distribution approximates a 2-dimensional circular-symmetric Gaussian, $P(f_4) \propto \exp -[Re(f_4)^2 + Im(f_4)^2] = \exp -|f_4|^2$. Thus the power, $|f_4|^2$, has a negative exponential distribution, as shown in (d). It is a property of the negative exponential distribution that its standard deviation equals its mean. The dot in (d) represents the position of the mean, and the horizontal bar extends one standard deviation to the left and one to the right. Therefore, in measuring power spectra of a noise waveform, the standard deviation of the measured power is equal to the actual power. To make a more reliable estimate of the power at a certain frequency one can for example average the power spectra from independent samples of the waveform. See also Fig. 3.8.




at 10.01666 ··· Hz—is almost the same as at 10 Hz. This means that $\sigma^2(\omega_n)$ is becoming a smooth function of the $\omega_n$, and the fact that we sample this function at discrete frequencies becomes less and less important as the time window $T$ gets larger. To rid ourselves of a dependence on the window size, we would like to get at this underlying smooth function.

To see how this all works, let's go back to our formula for the variance of $f(t)$, Eq (3.59). The idea is that although we are summing over $\sigma^2$ evaluated at discrete frequencies $\omega_n$, there is really some smooth function $\sigma^2(\omega)$, and



**Figure 3.8**
Construction of the power spectral density. Panel (a) shows the mean square value $\langle |f_n|^2 \rangle$ of positive frequency Fourier components computed from the same data as used in Fig. 3.7. This was done for different window lengths $T$, specified here by the number of bins, $n$ (with $T = n/1024$ s). The total length of the noise trace was 155638 bins, so for the 8-bin window we have 155638/8=19456 examples of each Fourier coefficient. The means computed from this large number of examples have correspondingly small errors, but with these small windows we have only 8 independent Fourier components. The number of independent Fourier components increases linearly with $n$, while at the same time the number of independent examples of each component decreases, going down to 19 for $n$=8192. The decreasing number of examples means that the statistical errors in computing the means are larger, and hence the spectra look noisier for higher $n$. With one exception, the mean square value of each coefficient is proportional to $1/n$, as expected from the discussion of Gaussian random functions; this is demonstrated for the 256 Hz component (black dots) in panel (b). The exception to this behavior is found at 50 Hz, where there is a component that behaves as if $f_n$ has a nonzero mean value, independent of the window size. The experiment was done in the United Kingdom, which has a 50 Hz line frequency. The 50 Hz component can be clearly seen for the higher $n$ in (a), and its mean square value is essentially constant, as shown in (b). This distinguishes a pure sine wave, for which $\langle |f_n|^2 \rangle$ measures the square of the sine wave amplitude, from random noise, for which $\langle |f_n|^2 \rangle$ measures the variance of the Fourier coefficients (because $\langle f_n \rangle$ is zero for random waveforms). Panel (c) shows a section of the data in (a) on linear scales, making it clear that increasing the size of the window leads to decreasing variance per component, and to an increasing density of components. If we multiply the values of the variance in (c) by the density of frequency components, we finally arrive at the power spectral density shown in (d). This is a physically meaningful result in the sense that the value of the power spectrum stabilizes at large $n$. The exception is again the peak in the neighborhood of 50 Hz, which becomes narrower and higher as $n$ increases, preserving its total area. This makes sense, because a pure sine wave is represented by a delta function in the frequency domain (see Lighthill 1958).

we can think of the sum as an approximation to the integral of this function, as shown in Fig. 3.8. Thus we have

$$\langle [f(t)]^2 \rangle = \sum_{n=-\infty}^{\infty} \sigma^2(\omega_n) \tag{3.61}$$

$$= \sum_{n=-\infty}^{\infty} (\omega_{n+1} - \omega_n) \frac{\sigma^2(\omega_n)}{\omega_{n+1} - \omega_n}, \tag{3.62}$$

but we notice that the frequency differences are related to the size of our time window,

$$\omega_{n+1} - \omega_n = \Delta\omega = 2\pi(n+1)/T - 2\pi n/T = 2\pi/T, \tag{3.63}$$

so we can write the variance as

$$\langle [f(t)]^2 \rangle = \sum_{n=-\infty}^{\infty} \Delta\omega \frac{1}{2\pi} [T\sigma^2(\omega_n)]. \tag{3.64}$$

Finally, we replace the sum by the integral it approximates, a replacement that is strictly valid only in the limit that $T$ becomes very large. Thus, as $T \to \infty$,

$$\langle [f(t)]^2 \rangle \to \int_{-\infty}^{\infty} \frac{d\omega}{2\pi} \left[ T\sigma^2(\omega) \right], \tag{3.65}$$

where the natural object that emerges from the calculation is the smooth function

$$S(\omega) = \lim_{T\to\infty} T\sigma^2(\omega), \tag{3.66}$$

so that

$$\langle [f(t)]^2 \rangle = \int_{-\infty}^{\infty} \frac{d\omega}{2\pi} S(\omega). \tag{3.67}$$

The function $S(\omega)$ is the *power spectrum*, also called the power spectral density, and we see that it has units of the variance of $f$ times a factor of time, or "variance per Hz." Thus the units for the power spectrum of voltage fluctuations are (Volts)$^2$/Hz. For the power spectrum of fluctuations in firing rate of a neuron the units are (spikes/sec)$^2$/Hz or (spikes)$^2$/sec. The relation between the variance and the power spectrum, Eq. (3.67), is a special case of Parseval's theorem (Lighthill 1958).[8]

The interpretation of the power spectrum is as follows. Imagine that we look at signals in the neighborhood of frequency $\omega$ and average for about $\tau$ seconds, so that we are looking through a bandwidth of $\Delta\omega \sim 1/\tau$. Then we will see a variance $S(\omega)\Delta\omega \sim S(\omega)/\tau$. Note that this is dimensionally correct—if we measure voltage fluctuations, for example, the product $S(\omega)\Delta\omega$ has units [(Volts)$^2$/Hz]/[sec] = (Volts)$^2$. As we average for a longer and longer time $\tau$, the variance that we see through our averaging filter goes down in proportion to $\tau$, which means that the standard deviation of the fluctuations goes down in proportion to $\sqrt{\tau}$. This is the continuous time version of the familiar idea that we can reduce fluctuations by $\sqrt{N}$ if we make $N$ independent measurements, and is illustrated in Fig. 3.8b. If we make measurements at lots of different frequencies and add up the results, we obtain the total variance of Eq. (3.67). This is the same result as if we don't filter or average, because then we are sensitive to all frequencies instead of just a narrow band.

8. In modern experiments power spectra are almost always measured using numerical analysis of digitized data (Press et al. 1992). In this process it is very easy to lose track of the units in which the spectrum should be measured. This is unfortunate, because then we also lose the possibility of checking that the integral of the power spectrum is equal to the total variance.



If the signal we are measuring is a velocity, say in cm/sec, then the power spectrum has units of a diffusion constant, (cm/sec)$^2$ × sec = cm$^2$/sec. Indeed the diffusion constant of a particle can be computed from the low frequency limit of the power spectrum for velocity fluctuations. In the same way, the power spectrum of the spike train has units of spikes$^2$/sec, which is the diffusion constant for the spike count. Thus, if we count spikes in a window of size $T$, the variance of the spike count will grow in proportion to $T$ at large $T$, and the coefficient of proportionality is the low frequency limit of the power spectrum of the spike train. Some details are given in section A.2.

When we think about ordinary Gaussian random variables, the natural generalization of the variance is the covariance matrix. Thus, if we have variables $x$ labeled by an index $n$, and these variable have zero mean, then $C_{nm} = \langle x_n x_m \rangle$ is the covariance matrix. This matrix can be diagonalized by changing coordinates, that is, by considering appropriate linear combinations of the $x_n$. Once we transform to these new coordinates, the different variables are *independent* and their variances are given by the eigenvalues of the covariance matrix. The underlying independent variables are sometimes called principal components. How does this structure relate to our discussion of power spectra?

When we discuss continuous functions of time, we can think of the time $t$ as the analog of the index $n$ used for ordinary random variables; if we digitize the signal, then this analogy is exact. Then the analog of the covariance matrix is the *correlation function*

$$C(t,t') = \langle f(t) f(t') \rangle, \tag{3.68}$$

which we notice has the same units as the variance of $f$. Stationarity tells us that this function cannot depend on the absolute times (remember, they are measured relative to an arbitrary time zero) but only on the time differences. Thus $C(t,t') = C(t-t')$ is a function of only one time variable. We can compute this function in terms of the Fourier coefficients, as outlined in section A.14, and we find

$$C(\tau) = \int_{-\infty}^{\infty} \frac{d\omega}{2\pi} S(\omega) \exp(-i\omega\tau). \tag{3.69}$$

Thus the correlation function is the Fourier transform of the power spectrum, which is called the Wiener–Khinchine theorem (Papoulis 1965). More important for our discussion, we see that although the values of $f(t)$ at different

times are *not* independent, the different frequency components are independent. This means that the frequency components provide the coordinate system in which the covariance matrix is diagonal (principal components), and the power spectrum measures the variances of these independent variables.

As with the power spectrum, the correlation function has a simple phenomenological interpretation. If we imagine observing the function $f(t)$ for all time up to some point $t_0$, the correlation function tells us how far beyond $t_0$ we can predict $f(t)$. Indeed, for Gaussian random functions with zero mean, if you observe only the value of $f$ at $t_0$, then your best guess about the value of $f$ at some other time $t$ is $f_{\text{guess}}(t) = C(t - t_0)f(t_0)/C(0)$. As $t$ moves away from $t_0$ into the past or future one's knowledge at $t_0$ becomes less and less predictive, and the fluctuations of the true $f(t)$ around the value $f_{\text{guess}}(t)$ become larger. These ideas are illustrated in Fig. 3.9. One often talks about the *correlation time*, which measures the (approximate) width of $C(\tau)$, and hence the time scale over which knowledge of the function $f$ can be extrapolated.

To summarize, the power spectrum, which is a function of frequency, generalizes the concept of variance to the case of time dependent signals. For Gaussian distributions, stationarity tells us that different frequency components are statistically independent, which means that Gaussian random waveforms are really just a large collection of independent Gaussian random variables, and this makes things easy. Equivalently, we can say that the Fourier series provides us with a coordinate system that diagonalizes the covariance matrix, or correlation function, for time dependent signals. The Gaussian distribution is again singled out as the distribution that has maximal entropy consistent with the power spectrum.

To compute the information transmission for signals in the presence of noise, we note that since different Fourier coefficients are independent, the information carried by each coefficient can just be added up to give the total information. Once more we look at signals and noise in a fixed time window $0 < t < T$, and then we define a signal to noise ratio at each of the discrete frequencies $\omega_n$, $SNR(\omega_n)$. Then the information transmitted by each Fourier component is, from Eq. (3.37), $I_n = \frac{1}{2}\log_2[1 + SNR(\omega_n)]$, and the total information transmission is

$$I = \sum_{n=-\infty}^{\infty} I_n = \frac{1}{2}\sum_{n=-\infty}^{\infty} \log_2[1 + SNR(\omega_n)]. \tag{3.70}$$

If we now let the time window become large, we can use the same trick as in Eq's. (3.62) to (3.65), replacing the sum over discrete frequencies by an integral, so that



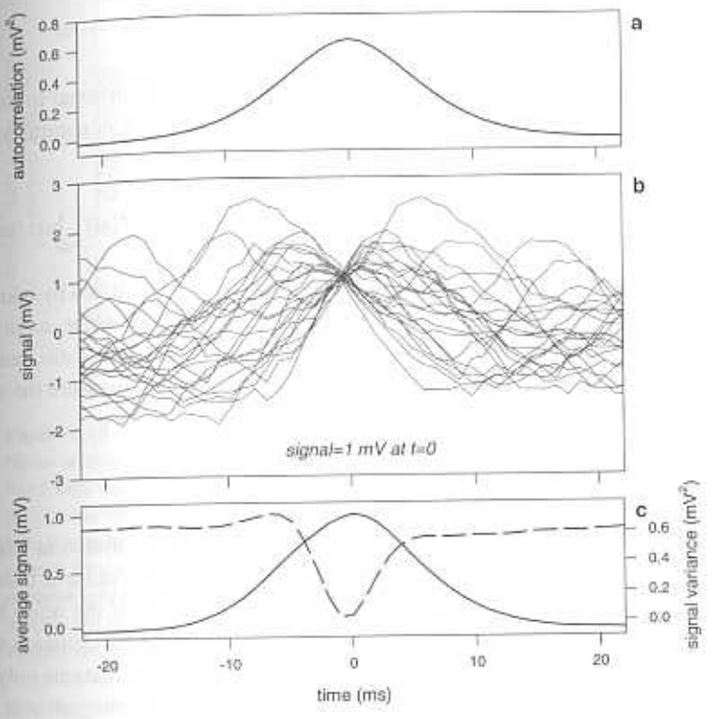



**Figure 3.9**
Correlation function and correlation time. The correlation function $C(t - t') = \langle s(t)s(t')\rangle$, shown in the top panel, measures the extent to which the signal $s$ at time $t$ is correlated with the signal at time $t'$. Because of stationarity, the correlation function depends only on the time difference $t - t'$, and not on the absolute time $t$. The correlation function is often condensed into a correlation time, which measures how quickly structure in the correlation function dies out as $t - t'$ increases. The middle panel illustrates this. Here we selected portions of the signal that passed through 1 mV between $t = -1$ ms and $t = 1$ ms. It is clear that, on average, the signal around $t = 0$ is higher than the mean. Also, near $t = 0$ the signals are closer to one another than they are at times far from $t = 0$. This is summarized in the bottom panel, which shows that the average waveform peaks at $t = 0$, whereas the variance of the set of waveforms has a minimum. The shape of the conditional mean waveform vs. time in the bottom panel matches the shape of the correlation function in the top panel, as expected for Gaussian noise, and this demonstrates that the correlation time defines a window of predictability in the waveform. The data are from the same photoreceptor cell as used in the previous two figures, but at a much lower light intensity, for which the cell's voltage noise has a substantially longer correlation time.

$$I \to \frac{1}{2} T \int_{-\infty}^{\infty} \frac{d\omega}{2\pi} \log_2 \left[ 1 + SNR(\omega) \right] \text{ bits.} \tag{3.71}$$

Thus, the amount of information transmitted is proportional to the time over which we observe the signal, which makes sense. It is natural to define the information transmission rate,

$$R_{\text{info}} = \lim_{T \to \infty} I/T = \frac{1}{2} \int_{-\infty}^{\infty} \frac{d\omega}{2\pi} \log_2 \left[ 1 + SNR(\omega) \right] \text{ bits/sec.} \tag{3.72}$$

For a single variable, the signal to noise ratio is constructed by taking the ratio of the signal variance to the variance of the effective noise, as in Eq. (3.37). In the case of time dependent signals, we form the same ratio, but now using power spectra rather than variances, since the spectra measure the variances of the Fourier coefficients, so that

$$SNR(\omega) = \frac{S(\omega)}{N_{\text{eff}}(\omega)}. \tag{3.73}$$

We have seen that to maximize information transmission at fixed variance we should choose Gaussian signals, assuming that the noise is Gaussian. This result generalizes to time dependent signals, so that if the noise is Gaussian we can maximize information transmission by choosing Gaussian signals. But how should we choose the power spectrum? Suppose that our only constraint is on the total signal variance. This variance is an integral over the power spectrum, as in Eq. (3.65), and we would like to maximize $R_{\text{info}}$ in Eq. (3.72) while holding this integral fixed. The solution to this optimization problem is quite remarkable (Shannon 1949), and is shown graphically in Fig. 3.10; for the derivation, see section A.15. The optimal signal spectrum is one that exactly complements the noise spectrum over a limited bandwidth, so that the total of signal plus noise is flat over that bandwidth. We must choose the bandwidth so that the variance comes out to the correct value. This means that optimal coding schemes produce an output that looks as much like white noise as possible, given the constraints.

The result in Fig. 3.10 is a special case of something more general. If one wants to maximize information transmission subject to some constraints, the optimal strategy is always to make the output of the communication channel look as random and "noise-like" as possible. Thus we have seen that if the noise comes from a Gaussian distribution, maximum information transmission occurs when the signal also comes from a Gaussian distribution. Similarly, for time dependent signals, maximum information transmission occurs when the



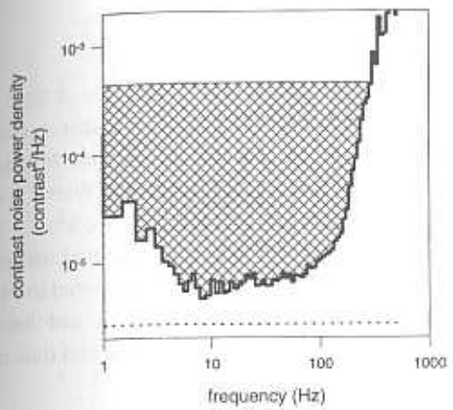



**Figure 3.10**
Noise whitening. To transmit the maximum amount of information possible given a fixed total signal variance, the power spectrum of the signals to be coded should be matched to the power spectrum of the noise in the system. The optimal situation shown here is one in which the total power—the signal power plus the noise power—is equal at each frequency. The power spectral density, given by the solid line, is the effective contrast noise power density of a blowfly photoreceptor. The cross-hatched area represents the optimal distribution of a fixed amount of signal contrast power over the various frequencies. This contrast power is distributed as if it were water in a vessel shaped as the effective contrast noise power spectrum. Hence the procedure is sometimes referred to as the "water filling analogy." See also de Ruyter van Steveninck and Laughlin (1996a).

sum of signal and noise is as close as possible to the "completely random" white noise. This is a crucial result, since it means that if one were recording from a neuron that optimized transmission from one point in the brain to another, its spike trains would look like complete junk! This is a cautionary tale, since it tells us that the observation of highly random spike trains might indicate extremely noisy neurons or it might indicate optimal coding. The way to distinguish these limiting pictures is to *measure* the information transmission rates, and see if they are in any sense optimal. We will show how to do this in the following sections.

As an example of these ideas, consider a single photoreceptor cell in the fly's eye. At reasonable background light intensities, these cells respond linearly to changes in the contrast $C(t)$, so that the cell voltage has a time dependence

$$V(t) = \int_0^\infty d\tau T(\tau) C(t-\tau) + \delta V(t), \tag{3.74}$$

where $T(\tau)$ is the response of the cell to a contrast pulse at time $\tau = 0$, and $\delta V(t)$ is the voltage noise. This is an example of the linear response models discussed in section 2.1.3, and we can measure the response function $T(\tau)$, or its Fourier transform $\tilde{T}(\omega)$, with methods in the spirit of Wiener's white noise technique. The important addition to the discussion in section 2.1.3 is that we want to characterize not only the deterministic component of the response, but also the random component $\delta V(t)$. This analysis is illustrated in Fig. 3.11.

If we present a time dependent contrast stimulus $C(t)$, and then repeat this input signal many times, we can average away the noise and thus measure the average voltage response,

$$\langle V(t) \rangle = \int_0^\infty d\tau T(\tau) C(t-\tau). \tag{3.75}$$

As in the discussion of electrical impedance in section 2.1.3, we can Fourier transform both sides of Eq. (3.75) to find that each Fourier component of the average voltage is proportional to the corresponding Fourier component of the stimulus, and the proportionality constant is just the response function $\tilde{T}(\omega)$:

$$\langle \tilde{V}(\omega) \rangle = \tilde{T}(\omega) \tilde{C}(\omega). \tag{3.76}$$

This equation relates the Fourier components of the average voltage, which we measure, to the Fourier components of the stimulus, which we control. Thus we know both $\langle \tilde{V}(\omega) \rangle$ and $\tilde{C}(\omega)$, and can find the response function $\tilde{T}(\omega)$ by taking their ratio. Obviously we have to choose a stimulus in which none of the Fourier components $\tilde{C}(\omega)$ are zero.

Once we have characterized the average voltage response, we can take the responses to each individual presentation of the stimulus and subtract the average, as in Fig. 3.11, which leaves us with many examples of the noise waveform $\delta V(t)$, one for each presentation. We can then Fourier transform each example, and compute the variance of the Fourier coefficients across our ensemble of examples. When the variances of the Fourier coefficients are normalized by the time window, as in Fig. 3.8 and Eq. (3.66), we obtain the power spectrum of the voltage noise $N_V(\omega)$ shown in Fig. 3.12. This analysis has been done for both the photoreceptors and the large monopolar cells in the fly's visual system (de Ruyter van Steveninck and Laughlin 1996a); the large monopolar cells are second order neurons that receive synaptic input from the photoreceptors. Results from both types of cells are shown in Fig's. 3.11 and 3.12 .



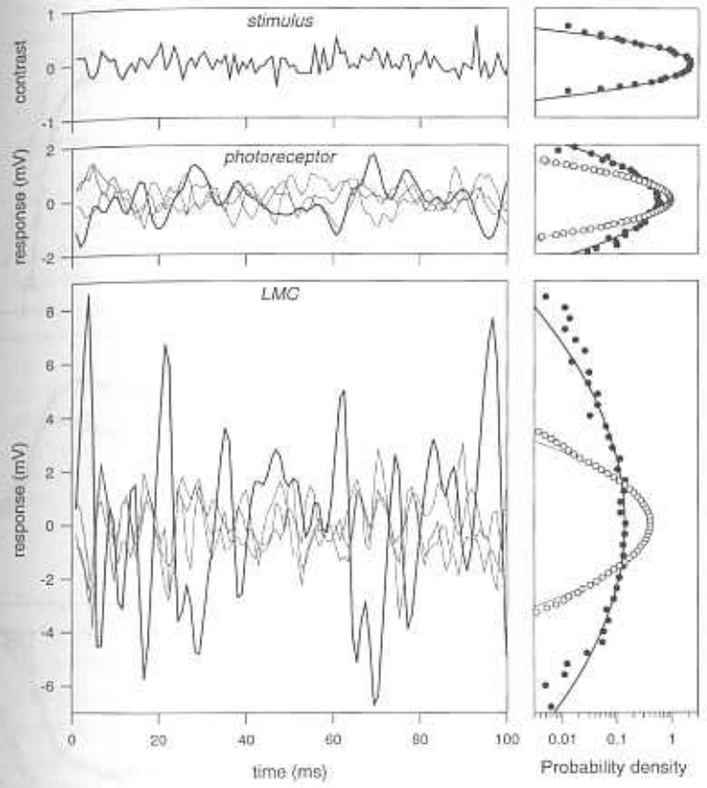
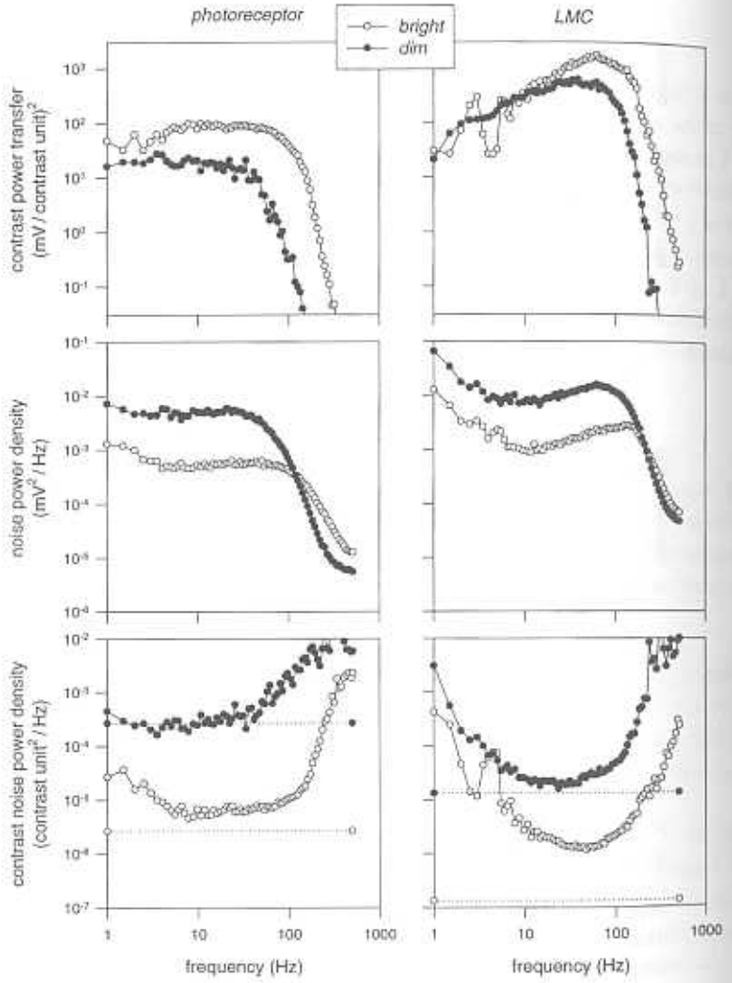



**Figure 3.11**
Stimulus and voltage responses in photoreceptors and LMCs. Left column: portions of the stimulus contrast signal $c(t)$ (top), and the averaged voltage responses (thick lines) with three samples of the fluctuations around the average (thin lines) recorded respectively from a blowfly photoreceptor (middle) and an LMC (bottom). LMCs, short for large monopolar cells, are directly postsynaptic to six photoreceptors in parallel. In the experiment the fly saw a light intensity modulated with a function $[1 + c(t)]$, with $c(t)$ the contrast waveform. This stimulus was repeated many times, and the responses $v_i(t)$ to $c(t)$ were averaged to get the ensemble-average waveform $\langle v(t) \rangle$. The fluctuations are the individual traces minus this average waveform: $\delta v_i(t) = v_i(t) - \langle v(t) \rangle$. The right column gives the probability distributions corresponding to these signals (filled circles: average waveforms, open circles: fluctuations around the average).




Following the discussion of Eq. (3.36), we can characterize the photoreceptor noise by referring it to the input, generating an equivalent contrast noise, shown in the bottom panels of Fig. 3.12. The spectrum of this equivalent contrast noise, $N_C^{\text{eff}}(\omega)$, is determined by the response function and the voltage noise power spectrum defined above,

$$N_C^{\text{eff}}(\omega) = N_V(\omega)/|\tilde{T}(\omega)|^2. \tag{3.77}$$



**Figure 3.12**
Contrast power transfer functions (top), noise power spectral densities (middle), and effective contrast noise power spectral densities (bottom) for a blowfly photoreceptor (left) and an LMC (right). Data are shown for two light intensities, a factor of 100 apart. The transfer function is defined as the Fourier transform of the average waveform divided by the Fourier transform of the stimulus contrast waveform. The noise power spectral density is the average power spectral density of all the fluctuation traces. And finally, the effective contrast noise power spectral density is the measured noise power spectral density divided by the square of the contrast transfer function. This simply transfers the measured noise power spectral density into an equivalent stimulus contrast. Shot noise analysis shows that an ideal photon detector should have an equivalent contrast noise power spectral density equal to one divided by the photon capture rate, independent of frequency. The dotted lines in the lower two panels give these physical limits. They represent $3.8 \times 10^5$ and $3.8 \times 10^3$ quantum bumps per second for the photoreceptor, and $7.5 \times 10^6$ and $7.5 \times 10^4$ bumps per second for the LMC. They are obtained by counting single photon absorptions in each cell at light levels that were a large calibrated factor lower than those used here. Over an appreciable frequency range the photoreceptor comes quite close to the ideal detector for both light intensities. The LMC starts to depart from ideal behavior only at the highest light intensity, probably as a result of the limited information capacity of the chemical synapse.

If each photon counted by a receptor cell triggers a stereotyped voltage pulse, or "quantum bump," then if the photons arrive from an ordinary light source we will have $N_C^{\text{eff}}(\omega) = 1/R$, where $R$ is the photon counting rate. The effective noise level cannot be lower than this physical limit, often called the shot noise limit. This limit is approached over a wide range of light intensities, at least up to some cutoff frequency where excess noise begins to appear due to randomness in the transduction mechanism itself. In the large monopolar cells one also expects to see excess noise due to randomness in synaptic transmission. To make a clean comparison between effective noise levels and the photon shot noise limit, one must calibrate the photon counting rate of each cell individually. This is done by maintaining the cell in a dark adapted state and counting the quantum bumps in response to dim, steady lights.

Given the effective contrast noise spectra of the receptor cells and of the large monopolar cells, how much information can the voltage responses of these cells provide about the visual world? This is an interesting computation because information that arrives at the large monopolar cells must have crossed a synapse, and so in this way we can make an estimate of the information capacity of the synapse itself. The difficulty is that we don't know the real ensemble of signals the fly encounters as it flies through the world. Recent work has characterized some of the spatial structure in randomly chosen snapshots of natural scenes (Field 1987; Ruderman 1993; Ruderman and

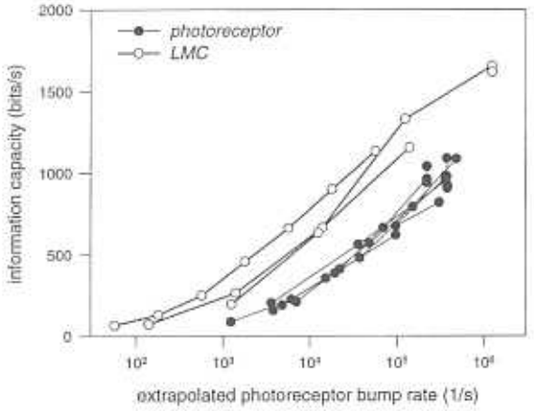



**Figure 3.13**
Information capacities of photoreceptors and large monopolar cells as a function of average light intensity. The information capacity is determined from the effective noise spectra in Fig. 3.12, using a total stimulus contrast variance of 0.1, and optimizing the stimulus contrast spectrum using the water filling analogy depicted in Fig. 3.10. We then get the rate of information transfer from Eq. (3.72). Information about light level conveyed by the LMC voltage must be transmitted through the photoreceptor-LMC synapse; thus these measurements determine a lower bound on the amount of information the synapse is capable of transmitting. The information transmitted by the LMCs is higher as a result of the convergence of six photoreceptors. See also de Ruyter van Steveninck and Laughlin (1996a).

Bialek 1994), but this doesn't tell us about the dynamics seen by any single receptor cell in time. What we do know is that the variance of the contrast in reasonably natural environments is about 0.1 if we look through an aperture the same size as that of the fly photoreceptor (Laughlin 1981; Ruderman and Bialek 1994). Thus we can formulate the problem as before: Given the effective noise spectrum and the contrast variance constraint, how can we choose the signal spectrum to maximize the information rate?

The procedure for maximizing the rate of information transmission problem is given again by the schematic of Fig. 3.10, and in Fig. 3.13 we see the resulting information rates plotted versus the photon counting rates for each type of cell. The information rates can be enormous, up to $1.65 \times 10^3$ bits/s in the large monopolar cells. Again we emphasize that this information must cross the synapse, and one can separate the LMC voltage noise into a component that is transmitted from the receptor cell and a component that is added by the synapse, and it is this synaptic noise that defines the limiting information capacity of the synapse itself. The final result is that receptor/LMC synapse can transmit information at a rate of $2 \times 10^3$ bits/s (de Ruyter van Steveninck and Laughlin 1996a).



Both the photoreceptor and the large monopolar cell produce graded responses, but there is an element of discreteness to synaptic transmission itself. Chemical synapses such as this one work by releasing transmitters, which are packaged in vesicles (Katz 1966). By analogy with our discussion of spike counting, we can imagine that the postsynaptic cell is a vesicle counter, and that vesicles are released (or counted) with some limited time resolution $\Delta\tau$. Then, if we know the mean rate of vesicle release, we can compute the maximum possible entropy of the vesicle count distribution, which is the same as for the spike count distribution in Eq. (3.24).

The synapse from photoreceptor to large monopolar cell has been very well studied anatomically (Meinertzhagen 1993). There is a convergence of six receptors onto one monopolar cell; because of the optics of the fly's eye, these six cells "look" in the same direction in space, and their signals are pooled. Each receptor makes roughly 200 synaptic contacts with the monopolar cell, so that the synaptic input to this one cell reflects a superposition of $1.2 \times 10^3$ boutons or active zones. This is comparable to the number of synaptic inputs converging onto a neuron in primary visual cortex.

Recent experiments suggest that the rate of vesicle release from a single active zone (albeit in a different species) does not exceed 150 vesicles/s (von Gersdorff and Matthews 1994). Using these observations as a guide, we expect that, with $1.2 \times 10^3$ active zones, the vesicle counting rate at the large monopolar cell is less than $1.8 \times 10^5$ s$^{-1}$. The effective contrast noise level of the cells saturates at photon counting rates of this order, suggesting that information transmission can be limited either by the discreteness of photon arrivals or by the discreteness of vesicle release. What is the time window the monopolar cell uses in counting the vesicle arrivals? The observed frequency response of these neurons suggests that this basic time resolution is of order 3–5 ms (de Ruyter van Steveninck and Laughlin 1996a, 1996b). We conclude that, viewed as a vesicle counter, the large monopolar cell is sampling time windows which contain an average count of (at most) 540–900 vesicles. From Eq. (3.24), the entropy of the vesicle count distribution must therefore be at most 10–11 bits in each time window. With 3 ms time resolution this corresponds to $\sim 3.7 \times 10^3$ bits/s, and of course information capacity declines as the effective time windows become larger. We see that the observed information capacity of the large monopolar cell is within a factor of two of the limit set by vesicle counting statistics.

This example of the first synapse in fly vision brings together several issues: the definition of effective noise levels; the comparison to physical limits set by photon shot noise; the use of maximum information arguments to determine the information capacity; and, finally, the use of maximum entropy arguments to determine the physical limit set by the discreteness of vesicles. We emphasize that, in computing the limit to information transmission imposed by the need to count vesicles, we have not used any model for the statistics of vesicle release. The bound that we calculate may be very generous if there is no simple mechanism to encode the receptor cell voltage variations into vesicle counts with the appropriate statistics. Nonetheless, the observed performance of the synapse comes close to the physical limit, encouraging us to think that the theoretical limits on information transmission are relevant to real neurons. We shall see this conclusion borne out in the analysis of spiking neurons as well.

## 3.2 INFORMATION TRANSMISSION WITH SPIKE TRAINS

Since the appearance of Shannon's pioneering papers in 1948–49, many authors have expressed the hope that information theory would provide a natural language for the discussion of neural coding, and perhaps even for the analysis of higher computational functions of the brain. Despite considerable effort, even measuring the information carried by a single spike train has been difficult. Some of the problems are related to the usual plague of insufficient data, but there are some more fundamental questions about what it means to make such measurements. Here we try to clarify these questions, then review some experiments that build an information theoretic analysis on top of the classical Adrian–Hartline experimental design. Finally we discuss the use of the stimulus reconstruction technique to quantify information transmission rates under conditions of continuous sensory stimulation.

### 3.2.1 Can we really "measure" information transmission?

Information theory as formulated by Shannon (and as taught in modern courses) is not an experimental science. Information theory is concerned primarily with calculating the limits on information transmission in a physical system starting from a hypothesized *model* of that system. The problem in neural coding is that we are given a real physical system (a sensory neuron) and asked to *measure* the information the system can transmit. Shannon didn't really tell us how to do this.



This problem is fundamental, as can be seen through connections between information theory and thermodynamics. Shannon's measure of available information is the thermodynamic entropy, so that the entropy of a gas is the amount of information we would gain if we learned the positions and velocities of each individual gas molecule (Brillouin 1962). Measuring the information carried by a neural spike train must thus be something like measuring the entropy of a box of gas or liquid. The problem is that, strictly speaking, the entropy is not an "observable" of the system—there are no entropy meters. One can measure entropy changes, but only because of the connection between entropy change and heat flow. As far as we know there is no information theoretic analog to heat flow that can be measured by universal instruments.

Given the thermodynamic analogy, we see that we cannot, in fact, "measure" information transmission in the literal sense. We can, however, try to *estimate* the relevant entropies. We would like to make controlled estimates, so that we know, for example, that the true value of the information transmission rate is larger than some directly measurable quantity.

Making reliable, controlled estimates requires that we understand something about how the system works. In the thermodynamic setting, we might start by approximating our system as an ideal gas, in which motions of all the molecules are independent. Of course, for many fluids this is a terrible approximation; near a critical point, for example, macroscopic numbers of molecules participate in correlated motions. To make meaningful estimates of the entropy, we thus need some ideas about the structure of the correlations in the fluid. Similarly, to make meaningful estimates of the information transmission in sensory neurons we will need to understand something about the structure of the neural code.

The fact that one can estimate but not measure information transmission thus means that there is no automatic information theoretic approach that boils the behavior of a neuron down to one number, the information transmission rate. On the contrary, the attempt to pull such a number out of experiments forces us to examine our understanding of the neural code itself.

### 3.2.2 Information transmission with discrete stimuli

The earliest application of information theory to neural coding is, as far as we can find,[9] the theoretical paper of MacKay and McCulloch (1952), which we

9. It is of some historical interest that Shannon himself used information theory to analyze the results of psychological experiments, using the linguistic knowledge of native speakers to place bounds on the entropy of English (Shannon 1951).

have described in detail. They estimated the limits on information transmission in spiking cells, that is, what is *possible* according to different hypotheses about the structure of the neural code. There is, of course, a difference between what is possible and what actually happens, and even MacKay and McCulloch expressed some skepticism about whether the limits on information transmission that they derived would be relevant to real neurons.

Here we review a few attempts to quantify information transmission in experiments with real neurons; we make no pretense to completeness. All of these examples share a common feature, namely that sensory signals are chosen from a discrete set, so that one might equally well view the experiments as discrimination experiments more analogous to traditional psychophysical methods. These discrete stimuli are presented (with one exception) in the usual fashion established by Adrian and Hartline; that is, the stimulus is turned on and left on with stationary parameters. We will see that one of the most interesting issues in these experiments is the way in which information transmission depends on the size of the time window used in analyzing the resulting neural responses.

Before discussing any experimental results we must issue a general caveat. Information is a probabilistic concept, which means that to attach a numerical value to information transmission we will need to manipulate probability distributions. But, any real experiment produces a finite amount of data, and from a finite number of samples we *never* know the true shape of a probability distribution. If we have some idea about the shape of the distribution we can try to describe this shape with a small number of parameters—e.g., the mean and variance of a Gaussian distribution—and then fit these parameters to the data. Even with such simplifications, there may be a large number of relevant parameters, since the input signals live in very high dimensional spaces, as illustrated by the response-conditional ensembles of section 2.2.3.

One way to make progress on the measurement of information transmission is thus to simplify the structure of the space from which the signals are drawn. Working in the cat visual system, Eckhorn and Pöpel (1974, 1975) chose to discretize both the spike arrival times and the stimulus waveform, so that each could be described as a binary string. In the case of spikes this is a reasonable scheme, since all practical analyses are done with spikes registered into time bins—although care must be taken to study the effects of different bin sizes. In the case of the stimulus waveform, this approach is limited to telegraph signals that jump between two discrete values, and this is highly restrictive unless the time bins are very small. Nonetheless, in this reduced stimulus space one can make considerable progress toward a complete probabilistic



characterization, and for cells in lateral geniculate nucleus Eckhorn and Pöpel found information transmission rates of up to ~ 50 bits/s, somewhat more than one bit per spike.

In a similar spirit, Richmond, Optican, and their collaborators have used principal component analysis to simplify the description of time dependent neural responses in the monkey visual system. The first experiments were done on cells in the inferior temporal cortex (Richmond et al. 1987; Richmond and Optican 1987; Optican and Richmond 1987), and subsequent experiments have used the same methods to study neurons in several different visual areas (McClurkin et al. 1991a, 1991b, 1991c; Richmond and Optican 1990; Eskandar, Richmond, and Optican 1992). In this approach one first filters the spike train to generate a smooth function of time, then approximates this smooth function as a sum of successively more complex waveforms (the principal components). Each instance of the spike train is then transformed into a set of coefficients, in much the same way that the Fourier series transforms a function of time into the discrete set of Fourier coefficients. In the analysis of spike trains, the first coefficient corresponds roughly to the spike count, and subsequent coefficients describe the tendency of the spikes to cluster at different times following the onset of the static visual stimulus.

The principal components method allows us to move beyond the description of neural responses in terms of spike count or firing rate. One can argue that this attempt is systematic and appropriately incremental, in that one adds additional terms in order of "importance," that is, in order of their contribution to the overall variance of the response. The essential claim of the Richmond–Optican experiments is that principal components beyond the first component carry significant information about the spatial structure of the visual stimulus.

Perhaps the most direct demonstration that there is information available beyond the total spike count is provided by experiments where two different spatial patterns give the same average number of spikes, as shown in Fig. 3.14. Looking at the raw data, it seems unlikely that one could distinguish between these stimuli based solely on the spike count from this neuron. On the other hand, it is clear that the temporal pattern of spikes during the stimulus presentation is very different in the two cases. Thus, by keeping track of the occurrence times of the spikes, rather than just their total number, one can discriminate between the two spatial patterns. This is a clear and rather robust result. In the context of a principal components analysis, one finds that these two different stimuli elicit, on average, the same first principal component in the neural response, but the second (or higher) components are different. From an information theoretic point of view, if we ask how much information the

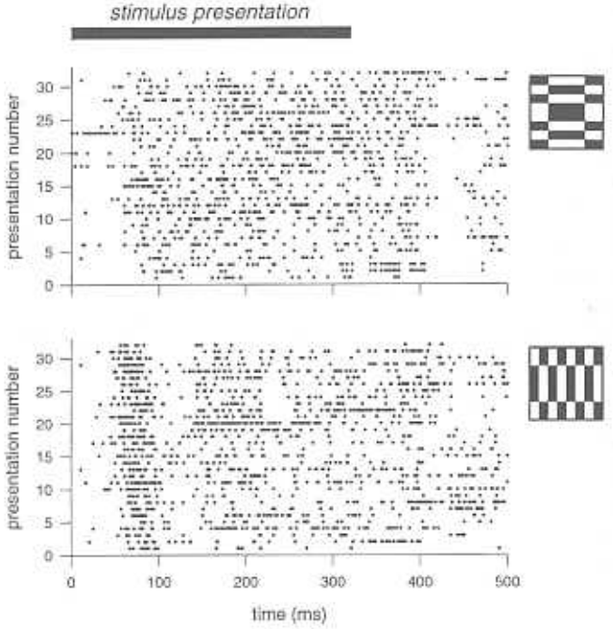



**Figure 3.14**
Responses of a complex cell in monkey striate cortex to two different spatial patterns. The duration of the stimulus presentation is given by the dark horizontal bar, and individual spikes are represented by dots as in Fig. 2.1. The mean number of spikes generated during the 400 ms following onset of the stimulus is similar in response to the two patterns. However, the probability of firing clearly has a transient peak shortly after stimulus onset for the pattern in the bottom panel; this peak is largely missing for the stimulus shown in the top panel. Furthermore, this distinction between the transient behaviors in response to the two stimuli is clear in each individual trial. Redrawn from the original data of Richmond, Optican, and Spitzer (1990), with our thanks to the authors.

neural response gives about the stimulus we will find that the inclusion of higher components provides more discrimination power and hence more information. Figure 3.14 conveys the basic point that the timing of spikes following stimulus onset is specific to the stimulus, and we emphasize that this point is independent of the use of principal components or even the technical apparatus of information theory. We suspect that similar results would be found in many systems.



For example, in the auditory nerve it is known that the transient response to the onset of a pure tone can depend on the frequency of the tone even when comparing two tones that give the same steady firing rate after the transients have settled. Carrying the point a little further, the conventional analysis of auditory neurons maps out contours in the amplitude–frequency plane corresponding to constant steady firing rate, as shown in Fig. 2.7. But, if we modulate a tone so that it moves along this amplitude–frequency trajectory, we don't really expect that the modulations will be undetectable in the cell's output. The fact that a cell gives the same steady firing rate all along a particular stimulus dimension does not mean that the cell is blind (or deaf, in this case) to realistic dynamics of the stimulus along this dimension.

Returning to the visual system, it is an old idea that the center and surround of classical receptive fields (Barlow 1953b; Kuffler 1953) have different dynamics, so that inhibition associated with the surround is slower than excitation in the center. Again this means that if we design two stimuli that each differentially excite center and surround, we can arrange that the steady firing rates for the two spatial patterns are the same, but the timing of the spikes at the onset of the stimuli will be different. Recently Golomb et al. (1994) have shown how this idea can be formalized for cells in the lateral geniculate nucleus. They studied a model in which neural firing is a Poisson process (see section 2.1.4) whose rate is modulated by the visual stimulus as seen through the spatiotemporal receptive field. For simplicity they assumed that modulations are linear, and the receptive fields were measured using the reverse correlation technique (see section 2.1.3). Golomb et al. found that many of the results from a principal components analysis of these cells (McClurkin et al. 1991a, 1991b) can be reproduced by the model. In addition, the model allows us to go back and study the dependence of information transmission on time following stimulus onset, as was also done in the experiments of Tovee et al. (1993).

Also working in IT cortex, Tovee et al. did a principal components analysis very similar to that of Richmond, Optican, and their collaborators but with varying time windows. Remarkably, they found that the majority of the available information could be extracted from observations of the spike train in very small windows, as small as 20 ms (Tovee et al. 1993). In their model for lateral geniculate nucleus responses, Golomb et al. (1994) found, in qualitative agreement with Tovee et al., that the bulk of the information the spike train carries about the "static" visual pattern is conveyed rapidly, within 100 ms of the stimulus onset. Of this 100 ms, the first 30 ms would seem to be true latency, since no information is conveyed in this time. In these time windows

the cells fire on average just a few spikes, so that an analysis that coarse-grains the neural output into firing rates or even principal components may be conceptually misleading. Both the work of Golomb et al. and that of Tovee et al. support our claim in section 2.2.1 that, for many sensory neurons, the natural time windows contain on the order of one spike.

Further evidence for the transmission of substantial information by small numbers of spikes can be found in the classic paper of Werner and Mountcastle (1965), who studied the skin mechanosensors in monkeys. This remains one of the standard references on the statistics of interspike intervals as a function of firing rate, and it is one of the first examples of a connection between the reliability of neurons and the reliability of perception, a topic we explore in detail in the next chapter. Part of their study, however, had an information theoretic flavor.

Werner and Mountcastle chose the amplitude $A$ of the stimulus—a static deflection of the skin—at random from $K$ possibilities. They then counted the spikes in some fixed time window following the stimulus onset, and repeated the stimuli enough times to obtain a reasonable estimate of the conditional probability distribution $P(n|A)$ of spike counts given the stimulus amplitude. Then the information the spike count provides about the stimulus amplitude is, by analogy with Eq. (3.28),

$$I = \frac{1}{K}\sum_n \sum_A P(n|A)\log_2\left[\frac{P(n|A)}{P(n)}\right], \tag{3.78}$$

where the overall probability of observing $n$ spikes is

$$P(n) = \frac{1}{K}\sum_A P(n|A). \tag{3.79}$$

By changing the number of possible stimuli $K$ it was possible to saturate the value of $I$, suggesting that a limit of performance for the neuron had been reached, rather than a limit imposed by the choice of stimulus ensemble. This saturation value of $I$ is approximately 3 bits, so that the spike count is sufficient to distinguish reliably among $\sim 8 = 2^3$ different stimulus amplitudes, and very similar results were obtained for several different neurons.

Werner and Mountcastle tried to connect the observation of 3 bits of information with the apparent limits to human cognitive processing studied by Miller (1956) in his famous paper "The magical number seven, plus or minus two." In retrospect this seems a bit far fetched, and perhaps this is one of the reasons that the information theoretic analysis in this paper is not widely cited.



For our present discussion, however, the most interesting aspect of the paper is that the authors studied information transmission by spike counts taken in windows of different sizes. In particular, they found that more than two bits of information could be gained by counting spikes in windows as small as 20–50 ms. But, for the typical firing rates of these neurons, these small windows contain, on average, just 5 to 15 spikes. From Eq. (3.24) we know that the entropy of the spike count distributions must therefore be less than 4 to 5 bits, so that the information about the stimulus is roughly half the available entropy. In one case (cell 25-1) Werner and Mountcastle provide the raw data from which we can compute the spike count entropy, and we find 4.2 bits under conditions where the information about the stimulus is reported as 2.5 bits.

We have emphasized the fact that the spike train entropy provides an absolute upper bound on the amount of information a neuron can transmit. Coding schemes that look only at the spike count are further limited by the entropy of the counting distribution. Werner and Mountcastle found that the amount of information carried by the spike count distribution could be as large as 60% of the count entropy, so that most of the variability in the spike count is in fact being used to encode the stimulus amplitude. Unless we look at the timing of spikes within the 20 ms windows, it is thus impossible to observe significantly higher information transmission rates. In addition, the information transmission saturates with integration windows of order 100 ms, comparable to the transient response time of the neurons, suggesting that the cell is optimized for transmitting information about rapidly varying signals.

If we return to our discussion in the introduction, we see that the Werner and Mountcastle results are very much in the tradition established by Adrian: Stimuli are static displacements, and responses are quantified by counting spikes. But now we see that cells can transmit almost all of the available information before the transient responses to stimulus onset have settled, and that this information is comparable to the limits imposed by our choice to count spikes. To determine the true information content of the spike train, then, we must choose stimuli with more realistic time dependencies and examine the neural response spike by spike.

We end this section with a note of caution concerning the amount of data required to draw meaningful conclusions about neural information transmission (Kjaer, Hertz, and Richmond 1994; Treves and Panzeri 1995). If we describe neural responses in terms of spike counts or firing rates, then it is relatively easy to collect data sets large enough to sample the distribution of spike counts conditional on each stimulus, as in the work of Werner and Mountcastle (1965). On the other hand, once we consider the possibility that the

arrival time of each spike carries information, the number of dimensions of our description increases dramatically and the number of "bins" in which we have to sample the probability distribution explodes. The spike train entropy quantifies this explosion. If we look at a time window of size $T$ and mark the arrival of spikes to accuracy $\Delta\tau$, then the entropy is, from Eq. (3.22), $S = \bar{r}T \log_2(e/\bar{r}\Delta\tau)$ bits, where $\bar{r}$ is the average firing rate over the window. We recall that the entropy measures the *logarithm* of the number of possible spike trains, so that there are roughly $2^S \sim (e/\bar{r}\Delta\tau)^{\bar{r}T}$ distinguishable spike trains, and in principle we need to see every one of these responses a few times before we can make model–independent estimates of the full probability distribution. But with $\bar{r} \sim 30$ s$^{-1}$ and $T \sim 300$ ms, a time resolution of $\Delta\tau \sim 5$ ms gives us $2^S \sim 10^{11}$. Thus the number of possible spike trains is of the same order as the (1995) United States government budget measured in dollars. No experiment will ever sample even a tiny fraction of these possibilities; no experimental animal (and few experimenters!) will experience even one-tenth this number of 300 ms intervals in a lifetime. Clearly one cannot jump from counting spikes to a "complete" analysis of timing codes without some hint about how to control this explosion of possibilities.

### 3.2.3 Stimulus reconstruction and information rates

In this section we will show that by learning to read the neural code we can place a rigorous lower bound on the amount of information a sensory neuron is transmitting (Bialek et al. 1993). The key idea is that our estimate of the stimulus cannot contain any information that wasn't actually present in the spikes. But the estimate is a continuous waveform, and we have lots of analytical techniques for asking how much information one analog signal (the estimate) provides about another (the stimulus). The result, Eq. (3.95), is that the rate of information transmission by the spike train is larger than a simple quantity that measures the quality of the reconstruction. This in turn provides us with an experimental strategy for estimating the rate of information transmission in real neurons, as schematized in Fig. 3.15. In the following sections this approach is used as an experimental tool.

This section is the most mathematical of all the main text, and it is really all about avoiding a mistake in the analysis of experiments. Specifically, we will discuss experiments that claim to estimate the rate of information transmission by sensory neurons. These experiments indicate that information rates are very large, close to the physical limits imposed by the spike train entropy. But we have emphasized that one cannot really measure information rates, only estimate them. Estimates contain random errors, which we can control by the



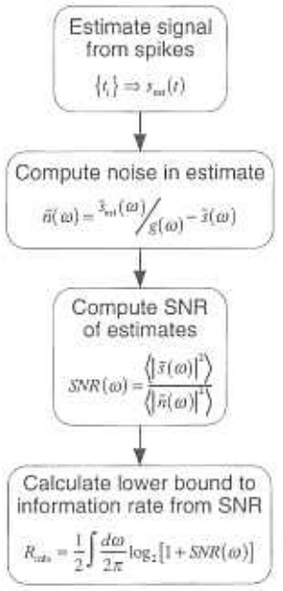



**Figure 3.15**
Strategy for measuring information rate. We can place a lower bound on the rate at which observation of the spike train provides information about the input signal by the procedure outlined here. First we estimate the input signal. We then measure the random errors, $\tilde{\eta}(\omega)$, in the estimate. These random errors determine the signal to noise ratio (SNR) of the estimate, and from the SNR we compute a lower bound to the information transmission rate. This bound will be close to the actual information transmission rate if the errors in our estimate are nearly Gaussian and if our estimation strategy adequately captures the structure of the code.

usual methods. But the entire estimation procedure may be biased, so that even with infinite data sets we do not converge to the correct answer. We want to control these systematic errors, and in particular we would like to be sure that we do not overestimate the performance of the neuron. Thus we want to make the conservative statement—this neuron transmits *at least* so many bits per spike—and be sure that the statement is correct. Technically this means that we want to give a lower bound on the information rate, and all of the mathematics in this section is concerned with constructing this lower bound.

From the general definition in Eq. (3.28) the information that the spike train provides about the stimulus is given by

$$I[\{t_i\} \to s(\tau)] = \int Dt_i \int Ds P[s(\tau); \{t_i\}] \log_2 \left( \frac{P[s(\tau), \{t_i\}]}{P[s(\tau)]P[\{t_i\}]} \right), \quad (3.80)$$

where we use the shorthand notation in which $\int Dt_i$ stands for integration over all arrival times $t_1, t_2, \cdots, t_N$ and summation over all spike counts $N$ on the time interval $0 < t < T$. In a similar shorthand, $\int Ds$ denotes an integration over all functions $s(t)$ for $0 < t < T$. $P[s(\tau)]$ is the a priori distribution from which the signal is drawn in a given experimental or natural situation. The goal of the calculations here is to massage this expression for information into a form in which we recognize terms that are accessible to experimental observation.

We can rewrite the information by remembering that the joint distribution of signals and spike trains, $P[s(\tau), \{t_i\}]$ can be decomposed into the conditional distribution of signals given the spikes—the response-conditional ensembles of section 2.2.3—and the distribution of spike trains $P[\{t_i\}]$:

$$P[s(\tau), \{t_i\}] = P[s(\tau)|\{t_i\}]P[\{t_i\}]. \quad (3.81)$$

But, with this factoring, the distribution of spike trains cancels from inside the logarithm of Eq. (3.80), and we expand the log into its two terms, obtaining

$$I[\{t_i\} \to s(\tau)] = -\int Ds P[s(\tau)] \log_2 P[s(\tau)] \quad (3.82)$$

$$- \int Dt_i P[\{t_i\}] \left[ - \int Ds P[s(\tau)|\{t_i\}] \log_2 P[s(\tau)|\{t_i\}] \right].$$

The first term is the entropy of the stimulus distribution, and the second term is just the entropy of the conditional distribution $P[s(\tau)|\{t_i\}]$, averaged over the distribution of spike trains $P[\{t_i\}]$. The entropy of the signals is determined by the setup of the experiment, so we need to work on the second term.

In the discussion of response-conditional ensembles (section 2.2.3), we illustrated the structure of the conditional distributions $P[s(\tau)|\{t_i\}]$ for simple choices of the spike sequences $\{t_i\}$. Even with these simple choices, we could only approximate these distributions as multidimensional Gaussians, or else we run out of data almost immediately. To evaluate the information transmission in Eq. (3.2.3) we really need to characterize the response-conditional ensembles for arbitrarily long spike sequences. This is, at first sight, completely hopeless. The escape from this apparent dead end begins with the maximum entropy idea.

The entropy of a distribution is always less than that of a Gaussian distribution having the same mean and variance. To exploit this fact we need to work



with the mean stimulus waveform and the variance of these waveforms *given* some particular spike train. Then we can put a *lower* bound on $I$ by approximating the conditional distribution as Gaussian. The mean stimulus waveform given a particular spike train is defined to be

$$\bar{s}(t; \{t_i\}) = \int Ds P[s(\tau)|\{t_i\}] s(t). \quad (3.83)$$

To define the variance of the conditional distribution, we have to remember that the fluctuations in $s(t)$ given the spike train are not stationary, since the spikes pick out certain specific times; surely the fluctuation in $s(t)$ would be large in a long time interval completely devoid of spikes. Thus we need a full covariance matrix, which we write as

$$\hat{N}(t, t'; \{t_i\}) = \int Ds P[s(\tau)|\{t_i\}][s(t) - \bar{s}(t; \{t_i\})][s(t') - \bar{s}(t'; \{t_i\})]. \quad (3.84)$$

If we think about digitized signals, where time is measured in discrete ticks of a clock, this is a large but quite ordinary matrix that can be manipulated with the usual rules of algebra. Indeed, this is the same sort of nonstationary correlation matrix that appeared in the discussion of the response-conditional ensembles. More generally, Eq. (3.84) describes a correlation function for the fluctuations $\delta s(t) = s(t) - \bar{s}(t; \{t_i\})$ under conditions where the usual invariance with respect to time translation is broken by the choice of particular spike times.

In the simple case that the ensemble of stimuli is itself Gaussian (although not necessarily white noise) the prior distribution $P[s(t)]$ can be completely characterized by the power spectrum or correlation function, as we have described before (section 3.1.4). For the moment it is useful to think about the correlation function or covariance matrix, which we write as $\hat{S}(t, t')$. We recall that this matrix depends only on the time difference $t - t'$, that we can diagonalize the matrix by going to a Fourier representation, and that the resulting eigenvalues are just proportional to the power spectrum. These facts become important a bit later, but for now we just manipulate $\hat{S}(t, t')$ and $\hat{N}(t, t'; \{t_i\})$ as ordinary matrices.

For Gaussian signals our problem reduces to computing the difference in entropy between two multidimensional Gaussian distributions with covariance matrices given by $\hat{S}(t, t')$ and $\hat{N}(t, t'; \{t_i\})$. We've already done this in the case of one dimensional Gaussians, in Eq. (3.10) and the surrounding discussion. One can solve the analogous multidimensional problem by transforming to coordinates where the covariance matrices are diagonal, doing the entropy calculation for each independent degree of freedom, summing up the results,

and transforming back to the original coordinates. One then has to check that the result is independent of the choice of coordinates, and it is. The result is that

$$I \geq \frac{1}{2} \int Dt_i P[\{t_i\}] \mathrm{Tr} \left( \log_2 \left[ \int dt' \hat{S}(t,t') \hat{N}^{-1}(t',t''; \{t_i\}) \right] \right), \quad (3.85)$$

where $\mathrm{Tr}(\cdots)$ denotes the trace of the matrix $(\cdots)$. The logarithm in this expression has a matrix as its argument. This matrix log can be computed by changing coordinates once again to diagonalize the matrix, replacing each diagonal element (eigenvalue) with its logarithm, and then transforming back to the original coordinate system. Note that the expression

$$\int dt' \hat{S}(t,t') \hat{N}^{-1}(t',t''; \{t_i\}) \quad (3.86)$$

is the product of the two matrices $\hat{S}$ and $\hat{N}^{-1}$, and that $\hat{N}^{-1}$ is just the matrix inverse of $\hat{N}$, that is,

$$\int dt' \hat{N}^{-1}(t,t'; \{t_i\}) \hat{N}(t',t''; \{t_i\}) = \delta(t - t''). \quad (3.87)$$

We emphasize that our expression for the information, Eq. (3.85), has a simple interpretation: From the covariance matrix of the fluctuations in the signal given a particular spike train and the covariance of the signal itself we can calculate a lower bound on how much information that particular spike train gives about the signal. Then we average over all spike trains, weighted by their probability of occurrence, to find a lower bound on the average information transmission.

Although the idea behind Eq. (3.85) is simple, the mathematics is still a bit complicated, because we have to compute $\hat{N}$ for each spike train, then average over spike trains after taking a logarithm. We can make things a little bit better by noticing that for any distribution of $x$, the average of the logarithm of $x$ is smaller than the logarithm of the average, that is $\langle \log x \rangle \leq \log \langle x \rangle$. This relation is a restatement of the fact that a plot of $\log x$ versus $x$ curves downward, as illustrated in Fig. 3.16, and this means that we can take the average inside the logarithm and preserve the inequality. If we write $\log_2(\hat{N}^{-1}) = -\log_2(\hat{N})$, and take the average inside the log, then we need

$$\langle \hat{N}(t,t'; \{t_i\}) \rangle = \int Dt_i P[\{t_i\}] \hat{N}(t,t'; \{t_i\}). \quad (3.88)$$

This average covariance then measures the fluctuations of the signal around its conditional mean, but the fluctuations are themselves averaged over all



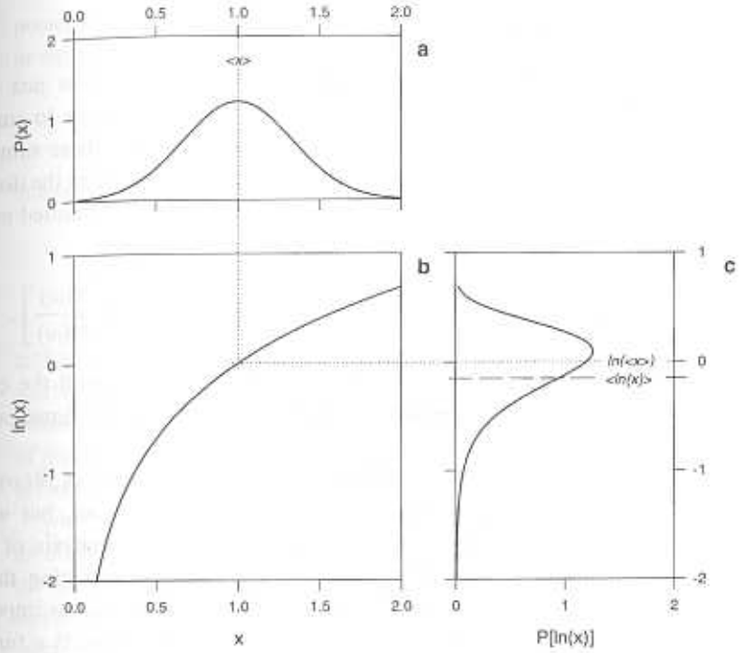



**Figure 3.16**
Concavity of $\ln(x)$ and the inequality $\langle \ln(x) \rangle < \ln\langle x \rangle$. The concave shape of $\ln(x)$ causes the average of $\ln(x)$ over $x$ to be less than the logarithm of the average of $x$. In the example shown in panel (a), $x$ has a symmetric distribution with mean 1. The natural logarithm of $x$ shown in (b) has the skewed distribution shown in (c). This concave transformation compresses the region of $x$ where $x > 1$, and expands the region where $x < 1$. As a result, the mode of $P[\ln(x)]$ is shifted up, but the mean of $P[\ln(x)]$ is lowered: $\langle \ln(x) \rangle < \ln(\langle x \rangle)$.

possible spike trains. These average fluctuations are stationary, since we are no longer keeping track of particular spike arrival times. Thus, the average covariance is a function only of the time difference $t - t'$, that is,

$$\langle \hat{N}(t,t'; \{t_i\}) \rangle = \bar{N}(t - t'). \quad (3.89)$$

Now the correlation matrix of the noise has the same structure as the correlation matrix of the signal, namely that it depends only on time differences, so we know that the eigenvalues of $\hat{S}$ are the power spectrum of the signal and the eigenvalues of $\hat{N}$ are the power spectrum of the noise. To fix the notation, we define explicitly the noise spectrum:

$$\bar{N}(\omega) = \int d\tau \exp(+i\omega\tau)\bar{N}(\tau). \tag{3.90}$$

Instead of fiddling with matrix manipulations, we can now just work with the eigenvalues. Then taking the trace means that we have to sum over all the discrete frequencies, and we have already seen how these sums turn into integrals when we let our time window $T$ become large, as in the discussion of Eq. (3.65). When all the dust settles, we find that the transmitted information is bounded by a simple expression:

$$R_{\rm info} = \lim_{T\to\infty} I[\{t_i\} \to s(\tau)]/T \geq \frac{1}{2}\int_{-\infty}^{\infty}\frac{d\omega}{2\pi}\log_2\left[\frac{S(\omega)}{\bar{N}(\omega)}\right], \tag{3.91}$$

where $\bar{N}(\omega)$ is the power spectrum of fluctuations around the conditional mean waveform, averaged over spike trains, and $S(\omega)$ is the signal power spectrum, as before.

We are still not quite finished, since the power spectrum $\bar{N}(\omega)$ describes fluctuations of the stimulus around the conditional mean, but we haven't explained how to find the conditional mean. In the analysis of response-conditional ensembles, the mean is constructed by searching through the experiment for many examples of each spike train, but this is impossible for long spike trains. What can we do? The conditional mean is a function that maps the spike arrival times $\{t_i\}$ into a value of $s$ at each time $t$, which we have written as $\bar{s}(t;\{t_i\})$. Suppose that we try to guess this function. Given our guess, we can subtract it from the real stimulus to form $\delta s(t)$ and compute the power spectrum of these deviations. What if our guess is wrong? Then the power spectrum we compute will always be *larger* than the true power spectrum $\bar{N}(\omega)$!

We can think of our guess at the conditional mean as being an estimate of the stimulus given that we have seen the spike train. The resulting $\delta s(t)$ is the error in our estimate, and the power spectrum of these errors is just the conventional mean square error measure (or $\chi^2$) taken frequency component by frequency component. The crucial point is that, to minimize the mean square error, the best estimator is the conditional mean—we have already used this in section 2.3.1, and details are in section A.7. Thus the mean square error of any other estimator will be larger than the fluctuations of the signal around its conditional mean. We emphasize that these statements are useful even though we don't know the true conditional mean. Specifically, suppose that we construct some *arbitrary* estimator that takes as input the spike train $\{t_i\}$ and returns some estimate of $s(t)$, which we call $s_{\rm est}(t;\{t_i\})$. Then the



power spectrum of errors in this estimate $N_{\rm est}(\omega)$ will always be greater than or equal to $\bar{N}(\omega)$. Hence,

$$R_{\rm info} \geq \frac{1}{2}\int_{-\infty}^{\infty}\frac{d\omega}{2\pi}\log_2\left[\frac{S(\omega)}{N_{\rm est}(\omega)}\right], \tag{3.92}$$

where the noise spectrum is defined by

$$N_{\rm est}(\omega) = \int d\tau \exp(+i\omega\tau)\langle[s(t) - s_{\rm est}(t;\{t_i\})][s(t') - s_{\rm est}(t';\{t_i\})]\rangle, \tag{3.93}$$

and now the average $\langle\cdots\rangle$ denotes an average over the ensemble of signals and spike trains that occur in response to those signals. If we choose a bad estimator, this bound will be far below the true information rate.[10] The ratio of this bound to the true information rate gives us the fraction of the available information about $s(t)$ that is captured by the estimated waveform.

We thus arrive at a strategy for experiments: Construct a box that takes as input the spike train $\{t_i\}$ and delivers as output an estimate $s_{\rm est}(t;\{t_i\})$ of the unknown, continuous stimulus chosen from the ensemble $P[s(\tau)]$. If we can parameterize this box, choose the parameters so as to minimize the mean square deviation ($\chi^2$) between the estimate and the true signal. Finally, the power spectrum of the errors will provide a lower bound on the information rate through Eq. (3.92).

We emphasize that the statements in the preceding paragraph are statements of mathematical fact. There is no assumption that the nervous system is "interested" in reconstructing the stimulus waveform exactly, nor that $\chi^2$ is the biologically relevant measure of error in the reconstruction. In the present context, stimulus reconstruction is just a tool to transform the variability of spike trains back into an equivalent variability of the stimulus, and the $\chi^2$ measure of error is singled out by the maximum entropy property of Gaussian distributions.

We have defined several different noises in this discussion, all of which are called $N$ (and we will define one more below): There is $\hat{N}$, that describes the true fluctuations in the signal $s(t)$ given some particular spike train. Then there is $\bar{N}$, which is the average of $\hat{N}$ over all possible spike trains. Finally

10. If we choose a sufficiently bad estimator it could even turn out that the right hand side of Eq. (3.92) is negative. Then we know that the true information rate is larger than some negative number. This is correct (we said that you could choose any estimator), just not very helpful.

there is $N_{est}$, which describes the errors in estimating the signal. Still, our final expression, Eq. (3.92), looks a bit funny when compared with the more conventional Eq. (3.37). The reason is that the "spectrum of errors" in Eq. (3.92) may include both systematic and random errors. Thus, if for some reason we fail completely in our attempt to reconstruct the signal from the spike train, the best we can do is to guess that the signal was zero, which is its average value in the ensemble $P[s(\tau)]$. But then, our errors are just as big as the signal, so $N_{est}(\omega) = S(\omega)$ and $R_{info} \to 0$, which makes sense. In this extreme example, our errors are not random, but rather are perfectly correlated with the signal.

Although estimating the information rate does not require us to distinguish random from systematic error, it is conceptually useful to make this distinction. Thus we would like to characterize the quality of the reconstruction in terms of an effective noise referred to the input, as in the discussion of Eq's. (3.36) and (3.37). To separate systematic and random errors, we write, for each frequency,

$$\tilde{s}_{est}(\omega) = g(\omega)\left[\tilde{s}(\omega) + \tilde{n}_{eff}(\omega)\right], \tag{3.94}$$

where $g(\omega)$ is a frequency dependent gain introduced to correct for systematic errors, and $\tilde{n}_{eff}(\omega)$ is effective input noise. To calculate $g$ and $n_{eff}$, we divide the experiment into segments, each of length $\tau_0$. Then we Fourier transform the stimulus and reconstruction in each segment. The result is a collection of Fourier coefficients $\{\tilde{s}^n(\omega), \tilde{s}^n_{est}(\omega)\}$, where $n$ numbers the segments. For each frequency $\omega$ we make a plot of stimulus Fourier coefficients $\tilde{s}^n(\omega)$ against reconstruction Fourier coefficients $\tilde{s}^n_{est}(\omega)$. Each segment contributes one point to this plot. The gain $g(\omega)$ is the slope of the best linear fit through these points, and the effective noise $\tilde{n}_{eff}(\omega)$ measures the scatter, along the stimulus axis, about this best fit line. Two such scatter plots are shown in Fig. 3.17, corresponding to the reconstructions of Fig. 2.20.

One reason it is useful to separate random from systematic errors is that the optimal estimator in the least squares sense *always* underestimates the true signal. In the limit that noise in the system is very large, this underestimation becomes very serious, and the estimate approaches zero. Under these conditions, the spectrum of errors $N_{est}(\omega)$ approaches the spectrum of the signal, as noted above. The effective noise, on the other hand, removes the systematic component of the errors, so that frequency bands where the reconstruction is very poor are revealed as having a large effective noise level. Furthermore, by removing systematic errors we obtain a noise measure that is uncorrelated with the stimulus.



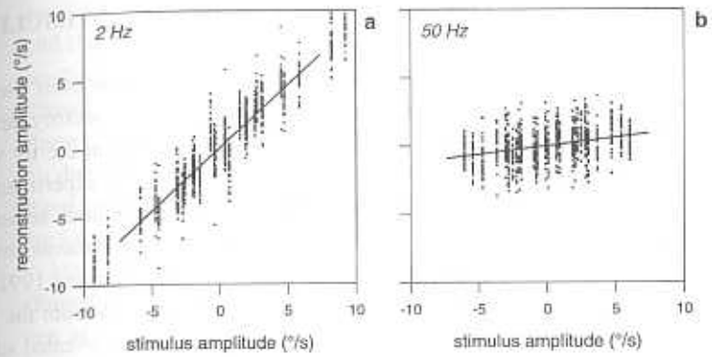



**Figure 3.17**
Scatter plots of estimation errors from H1 experiment at 2 Hz (a) and 50 Hz (b). The stimulus and estimate are each broken into segments of fixed duration, approximately two seconds in this example. We take the Fourier transform of each segment to generate a set of points $[\tilde{s}^i(\omega), \tilde{s}^i_{est}(\omega)]$, where the superscript $i$ counts the segments. We then create, at each frequency, a scatter plot of $\tilde{s}^i_{est}(\omega)$ against $\tilde{s}^i(\omega)$; each segment contributes one point to the scatter plot. The slope of the best-fit line is the gain, $g(\omega)$, and the scatter about this line measured along the $x$-axis is the effective noise, $\tilde{\eta}(\omega)$.

The effective noise level can be quantified by measuring its power spectrum $N_{eff}(\omega)$, which is just the variance of the Fourier components $n_{eff}(\omega)$ normalized by the time window $\tau_0$, $N_{eff}(\omega) = \langle|\tilde{n}_{eff}(\omega)|^2\rangle\tau_0$ (see the discussion in section 3.1.4 and Fig. 3.8). If we plot the effective noise spectrum as a function of frequency, we expect that the usual idea of a neuron being tuned to certain frequency bands, as in the auditory system, will be recovered as a band of frequencies where the effective noise is low. Finally, if we rewrite Eq. (3.92) for the information transmission rate in terms of the effective noise level, we find, by analogy with Eq. (3.72), that

$$R_{info} \geq \frac{1}{2}\int_{-\infty}^{\infty}\frac{d\omega}{2\pi}\log_2\left[1 + \frac{S(\omega)}{N_{eff}(\omega)}\right]. \tag{3.95}$$

Thus there is a natural notion of signal to noise ratio in the reconstructions, and this is the ratio of signal power spectrum to the effective noise spectrum, $SNR(\omega) = S(\omega)/N_{eff}(\omega)$. Then we can use the standard Shannon formula, Eq. (3.72), and calculate the rate of information transmission. The work of this section has been to show that this procedure is guaranteed to underestimate the true information transmission rate. This strategy for bounding the information transmission rate is summarized in Fig. 3.15.

## 3.3 ENTROPY AND INFORMATION WITH CONTINUOUS STIMULI

The approach described in the previous section has now been used to measure information transmission rates in a wide variety of sensory neurons. Initial experiments on the movement sensitive neuron H1 in the fly visual system (Bialek et al. 1991) were followed by a series of experiments on primary sensory neurons, the cells that first convert continuous sensory stimuli into discrete spikes. The first generation of these experiments involved the mechanosensitive cells in the cricket cercal system (Warland 1991; Warland et al. 1992) as well as the acoustic and vibratory sensors from the frog inner ear (Rieke 1991; Rieke et al. 1992). These experiments revealed surprisingly high rates of information transmission, approaching the physical limits set by the entropy of the spike trains themselves (Rieke, Warland, and Bialek 1993). Subsequent experiments have used the same methods to characterize information flow in higher order neurons of the cricket system (Theunissen 1993), and to explore the way in which more naturalistic stimuli are coded by the frog ear (Rieke, Bodnar, and Bialek 1992, 1995). Stimulus reconstruction methods have also been used to quantify information transmission by the array of ganglion cells in the tiger salamander retina (Warland and Meister 1993, 1995) and by the "probability coding" afferents of electroreceptive fish (Wessel, Koch, and Gabbiani 1996).

It is clear that each different sensory system poses new questions, but at the same time, measurements of information transmission have revealed some elements of universality across the systems studied. In trying to present these results, it thus makes sense to start with the "simplest" system and work our way toward the more complex. Of course, this is a dangerous classification to make, but there is a sense in which the cricket experiments are the cleanest probe of the coding problem. In this case no synapses intervene between the physical stimulus and the spike train of the first spiking cell in the sensory pathway, and one can deliver controlled stimuli by grabbing hold of the sensory hair. We thus study the encoding of time dependent signals in spike trains without the complications of preprocessing by a network of neurons, as in the retina, or by the mechanical structures of the vertebrate inner ear. In this section we therefore look first in some detail at the cricket cercal system, and then turn to neurons that are attached (distantly) to backbones.

### 3.3.1 Mechanical sensors in the cricket cercal system

Crickets, cockroaches, and related insects have two pronglike structures protruding from their rear ends (Huber, Moore, and Loher 1989). These are the cerci. Each cercus is covered with up to several thousand hairs of various sizes and shapes, each of which grows out of a single sensory cell; each cell in turn sends an axon into the abdominal ganglion. Hairs can be divided into three classes: touch sensitive bristle hairs that are too short to be significantly deflected by air flow; long, thin filiform hairs that are primarily responsive to air displacement; and massive clavate or "gravity sensitive" hairs. In addition, there are campaniform sensilla that sense deflection of the cercal surface; these are located in pairs near the sockets of the filiform hairs.



Many years ago there were serious debates about whether insects can hear (Pumphrey 1940). When humans hear, we sense the variations in sound pressure at our eardrums; these typically result from sound sources that are very far away. In particular, if we are listening to someone singing an A above middle C, the wavelength of the sound is three-quarters of a meter, so that the distance from our ears to the sound source is usually much greater than the wavelength of the sound. For many insects the situation is very different. A cricket listening to a potentially dangerous wasp beating its wings at 150 Hz is inside one wavelength as soon as the wasp is 2 meters away. This is the near field, where one cannot only "hear" the pressure variations but also "feel" the movements of the air. We can easily experience this phenomenon by cranking up the bass in a stereo system with a large woofer. The cercal sensory hairs serve to measure the air displacements, although the dynamics of coupling to the hair can be quite complex (Humphrey et al. 1993), as illustrated by measurements on analogous structures in spiders (Barth et al. 1993).

The crucial point about a near field acoustic sensor is that it provides the animal with a spatially filtered image of the world, but an image in which directional information is available from a single sensor. Each filiform hair is therefore confined to move in a plane, so that displacements of the hair signal air motions in a particular direction. The cercus is covered with hairs of different orientations, and their afferents project to a set of interneurons that construct a directional map (Miller, Jacobs, and Theunissen 1991; Jacobs and Nevin 1991).

Under natural conditions, the complex patterns of air movement that result from, for example, the wing beats of predators are transduced into hair displacements, and these signals are then encoded into spike trains. The cricket presumably uses these spike trains to construct at least a crude image of its surroundings, and to initiate behaviors appropriate to that image. We are interested here in a relatively simple question: How much information is contained in the spike train of a single sensory neuron? To address this issue, it makes sense to bypass the complex mechanics of coupling between air motion and

hair motion and just grab hold of the sensory hair while recording from its axon. The question is then very direct: By observing the spike train, how much information have we gained about the trajectory of the hair?

The basic experiment, schematized in Fig. 3.18a, is to deliver random displacements $s(t)$ of the sensory hair, and record (with an intracellular electrode) the resulting spike arrival times (Warland 1991; Warland et al. 1992). A long stretch of such recordings is then used to optimize the decoding filter $K_1(t)$, as in Eq. (2.25), and the decoding algorithm is tested on some new stretch of data that did not enter the optimization procedure. This insures that the filter $K_1$ generalizes to describe coding of the stimulus ensemble and not the particular stretch of the stimulus waveform which went into the filter calculation. In these experiments the stimulus ensemble was Gaussian noise with a flat spectrum from 25 to 525 Hz. An example of the stimulus, spike train, and reconstruction is shown in Fig. 3.18b,c. The reconstructed waveform clearly interpolates between the spikes and, in some places, gives a close match to details of the stimulus on very short time scales. This tells us that the bandwidth of the system is large. On the other hand, the typical errors are comparable to the stimulus itself, so the overall signal to noise ratio is about unity.

The quantitative analysis of the reconstructions proceeds, as described above, by separating the estimate of the stimulus into a signal that is deterministically related to the stimulus (but could be systematically biased) and an effective noise that is uncorrelated with the stimulus. The distribution of effective noise amplitudes seems very well approximated by a Gaussian, as shown in Fig. 3.19a. In relating the information transmission rate to the power spectrum of the effective noise, we made use of the maximum entropy property for Gaussian distributions to be sure that we are not overestimating the true information rate; the closer the real noise distribution is to being Gaussian, the tighter this bound becomes. The effective noise spectrum is shown together with the signal spectrum in Fig. 3.19b; we see that the signal to noise ratio $SNR \sim 1$ over a bandwidth of $\sim 300$ Hz. Finally, doing the integral, the information rate in this experiment is $294 \pm 6$ bits/s which corresponds to $3.2 \pm 0.07$ bits/spike (Warland et al. 1992).

The result that a single neuron could transmit nearly 300 bits per second came as quite a surprise. MacKay and McCulloch, however, had realized nearly forty years earlier that information rates of several bits per spike were possible, at least in principle. Still, one must careful in trying to visualize these results. Since 300 bits per second are being transmitted, one could argue that, over a time window of one second, the neuron is providing a spike train which uniquely identifies one signal out of $2^{300} \sim 10^{90}$ possible signals.



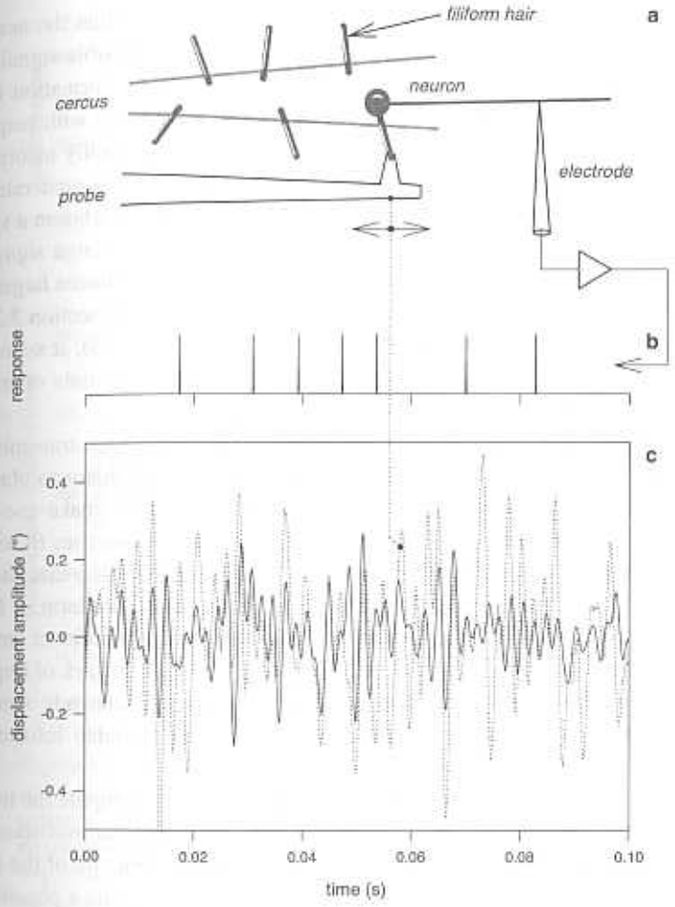



**Figure 3.18**
(a) Schematic of experiment on a cricket mechanoreceptor. A probe is placed over a filiform hair cell protruding from one of the two cerci. The probe is then moved horizontally while spike occurrences are measured with an intracellular electrode inserted into the neuron innervated by the hair (b). Sections of the stimulus (angular displacement of the hair, dotted line) and estimate (solid line) are shown in (c). Adapted from Warland et al. (1992) and Warland (1991).

This is true. A more useful point of view, however, is that the neuron is capable of telling us (roughly) about only one of two possible signals (positive or negative deflection of the sensory hair), but this information is updated every few milliseconds. These are equivalent statements with respect to the information transmission rate, but the second view explicitly incorporates the fact that the high information rate is achieved through a moderate signal to noise ratio across a broad bandwidth. Nature might have chosen a system that achieves the same information transmission rate with a large signal to noise ratio across a narrow bandwidth, but such a system introduces larger delays in the transmission of transient signals. As in the analysis (section 3.2.2) of somatosensory afferents by Werner and Mountcastle (1965), it seems that the cercal afferents are specialized for conveying large amounts of information about rapidly varying signals.

In building up our strategy for bounding the information transmission rates we emphasized that one could in principle use *any* estimator to place a lower bound on the bit rate. The results shown in Fig. 3.19 make use of a very simple estimator, namely the optimal linear filter. If one tries to improve the estimate by adding nonlinear terms, it doesn't seem to decrease the effective noise level. More precisely, after adding the second order term of Eq. (2.25), the bound on $R_{\rm info}$ is not increased by a statistically significant amount. The combination of high bit rates with the linear filter and the lack of improvement with the nonlinear filter certainly suggests that the relatively simple linear reconstruction strategy is capturing much of the available information. We shall see that this can be made more precise.

Let us recall that MacKay and McCulloch did not compute the information transmission in any model coding scheme, but rather the maximum possible information transmission allowed by the statistical structure of the spike train itself. As we have emphasized, the spike train entropy sets a physical limit to information transmission in much the same way that diffraction sets a limit to the spatial resolution of an imaging system. The entropy measures (roughly) the number of distinguishable spike sequences, and the information rate measures the number of distinguishable stimulus waveforms; clearly one cannot distinguish more waveforms than spike trains in any coding scheme. Thus we arrive at a measure of coding efficiency,

$$\epsilon = R_{\rm info}/(S/T), \tag{3.96}$$

where $S/T$ is the entropy per unit time of the spike train. Ideally, every variation in the spike train would correspond to a unique change in the input signal, and we would have efficiency $\epsilon = 1$. Notice that our experimental approach



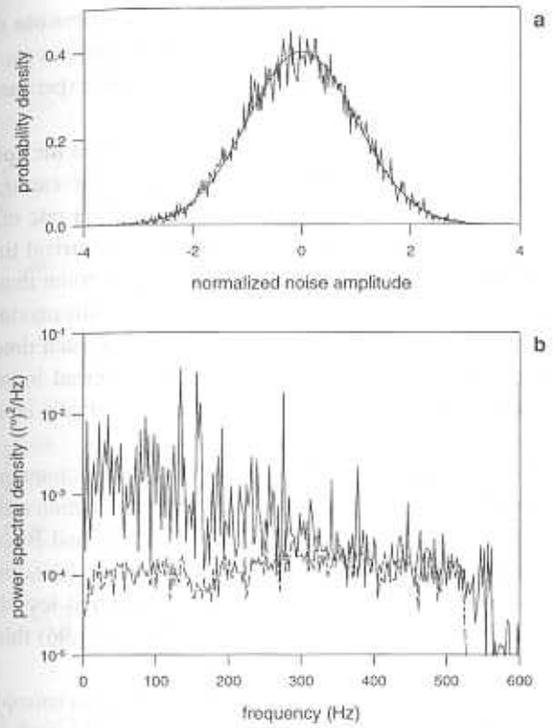



**Figure 3.19**
(a) Distribution of effective noise amplitudes from experiment on a cricket mechanoreceptor. The effective noise amplitude, $\tilde{\eta}(\omega)$, was measured in each of 120 sections of the experiment. Each effective noise amplitude was normalized by the standard deviation of the noise $\sqrt{\langle|\tilde{\eta}(\omega)|^2\rangle/2}$ at that frequency. A histogram is then constructed from these normalized noise amplitudes. The noise histogram constructed in this way is well fit by a Gaussian with a standard deviation of 1 (smooth curve). (b) Signal (solid line) and effective noise (dashed line) power spectra for cricket mechanoreceptor experiment. Although the signal to noise ratio is never particularly high, a signal to noise ratio close to one is maintained across a bandwidth of nearly 300 Hz.

gives us a lower bound on the information rate $R_{\text{info}}$. If we use the same experiments to generate an *upper* bound on the spike train entropy, then we can conclude that the coding efficiency is greater than some experimentally determined quantity.

The key to providing an experimental upper bound on the spike train entropy is again the maximum entropy idea. Although we can't measure the entropy, we can measure, for example, the average firing rate of the neuron. Then we can ask for a probability distribution of spike arrival times that has the largest possible entropy per unit time given the constraint that the average firing rate has to agree with experiment. The answer to this maximum entropy problem is exactly the model of independent spiking in each time bin, which was considered by MacKay and McCulloch and reviewed in section 2.1.3. Thus we know that the MacKay–McCulloch result is actually an *upper bound* on the entropy of a spike train given the mean firing rate.

More accurate bounds can be obtained by assuming that interspike intervals are independent but consistent with the measured distribution rather than just with the first moment $1/r$, and so on (Rieke, Warland, and Bialek 1993). In practice, one can seldom collect enough data to go beyond the double-interval distribution. The important point is that the real entropy is less than what we calculate from these low-order approximations. In Eq. (3.96) this means that we always *underestimate* the efficiency of coding.

The MacKay–McCulloch upper bound on the spike train entropy is the most fundamental, since it makes use only of the mean spike rate and no other information about the spike statistics. Thus *no* coding scheme could use the same number of spikes to transmit more information. Judging the performance of sensory neurons against this standard is the most stringent test of the idea that the neural code is an efficient code in the sense of information theory.

It is obvious that the spike train entropy depends on the assumed timing precision of the nervous system, $\Delta\tau$. In the absence of quantitative estimates for $\Delta\tau$ in this particular sensory system, we take an empirical approach. A given experimental setup has an implicit $\Delta\tau$ related to the properties of the electrode, the spike shape, and the noise level of the recording electronics. This finite precision is made explicit when we digitize the spike arrival times, defining bins within which all arrival times are viewed as equivalent. In the cricket experiments the bin size was 0.1 ms (Warland 1991). Thus, when we say that the reconstructions carry 3 bits per spike of information about the stimulus, this applies to conditions when $\Delta\tau = 0.1$ ms. When we increase $\Delta\tau$ we change both the information rate and the entropy rate, as shown in Fig. 3.20. To compute the information rate we redigitize the spike arrival times at lower resolution and carry through the same reconstruction analysis—find



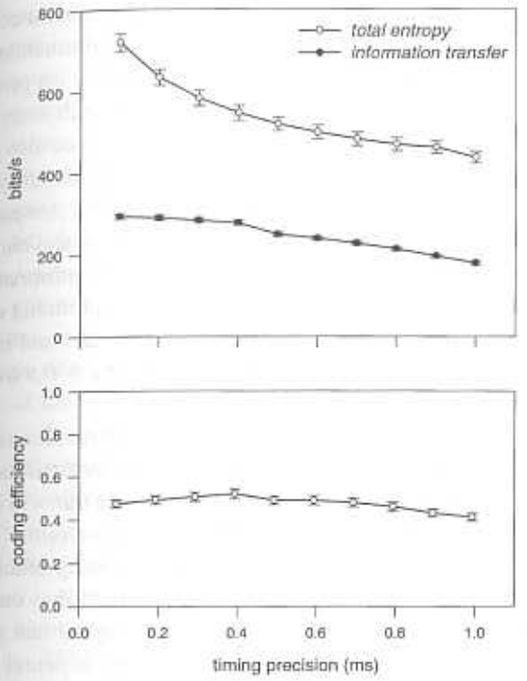



**Figure 3.20**
Ingredients of coding efficiency. Top panel shows the entropy rate (upper curve) and information rate (lower curve) as a function of timing precision for a cricket mechanoreceptor. The information rate at each timing precision $\Delta\tau$ is calculated by placing each measured spike at random within a bin of width $\Delta\tau$ centered on the original spike time. The estimation procedure is then repeated with this jittered spike train, and the quality of the reconstructions is again quantified by the effective noise level, leading to the information rate through Eq. (3.95). The entropy rate sets a fundamental upper limit to the information rate. The entropy rate is calculated from Eq. (3.22) using the mean firing rate of 96 $s^{-1}$, measured from the same section of data that went into the information rate calculation. Error bars are calculated by dividing the experiment into 4 sections and computing the standard deviation of each quantity among sections. The error bars on the information rate are obscured by the data points. Bottom panel shows the coding efficiency, defined in Eq. (3.96), for the cricket mechanoreceptor, calculated from the information and entropy rates above. The coding efficiency sets a lower bound on the fraction of the degrees of freedom in the spike train code that is used to represent the sensory input.

the optimal kernels, separate errors into systematic and random components, measure the noise power spectrum, and compute the information rate. The information rate is not terribly fragile—bins as large as 0.3 ms preserve 90% of the original information rate. The loss of information with decreasing time resolution is gradual and not catastrophic. By changing $\Delta\tau$ we also change the entropy rate, and from Eq. (3.22) we expect the effect to be fairly dramatic. Specifically, at small $\Delta\tau$ every factor of two increase in bin size causes a loss in entropy of one bit per spike. Comparing the information and entropy rates, we see that at very small $\Delta\tau$ the entropy is large but the information rate is bounded—we are not gaining anything by keeping track of timing on this fine scale. But the two curves approach each other at larger $\Delta\tau$, and in Fig. 3.20 we see that their ratio, the coding efficiency $\epsilon$, hovers at $\epsilon \sim 0.5$ over a broad range of $\Delta\tau$.

The coding efficiency results demonstrate that the information rates of 300 bits/s are within a factor of two of the physical limits set by the statistics of the spike train itself. No coding scheme could use these spike trains to carry more than twice as much information about the input signal waveform.

The results on information transmission rates and coding efficiency have three implications. First, the strategy followed here establishes only a lower bound on the information transmission rate, and we argued that the quality of this bound is directly related to our understanding of the neural code. Evidently the bound is very good, since it is close to the physical limit. Thus the linear decoding scheme not only "works" in the perturbative sense that small nonlinear terms don't seem to help, it works in the global sense that no decoding scheme could be much better. This is a very strong conclusion—we understand the neural code well enough to extract at least half of all the information that could possibly be present in spike trains sampled at reasonable time resolution. In experiments on the frog auditory system we will see information rates that come even closer to the physical limits, strengthening our claim of understanding the code.

The second conclusion concerns the contrast between rate and timing codes. We have argued that it is difficult to give a precise formulation of this distinction, and in particular that this distinction cannot be discussed without attention to the dynamics of the input signals. On the other hand, rate coding makes the clear qualitative prediction that the precise arrival times of the spikes do not carry information—although we can record spike times with submillisecond precision, this precision does not give us more knowledge about the identity of sensory signals. For a neuron that makes use of a timing code we expect, on the contrary, that more precise measurements of spike arrivals will yield



more detailed knowledge of the sensory world. The difficulty, then, is in attaching numbers to the ideas of "precision" and "knowledge," but these are exactly the problems addressed by information theory: The spike train entropy quantifies the effort we make to record spike times more accurately, and the information transmission rate quantifies what we gain in exchange for this increased precision; both quantities are functions of the time resolution itself, and are defined only in the context of the stimulus ensemble. We suggest that the distinction between rate and timing codes can be quantified by the dependence of coding efficiency on time resolution: If the intuitive rate coding picture is valid, then coding efficiencies must be very low when we sample the spike train at small $\Delta\tau$. Conversely, the intuitive notion of a timing code predicts that the efficiency should remain high even for $\Delta\tau$ a small fraction of the typical interspike intervals. The data of Fig. 3.20 show that, at least in this one stimulus ensemble, coding efficiency in the cricket cercal afferent remains near 50% at time resolutions just a few percent of the typical interspike intervals. The rather constant efficiency means that the effort of marking spike arrival times with increasing time precision is rewarded with proportional increases in information transmission, down to 0.4 ms resolution. In this quantitative sense, the submillisecond timing of action potentials provides knowledge about the sensory inputs.

The third and more speculative conclusion concerns the function of the primary sensory neuron. At the outset, one might imagine that the physical capacity of neurons to carry information is not relevant to the function of real organisms. On the contrary, the issue might be to parcel out the "interesting" pieces of the world as soon as possible, using a presumably large information capacity to convey only these important features with high reliability. At least under the conditions of this one experiment, it is clear that the limits to neural information capacity *are* relevant, and that the coding strategy adopted by primary sensory neurons involves the transmission of large amounts of information at relatively modest signal to noise ratio. If we take these conclusions seriously, it should be possible to state some general principles that govern the "design" of the neural code. But first let us look at a completely different system to see which (if any) of these conclusions can be generalized.

### 3.3.2 Amphibian eyes and ears

The ideas and methods of the previous sections have also been applied to two very different vertebrate sensory systems, the frog sacculus (Rieke et al. 1992; Rieke, Warland, and Bialek 1993) and the salamander retina (Warland and Meister 1993, 1995). In each case, the amount of information available from the afferent spike trains in these systems also is a substantial fraction of the

spike train entropy. We will return to the implications of these results, but first the data.

Frogs and salamanders are examples from two of the three orders of amphibians. Frogs are members of the order Anura, and salamanders are members of the order Urodela.[11] If one measures the success of an order by the diversity of the member families , then the order Anura is by far the most successful of the amphibians with 3,400 species (Frost 1985). In contrast, the order Urodela has about one tenth the number of species. Both orders are exquisitely adapted to their particular environmental niche. Part of this specialization is reflected in these creatures' sensory systems.

Vibrations of the ground are one source of information that frogs use to determine the location of predators. Precise sensing of these vibrations is done by a specialized sensory organ called the sacculus. The sacculus utilizes hair cells similar to those in the cochlea, and the primary afferent neurons innervating the sacculus join auditory afferents in the eighth nerve. The coding in these afferents can be studied by recording from a single afferent fiber while shaking the entire frog. Fortunately, the system is exquisitely sensitive and the shaking does not dislodge the electrode! The strategy in such experiments is similar to those already discussed: A random stimulus approximating Gaussian white noise is delivered to the system while the spike times are monitored. The stimulus waveform and spike times are then used to calculate the estimation kernels $\{K_n\}$. In the experiment discussed here, the inclusion of nonlinear terms in the estimation procedure improved the information rate by 7%; since this contribution was small we will restrict our discussion to linear estimation.

The power spectra of the stimulus and effective noise for an experiment on an afferent from the sacculus are shown in Fig. 3.21. The signal to noise ratio of the estimates is 3 to 4 for frequencies between 40 and 70 Hz, reflecting the tuning of the sacculus to low frequency vibrations. Outside this frequency range the quality of the estimates rapidly degrades. Thus, as in the cricket mechanoreceptors, these vibration sensors code at moderate SNR over a fairly wide bandwidth. Integrating over frequencies yields an information transmission rate of $155 \pm 3$ bits/sec, or nearly 3 bits/spike. Notice that the range of frequencies encoded by the sacculus is very different from that in the cercus, and that the total information rates and spike rates are different, but the number of bits per spike is very similar. The coding efficiency is plotted in Fig. 3.22. As in the cricket, the efficiency reaches 0.5–0.6 over a range of timing resolutions.

11. The other, relatively uncommon, amphibian order is appropriately named Apoda because the creatures do not have feet.



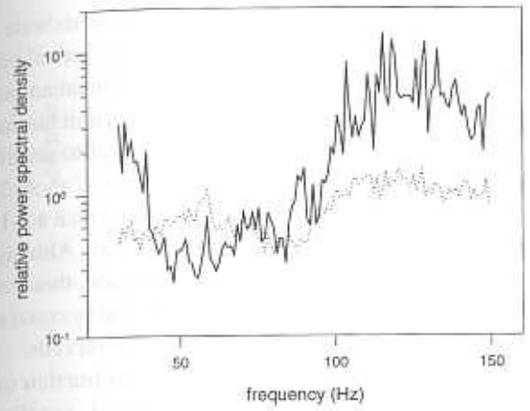



**Figure 3.21**
Power spectrum of stimulus (dashed line) and effective noise (solid line) for sacculus experiment. The signal in these experiments was Gaussian noise with an approximately flat power spectrum between 30 and 1000 Hz. The dip in the effective noise spectrum between 40 and 80 Hz reflects the tuning of the cell to vibrations in this frequency range. The signal to noise ratio of the estimate reaches a peak of about 3, and a signal to noise ratio greater than 1 is maintained from about 40 to 70 Hz.

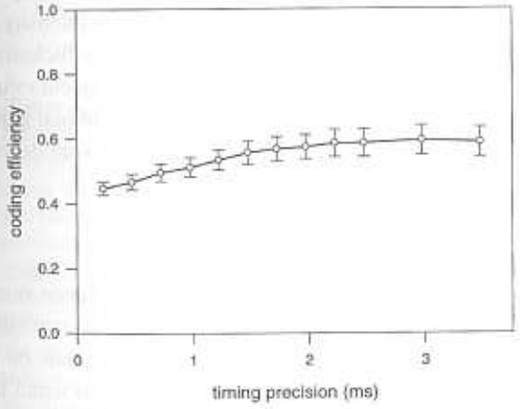



**Figure 3.22**
Coding efficiency in frog sacculus, computed as described in Fig. 3.20 for the experiment described in Fig. 3.21.

The salamander eye is widely used as a model for the vertebrate visual system. In part, this choice is motivated by large and robust retinal cells that are more experimentally accessible than those of their mammalian counterparts. While not as visually acute as frogs, salamanders still exhibit fascinating visually guided behavior. Most notably, salamanders respond to drifting gratings with a stripe separation of 0.1 degrees. This corresponds to about 2 $\mu$m when perfectly focused onto the retina and is somewhat smaller than a 5-10 $\mu$m photoreceptor diameter (Himstedt and Grüsser-Cornehls 1976). Although it can be difficult to convince yourself of this by casual observation, these experiments indicate that salamanders, like other animals, can respond to visual stimuli that vary on the scale of individual photoreceptors and ganglion cells.

The salamander retina has been the subject of reconstruction experiments using a preparation in which an isolated retina is placed, ganglion cell side down, on an array of extracellular electrodes (Meister, Pine, and Baylor 1994). Using this array it is possible to record action potentials from about 40 ganglion cells. This opportunity for simultaneous recording from several neurons allows an attack on many questions about the way in which information about spatially varying signals is shared among the spike trains of different cells. We defer these questions to section 5.1, and focus on a much simpler problem, the encoding of a spatially uniform stimulus that varies in time—full field flicker—so that the stimulus $s(t)$ is just the light intensity. In this particular experiment (Warland and Meister 1993), the light intensity was chosen at random from a Gaussian distribution every 15 ms. The flickering light was transduced by the photoreceptors and processed by subsequent retinal neurons, leading eventually to patterns of action potentials in the retinal ganglion cells.

The spike trains from this ensemble of ganglion cells were decoded using a generalized version of the decoding algorithm,

$$s_{est}(t) = \sum_n \sum_i K_1^n(t - t_i^n), \tag{3.97}$$

where $n$ denotes cell number and $t_i^n$ denotes the occurrence time of the $i^{th}$ spike in the $n^{th}$ cell. Notice that each cell has its own "private" filter $K_1^n$, but the choice of filters that give the best reconstructions depends on the correlations among the different cells. The filters were varied as usual to minimize the mean square error between the stimulus and the estimate, but now one has to find many filters, one for each cell. The top panel of Fig. 3.23 compares the resulting optimized reconstruction with the true stimulus as a function of time. The reconstruction was made with spike trains from 4 cells. About 18 spikes occurred in this 2 second window. Notice that in places where the true stimu-



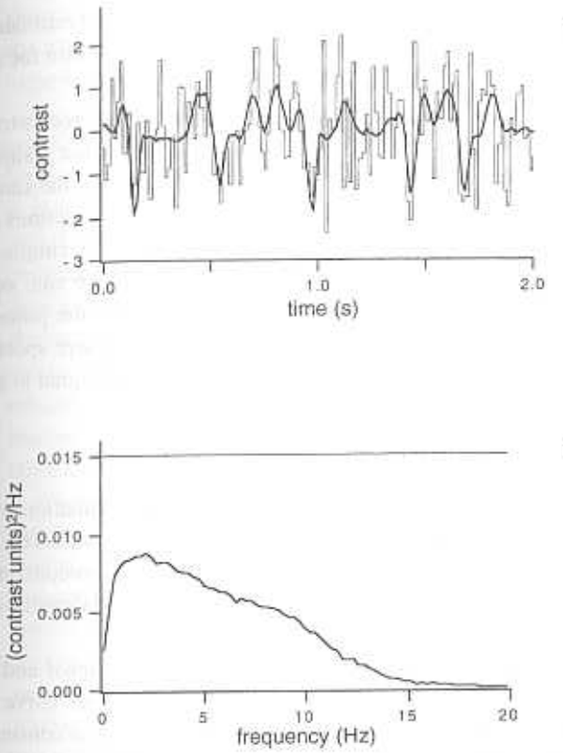



**Figure 3.23**
Multi-cell reconstructions from salamander retinal ganglion cells. Responses from multiple retinal ganglion cells were recorded simultaneously while presenting the retina with a dim, full field stimulus. (a) The stimulus intensity (thin line) was chosen randomly from a Gaussian distribution every 15 ms. Stimulus contrast is shown in units of its standard deviation. The time course of the resulting stimulus was estimated from the spike trains of 4 ganglion cells. Each spike train was filtered through a linear estimation filter calculated as described in the text. The outputs of the 4 filters were summed to obtain the stimulus estimate shown (thick line). (b) Power spectra of the signal (thin line) and estimate (thick line). At low frequencies the estimate captures much of the structure in the stimulus, and the power in the estimate approaches the stimulus power. As the frequency increases, filtering in the retinal cells upstream of the ganglion cells causes the power in the estimate to fall. From experiments by D. Warland and M. Meister.

lus is changing rapidly, the reconstruction does poorly and estimates the mean intensity, but in places where the light has been bright or dim for a while, the reconstruction does much better.

We have explained in section 3.2.3 how the quality of reconstructions can be quantified by defining an effective noise level, and hence a signal to noise ratio ($SNR$), at each frequency. This approach separates the random errors or noise from the systematic errors, but for linear reconstructions it turns out the magnitude of the systematic errors is related in a very simple way to the $SNR$. Instead of measuring the effective noise level, we can, equivalently, quantify the systematic errors: In linear reconstructions, the power spectrum of the estimated signal, $S_{\rm est}(\omega)$, will be less than the power spectrum of the real signal, $S_{\rm real}(\omega)$, by a factor that depends only on the signal to noise ratio,

$$\frac{S_{\rm est}(\omega)}{S_{\rm real}(\omega)} = \frac{SNR(\omega)}{1+SNR(\omega)}. \tag{3.98}$$

Note that as the signal to noise ratio becomes large, estimation becomes essentially perfect and the power in the reconstruction approaches the true signal power. At a signal to noise ratio of one, the reconstruction contains only half the power in the stimulus, and as the noise level becomes larger the reconstruction captures less and less of the total power.

In Fig. 3.23b we compare the power spectra of the signal and the reconstructed signal for the salamander experiment. The upper curve shows the power in the stimulus as a function of frequency, which is constant over the range 0–60 Hz. The lower curve shows the power spectrum of the reconstruction. This curve peaks at a frequency near 2 Hz and has power over a much smaller range of frequencies. At high frequencies, the power falls off and is nearly zero by 15 Hz. This decline at high frequencies parallels the frequency response of the photoreceptors.

From Eq. (3.98) we can convert the spectra of Fig. 3.23b into a measurement of signal to noise ratio, and then estimate the rate of information transmission as in the discussion of the frog and cricket experiments. The result is that this small group of cells in the salamander retina conveys about 9.6 bits/sec of information about the time course of full field flicker, corresponding to a coding efficiency of $\epsilon \approx 0.2$ if spike arrival times are measured with a precision of $\Delta\tau = 15$ ms.

Thus, in three quite different systems we have seen coding efficiencies in the 20–60% range, and we begin to be confident that coding at high efficiency is a general property of peripheral sensory processing. It is clear from these results that linear reconstructions capture about half of the information these spike trains could possibly carry about the sensory world. What happens with the other half? One answer emerges in the next section from the analysis of experiments with more naturalistic stimuli.



### 3.3.3 Frogs and frog calls

The world around us is, thankfully, a highly structured place. This structure is reflected in the fact that the signals that reach our sense organs are not completely random, but rather exhibit correlations in space and in time. What does the nervous system do with this structure? One possibility, first raised by Barlow in 1959, is that even the first stages of neural signal processing exploit the statistical structure of signals to create more efficient representations of the sensory world (Barlow 1961). In this section we explore the use of the reconstruction method to test this idea directly, comparing the information transmission and coding efficiency for single auditory afferents responding to stimuli chosen from different ensembles (Rieke, Bodnar, and Bialek 1995).

We discuss in section 5.2 the more general problem of characterizing natural signals in different sensory modalities. Here we focus on the bullfrog auditory system, which spends much of its time processing rather stereotyped sounds—frog calls—one of which is illustrated in Fig. 3.24. Frogs and toads use species specific communication signals, called advertisement calls, in their social and reproductive behavior. The power spectrum of frog calls consists of approximately 20 nearly harmonic bands, with a fundamental frequency near 100 Hz (Capranica 1965, 1968). This power spectrum endows the ensemble of call stimuli with a finite *correlation time*, which measures how far into the future the waveform can be predicted given knowledge of the past (see section 3.1.4). Roughly speaking, the correlation time is the inverse of the width of each spectral band, $\sim 30$ ms for these signals.

Though animals use the temporal correlations of natural stimuli in making behavioral decisions, it is not known at what stage in processing these correlations become important. We can imagine a family of different stimulus ensembles, starting with the "most random" Gaussian white noise and progressing toward the "most structured" sounds that actually occur at the frog pond. Ideally, we would like to understand how each of the naturalistic structures that can be built into the stimulus ensemble influences processing and coding.

The first step along this path is to take the Gaussian white noise and filter it, shaping its power spectrum into bands that match the spectrum of the naturally occurring calls. As a second step, one could replace these bands of Gaussian

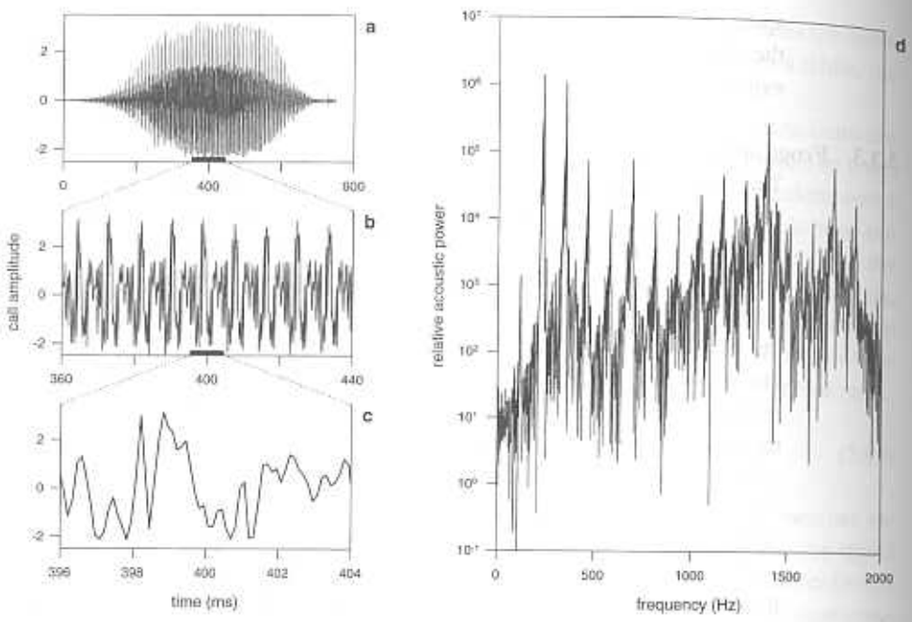



**Figure 3.24**
Bullfrog advertisement call. Single croak of bullfrog advertisement call recorded at Lake Carnegie in Princeton, New Jersey. Panels (a), (b) and (c) show the sound pressure waveform of the call on three different time scales. The periodic structure of the call, particularly clear in (b), indicates nonrandom phase relations between different frequency components. Panel (d) shows the power spectrum of the call. The call has a harmonic structure with a fundamental frequency close to 100 Hz. Thus the advertisement call has highly structured phase and amplitude spectra.

noise with tones that are randomly frequency or phase modulated to give the same power spectrum. Finally, one could try to give these modulations the same structure that occurs in real calls. Again we emphasize that to follow this program to completion one needs to understand much more about the statistics of natural sounds. In particular, to analyze information transmission using reconstruction methods we have to know the entropy of the distribution from which the signals are drawn. Here we review experiments (Rieke, Bodnar, and Bialek 1995) that take the first step, comparing the encoding of white noise stimuli with the encoding of stimuli that have naturalistic power spectra. Perhaps surprisingly, this small step toward natural sound has a big effect on the efficiency of the neural code.



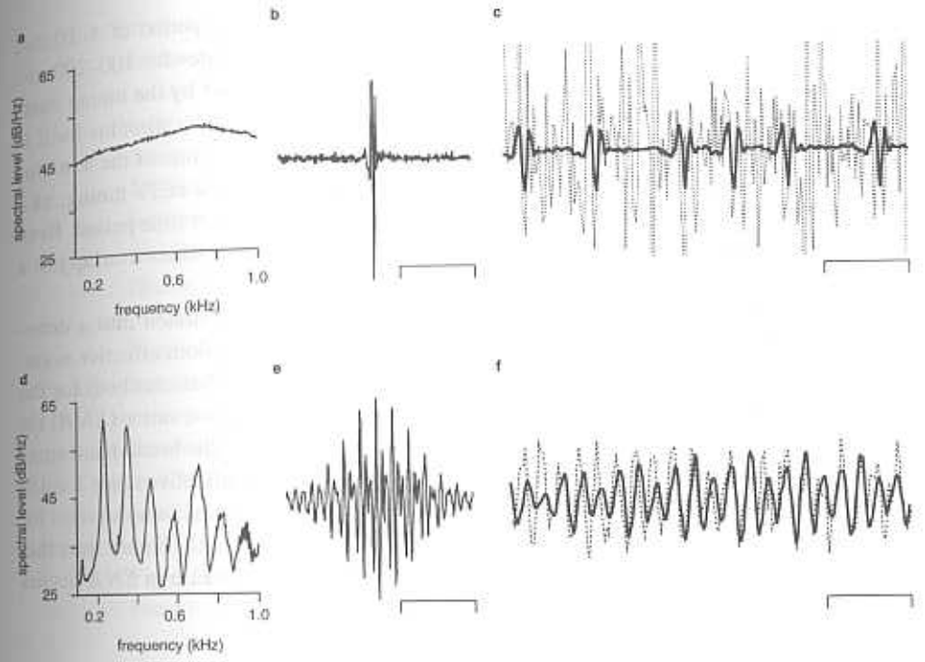



**Figure 3.25**
Stimuli, filters and estimates for experiments comparing coding of broad-band Gaussian noise and Gaussian noise shaped to have a natural power spectrum. Power spectra of the experimental stimuli are shown in (a) for the broad-band stimulus and in (d) for the call-spectrum stimulus. Reconstruction filters are shown in (b) and (e). Sections of the stimulus, spike train and estimate are shown in (c) and (f). Timing bars are 20 ms in (b) and (e) and 10 ms in (c) and (f). Redrawn from Rieke, Bodnar, and Bialek (1995).

Figure 3.25 shows results from an experiment on a single auditory nerve fiber that originates in the amphibian papilla, a frog auditory organ that is tuned to frequencies below 600 Hz. Neurons from this organ show a variety of response features that are broadly typical of the vertebrate auditory nerve—phase locking to low frequency sounds, two-tone suppression, and difference tone nonlinearities—as described in the review by Lewis, Leverenz, and Bialek (1985). We see immediately, even without quantitative analysis, that the attempt to reconstruct the stimulus is much more successful in the case of the stimulus with the shaped spectrum. Another obvious difference between the two experiments is the temporal width of the filters. A spike in the broad-

band noise experiment contributes to the estimate for a period of 5–10 ms, whereas a spike in the call-spectrum experiment contributes for 100–200 ms. The width of the filter in the broad-band experiment is set by the tuning characteristics of the neuron, because the correlation time of the stimulus itself is so short. In the call-spectrum experiment, the correlation time of the stimulus itself is longer than the correlation time introduced by the cell's tuning; as a result, a single spike contributes to the estimate for a longer time period. Even for the call-spectrum stimuli, however, the reconstruction filters overlap just a few interspike intervals.

As in previous analyses the reconstruction can be separated into a deterministic component related to the true stimulus and a random effective noise. The distribution of effective noise amplitudes is nearly Gaussian both for the broad-band and for the call-spectrum stimuli. Signal to noise ratios ($SNR$) for the experiment of Fig. 3.25 are shown in Fig. 3.26. With the broad-band stimuli, the frequency tuning of the cell measured from the effective noise level is similar to that measured using standard methods. The signal to noise ratio for reconstructions of the call-spectrum stimulus is significantly higher than that for the broad-band stimulus at most frequencies; this increase in $SNR$ occurs

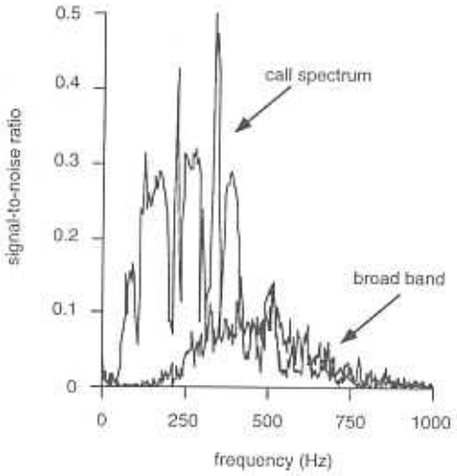



**Figure 3.26**
Signal to noise ratio for broad-band and call-spectrum stimuli. Integrating the signal to noise yields information rates of 133 bits/sec for the call-spectrum stimulus and 46 bits/sec for the broad-band stimulus. Redrawn from Rieke, Bodnar and Bialek (1995).



even at frequencies where the power in the call-spectrum stimulus is 10–15 dB below that of the broad-band noise.

As a final step in the analysis of the reconstructions, we use the signal to noise ratio at each frequency to compute the lower bound on the information transmission rate, as in Eq. (3.95). The result is $R_{\rm info} = 46 \pm 1$ bits/sec, or 1.4 bits/spike, for the broad-band stimulus, which increases to $R_{\rm info} = 133 \pm 5$ bits/sec for the call-spectrum stimulus. This last result corresponds to the cell transmitting 7.8 bits per spike. Similar results were obtained in many cells (Rieke, Bodnar, and Bialek 1995).

The dramatic improvement in information rate in response to a seemingly small step toward naturalistic stimuli leads us to ask how much additional increase in information rate is possible. It is physically impossible for a neuron to transmit sensory information at a rate greater than the entropy per unit time of the spike train. This entropy is in turn smaller than that calculated by MacKay and McCulloch (1952) and reviewed in section 3.1.2. The coding efficiency, or ratio of information rate to entropy rate which we defined in Eq. (3.96), is shown in Fig. 3.27 as a function of time resolution, by analogy with Fig. 3.20 in the cricket experiments. We see that, for the call-spectrum stimuli but *not* for the white noise stimuli, the efficiency reaches 90%, indicating that

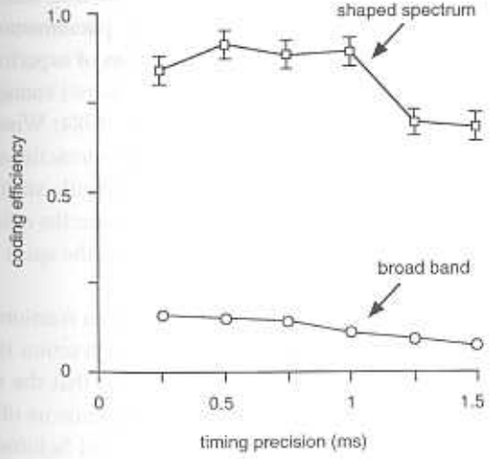



**Figure 3.27**
Coding efficiency for broad-band and call-spectrum stimuli. Error bars on the broad-band measurements are obscured by the data points. Redrawn from Rieke, Bodnar, and Bialek (1995).

the coding of the call-spectrum stimulus comes very close to the fundamental limits on information transfer.

These results indicate that the dynamics of the coding process in primary auditory neurons exploits the correlation structure of natural sounds to encode signals at higher information rates and efficiencies. This improvement would not be possible if the auditory system acted linearly with additive noise; if this were the case, lowering the power at a given frequency would lower the signal to noise ratio at that frequency, and broad-band stimuli would provide the highest possible rate of information transfer. Instead, nonlinearities in auditory processing increase the information rate and coding efficiency for naturalistic stimuli.

The idea that nonlinearities play an important role in the coding of complex sounds has come from many different experiments. In the frog, Schwartz and Simmons (1990) found that auditory neurons tuned to high frequencies phase lock to the fundamental frequency of the frog call, although the power of the second and third harmonics is considerably greater than the power at the fundamental frequency. Schwartz and Simmons suggest that the neurons derive the fundamental frequency from nonlinear interactions among the harmonics within their tuning curve, and this mechanism of "periodicity extraction" has been studied by Simmons and Ferragamo (1993) and by Simmons, Reese, and Ferragamo (1993). These nonlinear responses of the frog auditory system are strongly reminiscent of the "missing fundamental" phenomenon in human pitch perception, reviewed by de Boer (1976). In a series of experiments on the mammalian auditory nerve, Young, Sachs, and coworkers (Young and Sachs 1979; Sachs and Young 1980; Miller and Sachs 1983, 1984; Winslow, Barta, and Sachs 1987) have emphasized the role of nonlinear interactions in preserving information about spectral structure in complex sounds such as speech. The notion that these nonlinear interactions should enhance the efficiency with which the ear encodes more naturalistic sounds captures the spirit of the classical ethological studies (Capranica 1965, 1968).

Coding efficiency is a measure of reproducibility—what fraction of the neural response is uniquely related to the signal, and what fraction is noise? Increased coding efficiency for naturalistic signals means that the response to those signals is more reliable or reproducible. This reminds us of the results of Miller and Mark (1992) as well as those of Mainen and Sejnowski (1995), which showed how—in very different systems—signals with more natural time dependencies produce spike trains that are more reproducible from trial to trial, and Strong et al. (1996) show how this reproducibility can be quantified and used to provide an independent measure of information transmission. All



of these results emphasize that the full performance of even peripheral sensory neurons may be revealed only in response to the most natural of signals.

## 3.4 SUMMARY

We have seen that Shannon's information theory attaches a number to the "amount we can learn about the world" by observing certain signals. In the case of spike trains, by observing each spike to within a given temporal precision it is possible to gather *at most* an amount of information given by Eq. (3.22), whereas if we restrict ourselves to counting the number of spikes in large bins the amount of information is limited by Eq. (3.24). These limits to information transmission clarify the distinction between rate and timing codes, which we have argued is more subtle than one might have imagined. Similar limits apply to information transmission by vesicles at a chemical synapse, even if the presynaptic and postsynaptic neurons both give graded responses.

Experiments in several systems demonstrate that real neurons and synapses approach the limits to information transmission set by the spike train or vesicle entropy. Rather than throwing away information in favor of specific "biologically relevant" signals, these cells seem to pack as much information as possible into the spike sequences they send to the brain. The notion of biological relevance reappears as a matching of the coding strategy to sensory ecology—specifically, to the temporal features of naturally occurring signals—so that the same number of spikes can be used to transmit more information about the more structured signals that occur in the real world.

# Chapter 4
# Reliability of computation

We all share the qualitative impression that our perceptions are reliable. Illusions notwithstanding, this impression is supported by experience: we can run through the woods at relatively high speeds, avoiding collisions and missed footings—testimony to the reliability with which our senses signal the location of obstacles. There is a long tradition of quantifying the reliability and precision of our perceptions; this is the subject of psychophysics. In this chapter we explore several different experimental and theoretical approaches, all of which aim at understanding the neural basis of reliable perception. We shall see that, for several systems, there is agreement among two of three fundamental quantities: The reliability of behavior, the reliability of single neurons, and the reliability of an optimal processor that makes use of all relevant sensory input down to the physical limits imposed by noise in the sense data itself.

## 4.1 RELIABILITY OF NEURONS AND RELIABILITY OF PERCEPTION

Understanding the reliability of the nervous system is fundamentally a quantitative problem (Bullock 1970). Indeed, this is one of the few areas in the investigation of biological systems where it is clear from the outset that our qualitative view of function and mechanism will necessarily be influenced by quantitative experiments. Quantifying the reliability of spiking neurons is difficult, however, because we must decide whether a particular sequence of spikes is close to the "correct answer" to some computational problem. Clearly, this requires an understanding of the neural code, or else we run the risk of confusing a complex encoding with a wrong or random answer. It is equally important, however, to find a scale on which to measure the deviations between the result of a neural computation and the "correct answer" for that particular problem. One approach to developing such a scale is to compare the

reliability of individual neurons with the reliability of the nervous system as a whole. Alternatively, we can compare neural performance to the fundamental limits on reliability that arise from noise at the input to the computation.

### 4.1.1 Historical background

In the 1940s and 1950s, several investigators realized that understanding the reliability of computation in the nervous system posed significant theoretical challenges. Attempts to perform reliable computations with the available electronic computers certainly posed serious practical problems, and the possibility that the problems of natural and artificial computing are related was explored. Guided by the practical difficulties of electronic computing, von Neumann (1956) formulated the theoretical problem of "reliable computation with unreliable components." Many authors seem to take as self-evident the claim that this is a problem faced by the nervous system as well. The qualitative picture adopted in this approach envisions the nervous system as a highly interconnected network of rather noisy cells, in which meaningful signals are represented only by large numbers of neural firing events averaged over numerous redundant neurons. This has led to a widespread belief that neurons are inherently noisy, and ideas of redundancy and averaging pervade much of the literature. Interestingly, von Neumann himself did not seem to hold this view (von Neumann 1958).

Qualitatively, most sensory neurons *seem* unreliable in the obvious sense that repeated presentations of the same sensory stimulus do not lead to identical spike trains, as in the example of Fig. 2.1. But it is not so clear how we should quantify these observations, nor is it clear on what scale reliability should be measured: How much of the apparent noise in neural responses is the inevitable result of noise in the stimulus itself, how much resides in the mechanisms of spike generation, and how much is added by the many stages of processing as the signals pass through the brain?

There are a few strong voices objecting to the view of the brain as a noisy processor. For example, Barlow has emphasized the idea that neuronal processing of sensory signals is efficient in an information theoretic sense, minimizing redundancy between the signals carried by neighboring neurons (Barlow 1961) and reaching decisions with a reliability limited by the statistical structure of the sensory input (Barlow 1956, 1980). In a similar spirit, Bullock (1970, 1976) has collected a number of examples in which individual neurons provide highly reliable signals, or, more subtly, where apparent unreliability may serve to optimize the overall reliability of the system's function. From a theoretical point of view, these different sets of ideas about neuronal reliability



take us in very different directions: If cells are very noisy, then something like von Neumann's problem really is an issue for the brain. On the other hand, if neuronal operations maximize reliability and efficiency and effectively suppress the ever threatening sources of noise in single cells and synapses, then there should be a theory of this optimization process that predicts some of the essential features of neural function. Thus, as emphasized in section 1.3, our qualitatitive view will be influenced by the outcome of quantitative experiments. Rather than developing these theoretical ideas in isolation, we want to see if experiment can decide between the alternative directions.

Comparing the reliability of perception to the reliability of individual neurons is difficult because information of relevance to a particular behavior may be shared among a large number of cells. This distribution of information does not mean that signals from individual cells are unreliable, nor does it mean that the different neurons are redundant. It just means that we need to be careful in phrasing our questions. For example, information about the trajectory of a large moving object is distributed across many photoreceptors and retinal ganglion cells, and hence, even if the brain is an optimal and noiseless processor, the reliability of individual ganglion cells will be less than that of the organism if we pose the problem of discriminating small changes in trajectory.

There is a case for which these issues can be avoided, and this is the auditory system of the Noctuid moths, which was studied extensively by Roeder and coworkers. We describe it here not because it provides an example of extensive quantitative analysis, but rather because the simplicity of the system allowed Roeder to pose the problem of computational reliability in a stark and straightforward manner. Perhaps our review will stimulate someone to take up this preparation again and give a quantitative solution to Roeder's problem. For reviews of this work see Roeder and Payne (1966) and Roeder's monograph (Roeder 1963).

Noctuid moths have spectacular bat-evading flight strategies, acrobatics that are triggered by auditory inputs. In effect, the moth hears the bat coming and tries to fly away. Because the moth cannot outrun the bat, its only hope is to detect the bat's echolocation pulses when the bat is still sufficiently far away that returning echoes do not yet provide a clear image of the moth. The moth then flies away from the bat and hopes that the bat gets interested in something else. If this fails and the bat closes in, the moth can still try a power dive, which bats apparently prefer not to follow. Whereas the dive appears to be triggered just by the presence of a bat too close for comfort, the earlier "flying away" is a directional response. Clearly, an important step in making this strategy work

is for the moth to tune its ear to the frequency bands which dominate the bat's echolocation calls; this leads to an interesting natural history of coevolution in the sensory systems of predator and prey.

It is worth remembering that work on the *Noctuidae* began at a time when the whole question of whether insects "hear" in any way comparable to our own hearing was hotly debated (Pumphrey 1940). Much of the work was also done at the same time as the key experiments demonstrating that bats echolocate, using some of the early ultrasonic transducers (Roeder and Treat 1957). Not only do moths hear, they do it in an especially simple way. Roeder was able to show that each Noctuid ear has just two primary sensory neurons. One of these cells is activated by relatively quiet ultrasonic pulses, presumably corresponding to bats at a distance, and the second cell begins to fire once the pulses become much louder, presumably corresponding to one bat closing in. In the early experiments, Roeder was worried that his ultrasonic sources might produce artifactual signals outside the bands that bats use for echolocation, so that the moth would be "hearing" a sound in the laboratory that did not correlate with the approach of bats in the field. The solution was to record from the neurons while someone held a bat near the preparation. This experiment worked, and the cells respond to real bat calls.

Under the acoustic conditions where moths give directional responses, flying away from bats rather than just diving, it seems clear that only the more sensitive of the two auditory neurons is activated. Indeed, there are species of *Noctuidae* that have only one cell per ear—which is a remarkable fact all by itself (Surlykke 1984). This especially simple situation means that with careful dissection it is actually possible to record *all* of the auditory input to the moth's brain just by making extracellular recordings from two nerves. Roeder did this, not in the confines of the laboratory but in the field with bats flying around him (Roeder and Treat 1961).

Figure 4.1, taken from Roeder's "field physiology" experiment, is an excellent illustration of the twin problems of reliability and coding in spiking neurons. Let us emphasize that these two trains of spikes are all that the moth has to work with as it attempts to detect and evade oncoming bats. Together these two spike trains must convey enough information for the moth to decide that a bat is close enough to be threatening, and hence that evasive action is worthwhile. In addition, the spikes must provide a directional signal so that the moth turns away from the bat rather than flying into its clutches. The comparison of neural spike trains with behavior raises several questions that are at the heart of our discussion: How reliable are the neural signals? What is the code by which direction is represented? How does the reliability of the moth's




**Figure 4.1**
Roeder's field recordings: Binaural tympanic nerve responses of the moth *Feltia sp.* to the cries of red bats flying in the field. The slow waves on both channels are the electrocardiogram of the moth. Large spikes appear regularly, but without synchrony in both traces. (A): An approaching bat. Differential tympanic responses (latency, number of spikes) between right and left is marked at first, but has practically disappeared in the final response. (B) A "buzz" registered mainly by one ear. (C) A "buzz" registered a few seconds later by both ears. Redrawn from Roeder (1963).

directional response compare with the limits imposed by the reliability of its two afferent neurons?

Despite the apparent simplicity of the moth, most of the literature on the reliability of neurons is concerned with mammalian systems. We will try to redress this balance a bit in subsequent sections, but here we review the pioneering experiments and theoretical work, mostly from the 1960s and early 1970s.

### 4.1.2 Photon counting

Barlow and Levick (1969) set out to measure the reliability of retinal ganglion cells as they signal brief flashes of light superposed on a background (intensity discrimination) or in a completely dark adapted state (detection). One of the crucial motivations for this work was to compare the performance of single neurons with the known ability of human observers to count small numbers of photons, or, alternatively, to compare the performance of neurons with the absolute physical limits imposed by the random arrival of photons at the retina. It thus seems appropriate to review here the beautiful story of photon counting in the visual system.

In the nineteenth century several investigators measured the minimum energy (at the cornea) required for a human to see a dim flash of light against

a dark background. It was apparently the physicist Lorentz who first realized that this energy corresponds to less than one hundred photons (Bouman 1961); if 50 – 90% of these photons are lost due to scattering in the eye (which is not unreasonable), then we can see less than 10 – 50 photons at the retina. Can we test this idea directly? Could the true threshold of vision be just one photon?

Suppose that we are looking at dim light flashes from a conventional light source. Setting the 'intensity' $I$ of the flash really amounts to setting the *mean* energy delivered by the light source. This mean energy, in turn, determines the mean number of photons $\langle n \rangle$ counted by the photoreceptor cells, so that $\langle n \rangle = \alpha I$. The constant $\alpha$ includes lots of complicated things—the probability that a photon emitted by the light source actually reaches the retina, the probability that it is absorbed by the receptor pigment rhodopsin, and the probability that the receptor responds to the absorbed photon. Much of the cleverness in these experiments is designed to get around the lack of a precise value for this constant.

With conventional light sources, the actual number of photons $n$ in a single flash is a random variable with a Poisson distribution, $P(n) = e^{-\langle n \rangle}\langle n \rangle^n / n!$. This is the same Poisson distribution that we discussed in section 2.1.4 as a model for spike statistics; for photons under certain conditions this model is exact. If the observer is willing to say "I saw it" when at least $K$ photons are counted, then the probability of seeing is just the sum of $P(n)$ over all the counts $n$ greater than or equal to the threshold $K$:

$$P_{\text{see}}(I) = \exp(-\alpha I) \sum_{n=K}^{\infty} \frac{(\alpha I)^n}{n!}. \tag{4.1}$$

There are two key ideas here. First, the response of humans to dim light flashes is predicted to be probabilistic—there is a *probability* of seeing—but this randomness reflects the stochastic arrival of photons at the retina rather than some internal biological variability. Second, the function $P_{\text{see}}(I)$ is diagnostic of the threshold photon count $K$. Indeed, if we look at a plot of the probability of seeing versus the logarithm of the light intensity, its shape is invariant to variations in $\alpha$ but *does* depend on $K$.

In one of the classic experiments of modern biophysics, Hecht, Shlaer, and Pirenne (1942) measured $P_{\text{see}}(I)$ and found excellent fits to Eq. (4.1) with $K = 5$ to 7, as shown in Fig. 4.2. It would thus seem that humans can "see" as few as five photons, and under the conditions of these experiments these photons are distributed over many photoreceptors. This implies that single photon



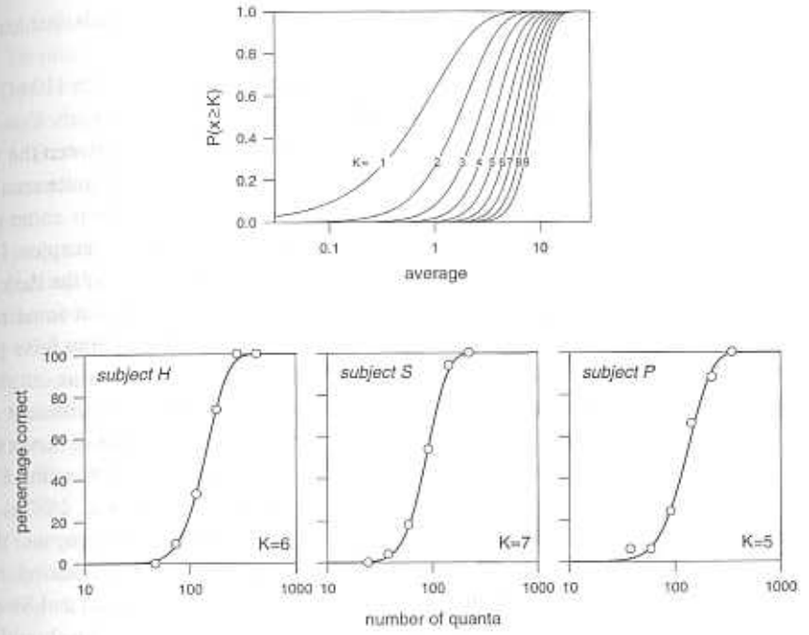



**Figure 4.2**
Frequency of seeing curves. Top panel illustrates the (theoretical) probabilities that more than $K$ photons will be counted at a detector illuminated by a light source with Poisson statistics, shown as functions of the average photon count. Hecht, Shlaer and Pirenne (1942) fit measurements of the frequency of seeing as a function of flash intensity at the cornea by choosing the best fitting member from this family of curves. Results are shown in the bottom panels. The broad transition from flashes which are never seen to flashes which are always seen arises from statistical fluctuations in the number of photons absorbed in each individual flash. The threshold $K$ for seeing determined in this way is 5–7 photon absorptions total. Because the flash stimulates photoreceptors drawn at random out of a large area on the retina, the probability of double hits is very low. From this one can conclude that individual photoreceptors reliably signal the absorption of a single photon. Adapted from Hecht, Shlaer, and Pirenne (1942).

arrivals at individual photoreceptors must generate signals that contribute to the detection process.

In an independent series of experiments, van der Velden (1944) also measured $P_{see}(I)$ and found that Eq. (4.1) was a good fit with $K = 2$. Barlow (1956) suggested a resolution of the apparent conflict between the two experiments by recognizing that there must exist a (probably quite small) level of noise in the visual system, so that even in the dark there is some probability of registering a nonzero photocount at the output of the receptor. In this case an observer could avoid reporting falsely that he "sees" in the dark by choosing a high threshold $K$; on the other hand, this means that some real flashes will be missed. Conversely, one could be unconcerned about false positive responses and then operate at a low value of $K$. It indeed turns out that the two measurements of $P_{see}$ were taken on observers with very different false positive rates, and more recently it has been found that when observers are asked to adopt different false positive rates, the threshold photocount $K$ varies as expected from Barlow's dark noise hypothesis (Teich et al. 1982a). These arguments are part of a more general realization that most apparent "thresholds" for sensation or perception in fact represent an adjustable criterion that is set to achieve a desired reliability in the presence of noise (Green and Swets 1966).[1] If we could change the statistics of photon arrivals, then we should be able to change this tradeoff between sensitivity and false alarm rate, and with modern optical techniques this prediction has been confirmed (Teich et al. 1982b).

It was against this background of psychophysical experiments and theoretical ideas that Barlow and Levick set out, in the late 1960s, to measure the reliability of detection and discrimination in retinal ganglion cells of the cat. These are the same ganglion cells where Kuffler (1953) had made the first measurements of receptive fields in mammalian vision, and where Barlow, FitzHugh, and Kuffler (1957) had demonstrated that receptive fields adapt to

1. As explained by Green and Swets (1966), this understanding emerged, in large part, through the application of modern statistical decision and signal detection theories to the analysis of psychophysical experiments. These theories, in turn, are often presented as having been constructed either as solutions to purely mathematical problems or as tools for the design of manmade systems. In fact much of the mathematical theory was worked out during World War II in connection with the development of radar. In the early radar systems, as for air traffic controllers today, the task was ultimately to make signals detectable to human observers, so in this sense the entire subject has its origins in a psychophysical context. The results of the American radar effort are described in a series of volumes orginally published by McGraw-Hill as the MIT Radiation Laboratory Series; the volume by Lawson and Uhlenbeck (1950) describes the development of detection theory as well as experiments done on both the electronic instruments and the human observers.



changes in mean light level. Note that at this point there had been no direct observations of single photon responses in individual cells from vertebrate visual systems, although Fuortes and Yeandle (1964) had seen "quantum bumps" in the response of receptors from the horseshoe crab *Limulus polyphemus*.

Barlow and Levick characterized the response of retinal ganglion cells by counting the number of spikes in a fixed time window following the presentation of a brief flash of light. For dim flashes, the spike count increases in proportion to flash intensity, so that there is a constant ratio between photons and spikes. Furthermore, if one examines the distribution of spike counts in the absence of a flash, the distribution is roughly Gaussian, as shown in Fig. 4.3a, so that it is characterized by its mean and its variance. Clearly, both these parameters vary with the width of the time window $\tau$, but it turns out that the variance $\sigma^2$ of the spike count in the steady discharge is proportional to the mean count $\langle n \rangle$.

The problem of intensity discrimination is that we observe a spike count $n$ and must decide if it represents background activity or if in fact a flash was present. If the distribution of background spike counts is Gaussian, and the extra spikes from the flash just add, then we have the situation shown in Fig. 4.3b. We can choose some criterion $n_c$ and guess that a flash occurred if $n > n_c$. Clearly if $n_c$ is very small, we never miss any flashes, but we also have many "false alarms" where we assign the background discharge to a real flash. On the other hand, if $n_c$ is very large, then we have a low false alarm rate but we also miss many of the real flashes. Thus by changing our criterion we can trade different types of errors against one another. This criterion dependent trading is the basis for Barlow's explanation of the difference between the Hecht, Shlaer, Pirenne and the van der Velden experiments.

If we don't know the absolute intensity of the light flash, we have to set our criterion in relation to the statistics of the background discharge. Thus we will say the flash is present if the spike count $n \geq \langle n \rangle + k\sigma$, where the constant $k$ determines the probability of a false alarm. This means that the flash will be detected if it produces $\Delta n = k\sigma$ spikes, but we know that the number of spikes is proportional to the number of absorbed photons. In addition, we know that the variance $\sigma^2$ is proportional to the mean spike count $\langle n \rangle$, which is in turn determined by the background light intensity. All of this means that we can turn the threshold change in spike count $\Delta n$ into a threshold change in light intensity $\Delta I$ and study the dependence of this increment threshold on the intensity of the background light $I$. The results are shown (for one cell) in Fig. 4.4.

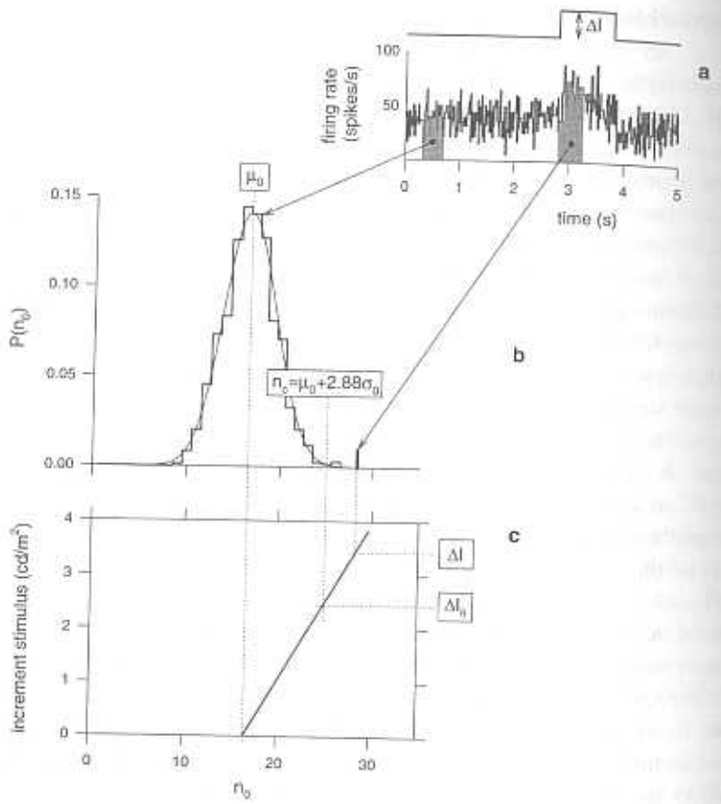



**Figure 4.3**
Spike counts and discrimination problem adapted from Barlow and Levick (1969). (a) Barlow and Levick measured the firing rate, as in Fig. 2.1, in response to a 1 s step of light of intensity $\Delta I$ superposed on a background. They then compared the mean rate during a 0.5 s period prior to the light step and during the light step. Measurements prior to the stimulus described the probability distribution for spike counts shown in (b). The criterion chosen for reliable detection was that the spike count during the light step should exceed 2.88 standard deviations from the mean of distribution. As the step intensity $\Delta I$ is increased, the spike counts increase, eventually reaching this criterion $n_c$. In (c) this criterion count is converted into a criterion intensity increment $\Delta I_\theta$.



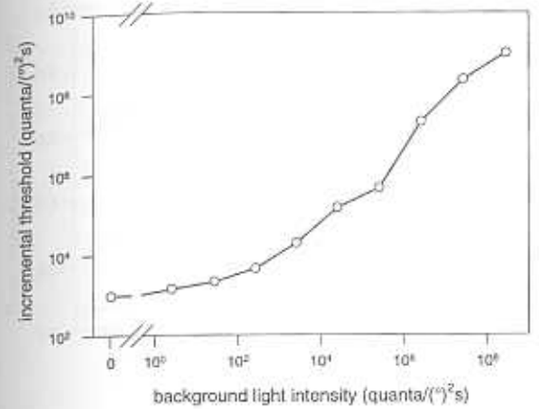



**Figure 4.4**
Threshold intensity as a function of background redrawn from Barlow and Levick (1969). The increment threshold determines the difference in light intensity $\Delta I$ such that flashes of intensity $I$ and intensity $I + \Delta I$ can be distinguished, as explained in Fig. 4.3.

Crudely speaking, we see that the threshold for reliable discrimination in a single cell is proportional to the light intensity, $\Delta I \propto I$, in bright backgrounds; this is the familiar "Weber's law" behavior of psychophysical thresholds. At intermediate intensities there is a hint that $\Delta I \propto I^{1/2}$, which is consistent with discrimination being limited by the random arrival of photons at the retina, as discussed by de Vries (1943) and Rose (1948). Finally, at very low light levels the increment threshold becomes constant, presumably limited by dark noise (Barlow 1956). Certainly these observations are in qualitative accord with the behavior of humans in response to similar stimuli. The fact that spike/photon ratios vary widely across intensities suggests that this basic unit of visual transduction is controlling the sensitivity of the whole animal, but the point is far from proven. In particular, one would like to work at very low light intensities where the retina is dark adapted, comparable to the conditions used in the Hecht, Shlaer, and Pirenne (1942) experiments. This was done by Barlow, Levick, and Yoon (1971).

A dramatic example of the sensitivity of dark adapted ganglion cells is shown in Fig. 4.5. Here we see the average response to flashes that deliver (on average) 5 photons at the cornea. The spike rate is elevated by more than a factor of three for a brief period, producing an extra 2.5 spikes in a 200 ms

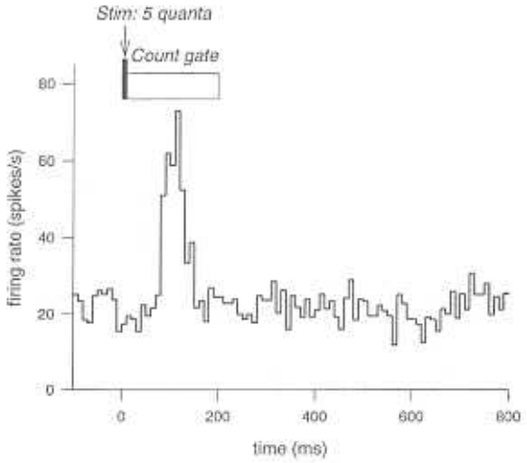



**Figure 4.5**
Post-stimulus time histogram for dark-adapted ganglion cell to flash delivering 5 photons on average at the cornea. The stimulus produces about 2.5 extra spikes during the counting window shown. The ability of this cell and others like it to respond to flashes producing only a few photon absorptions indicates that the dark-adapted retina may indeed respond to individual light quanta. Redrawn from Barlow, Levick and Yoon (1971).

window. If 2 (= 5/2.5) photons at the cornea are sufficient to trigger an extra spike, we have to take seriously the idea that individual photons arriving at the retina in fact produce more than one spike. Once again, estimating the exact fraction of light reaching the retina is difficult, but one can try various statistical arguments.

Let us suppose that each counted photon produces exactly $m$ spikes. Then, since photons arrive at random from a Poisson distribution, the change in mean photon count is necessarily accompanied by identical changes in variance (recall the discussion of the Poisson distribution in section 2.1.4). But, for the spikes, the changes in mean count are amplified by the factor $m$, while the changes in variance are amplified by $m^2$. Thus the variance of the spike count distribution should also vary linearly with the flash intensity, which it does, and by comparing the slopes of excess mean and excess variance we calculate the spike per photon ratio $m$. This analysis leads to values of $1.65 \leq m \leq 3.80$, clearly indicating that one photon produces several spikes.

If, for example, one photon produces 3 extra spikes, then by setting a criterion of 3 spikes we should produce "frequency-of-seeing" curves analogous to



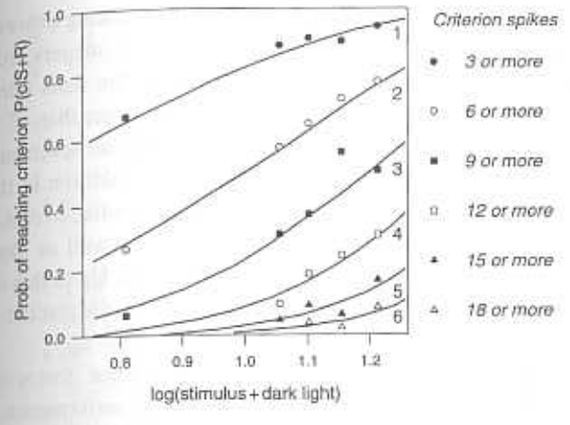



**Figure 4.6**
Frequency of seeing curves for retinal ganglion cells. Barlow, Levick, and Yoon measured the probability that a dim flash produced at least a criterion number of spikes in a retinal ganglion cell as a function of the flash intensity. Plotted here are the results of an experiment of this type for a number of different spike count criteria. These measurements of the probability of reaching a particular spike count are plotted against the intensity of the stimulus delivered plus an added number of random photon-like events (the "dark light") to account for noise generated in the retina. Both the number of photon absorptions produced by the stimulus and the number of random photon-like events are assumed to follow Poisson statistics. The resulting curves, analogous to the frequency of seeing curves of Hecht, Shlaer, Pirenne, can be fit assuming that each photon absorbed produces 3 spikes. Redrawn from Barlow, Levick and Yoon (1971).

those measured by Hecht, Slaer, and Pirenne, but these curves should be fit by Eq. (4.1) with $K = 1$. If we set a criterion of 6 spikes, we should find $K = 2$, and so on. As shown in Fig. 4.6, this is exactly what one finds. Thus we see that the statistics of decisions based on spikes from a single neuron, just like the decisions of human observers, reflect the statistics of photon arrivals at the retina. To get a precise fit to the data we must assume that the visual system has a finite level of dark noise, and that this noise is indistinguishable from random photon arrivals at some effective rate (Barlow 1956).

It is clear that, when recording from a retinal ganglion cell in the cat, we (as external observers) can manipulate our criterion, so that the apparent threshold photon count can take any value, including unity. Can human observers similarly lower their threshold down to one photon? Equivalently, can our brain change its rules for interpreting the spike trains from our own retinal ganglion cells?

Sakitt (1972) generalized the $P_{see}$ measurement by asking observers to rate their perception of the intensity of dim flashes using integers from 0 to 6. She found that the mean rating varied linearly with the flash intensity, but, more importantly, the probability of giving a rating larger than $K$ obeyed an equation of the same form as Eq. (4.1) for each $K$, with all seven curves being fit by the same $\alpha$. To achieve the best fit requires, in addition to the photons delivered in the flash, a small rate of spontaneous photonlike events, which we identify with Barlow's dark noise; this experiment (as well as many others) gives an estimate for the rate of these photonlike events. Up to this dark noise, the observer is behaving as expected if the rating number is just the number of photons counted!

By now it is clear that individual photoreceptors must give a reliable response to single photons if we are to understand the performance of human observers or of cat ganglion cells. As soon as it was possible to make good intracellular recordings of voltages in vertebrate photoreceptors, several groups searched unsuccessfully for single photon voltage responses. As it turns out, the receptor cells are electrically coupled into a network, so that currents produced in one cell are spread as voltage responses over many cells. To record single photon responses thus requires direct measurement of the receptor cell current, which was accomplished by Baylor, Lamb, and Yau (1979a). This led to the observation of highly reproducible single photon photocurrents in toads (Baylor, Lamb, and Yau 1979b) and in macaque monkeys (Baylor, Nunn, and Schnapf 1984); an example of these single photon experiments is shown in Fig. 4.7.

In recordings from single rod cells one also observes spontaneous photonlike events, which almost certainly reflect the spontaneous isomerization of the visual pigment rhodopsin (Baylor, Matthews, and Yau 1980). If we convert Sakitt's estimate of the dark noise into a rate per rod cell, the rate we obtain is in excellent agreement with the observed spontaneous event rate in monkey rods (Baylor, Nunn, and Schnapf 1984). This agreement strongly suggests that the limits to the reliability of night vision are set by noise in the photoreceptor array itself, not by noise or inefficiencies in the subsequent neural processing. It is yet another remarkable fact about the visual system that, to be consistent with the measured dark noise levels, the spontaneous isomerization rate per rhodopsin molecule must be less than once per 3,000 years at room temperature (in toads) or once per 300 years at mammalian body temperatures.

For cold blooded vertebrates it is possible to make a more direct comparison of dark noise levels at different stages in visual signal processing. In frogs and toads, Aho et al. (1988) measured the dark noise both in behavioral ex-



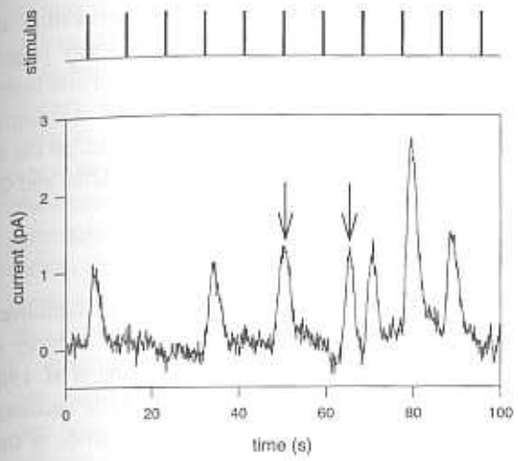



**Figure 4.7**
Single photon responses in a toad rod. Dim flashes, resulting in absorption of an average of less than 1 photon, were delivered at the times indicated by the stimulus trace, while the current flowing into the outer segment was measured with a suction electrode. Some flashes fail to elicit a response, some elicit a response of about 1 pA, and one elicits a response roughly twice this size. These responses reflect the absorption of zero, one, or two photons. This section of the experiment also shows two events, marked by arrows, due to the spontaneous or thermal activation of the photopigment rhodopsin. Such events occur at random, mimicking single photon absorptions, and provide the dark noise that limits the reliability of photon counting. From experiments by D. A. Baylor and F. Rieke.

periments and in recordings from retinal ganglion cells, which are the output cells of the retina. More importantly, one can vary the temperature of frogs and toads (this is difficult with people), and it was shown that the behavioral dark noise varies with temperature exactly as predicted from measurements on the activation energy of the spontaneous event rate in photoreceptors. If one extrapolates to mammalian body temperatures, this correspondence between behavioral and rod cell noise levels perfectly intersects the data point for humans and monkeys.

The agreement between behavioral and physiological data, especially over such a wide range of temperatures, strongly supports the conclusion that the organism is reaching a fundamental limit to the reliability of seeing, namely the noise in the photoreceptor itself (Barlow 1988; Donner 1989). It has been appreciated for many years that the ability of the visual system to count photons places important constraints on the mechanisms of phototransduction

within the rod cell. To account for the near optimal *processing* of these signals, neural computation must be very reliable. The many stages of neural processing that culminate in the behavioral response to a dim light flash must add little if any noise to the rod cell output, and the rod cell outputs must be processed in a maximally efficient manner so as to extract all the information available regarding the number of absorbed photons (Bialek and Owen 1990; Rieke, Owen, and Bialek 1991).

### 4.1.3 Auditory discrimination

In the early 1960s, a group at MIT set out to give a quantitative characterization of neural responses at various levels of the cat auditory system. As the group's first results were becoming available (Kiang et al. 1965), Siebert (1965) and Weiss (1966) set out to make models of the auditory nerve response and to connect these models both to the mechanics of the inner ear and to the phenomenology of auditory perception. This was an ambitious program, and probably there is no sensory system in which such a program has been brought to completion. On the other hand, these early papers presented a rather rigorous mathematical formulation of issues that are still with us—the role of temporal versus rate coding, the sharing of information among many neurons, and the connection between neuronal reliability and psychophysical discrimination performance. Here we focus on this last problem, although the issues are (perhaps inextricably) intertwined.

We should warn the reader that a straightforward attempt to relate the reliability of auditory discrimination to the statistics of auditory nerve responses actually fails. It seems that we should be able to discriminate frequencies, for example, with at least an order of magnitude more precision than is observed. The resolution of this paradox forces us to think carefully about the nature of the estimation problems the brain is solving, and along the way we also develop methods and ideas that will be useful throughout the remainder of the text.

As described in sections 2.1.2 and 2.1.4, the response of a primary auditory neuron to a pure tone stimulus approximates a Poisson process in which the rate is modulated by the stimulus waveform. There are corrections to this picture, but let us first try to understand the consequences of the Poisson approximation. We are interested in connecting the statistics of neural firing to the reliability of perceptual judgments, in the spirit of the Barlow–Levick experiment. Indeed, what follows is a generalized Barlow–Levick experiment performed on a model neuron. The advantage of the Poisson approximation,



as indicated in section 2.1.4, is that we can make considerable progress using pen and paper rather than computer simulation.

Siebert was concerned particularly with the problem of frequency discrimination. In the neighborhood of $f = 10^3$ Hz, humans can distinguish reliably between pure tones that differ in frequency by as little as $\Delta f \sim 1 - 3$ Hz. We know that the response of auditory neurons to pure tones changes as we change the stimulus frequency, but we also know that primary auditory neurons do not give deterministic responses. The Poisson model is an approximate description of this nondeterministic behavior. The fact that neurons do not always give the same response will limit our ability to make fine discriminations. How does this limit compare with human performance?

Imagine that we are trying to discriminate between two different stimuli, say stimulus + and stimulus −. We are going to observe the spike train of one neuron over a time window $0 < t < T$, and in this window the two stimuli give rise to time dependent firing rates $r_+(t)$ and $r_-(t)$, respectively. The discrimination problem is that we observe the sequence of spike arrival times $t_1, t_2, \cdots, t_N$ and must decide whether the neuron was driven by stimulus + or −. This is an example of the broad class of decision or discrimination problems that have often been used to probe the reliability of the nervous system, starting with the traditional psychophysical experiments on human observers (Green and Swets 1966). We measure performance by counting the fraction of decisions that are made correctly, so to find the limits of reliability we need to process the data $\{t_i\}$ so as to maximize this fraction. The optimal strategy in this sense is *maximum likelihood*—given the data $\{t_i\}$, choose the stimulus + or − to maximize the probability that the spike train was generated by that stimulus. The fact that maximum likelihood is the correct strategy to maximize the fraction of correct decisions is one of the most useful facts from the general mathematical theory of signal processing and decision making, so we sketch the proof in section A.16, and a simple example of the discrimination problem is shown in Fig. 4.8.

For our discussion here, the important point is that if we process the spike arrival times using the maximum likelihood decision rule we are guaranteed that we will calculate the maximum possible fraction of correct decisions that can be made using these spike sequences. In this sense the calculation gives us a true limit to the reliability of discrimination. Our task now is to carry through this maximum likelihood calculation in the case of a Poisson neuron.

We recall (from sections 2.1.4 and A.4) that, in the Poisson approximation, the probability that a neuron produces spikes at times $t_1, t_2, \cdots, t_N$ in response to stimulus + is given by Eq. (2.18), that is,

$$P[t_1, t_2, \ldots, t_N|+] = \frac{1}{N!} \exp\left[-\int_0^T dt r_+(t)\right] r_+(t_1) r_+(t_2) \cdots r_+(t_N). \quad (4.2)$$

Each factor $r_+(t_i)$ measures the probability that a spike occurred at the time we observe it, the exponential enforces the fact that no spikes occur at any other times, and the $N!$ divides out the number of different ways of assigning labels to the $N$ spikes. When we have seen a particular spike train $t_1, t_2, \cdots, t_N$, we

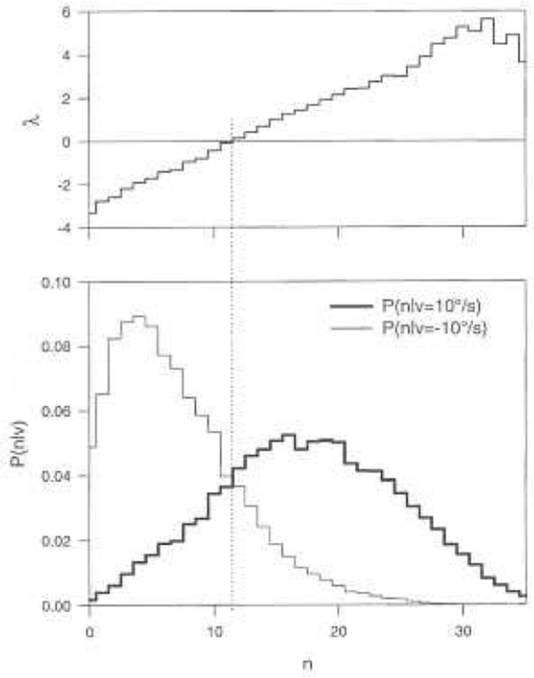



**Figure 4.8**
Construction of the log-likelihood function. The bottom panel shows the probability $P(n|v)$ of observing a spike count $n$ given a stimulus $v$ of either $10^\circ\,\mathrm{sec}^{-1}$ (thick line) or $-10^\circ\,\mathrm{sec}^{-1}$ (thin line), in an experiment on the fly H1 neuron. These conditional distributions are slices through the three dimensional plot shown in Fig. 2.2e. The decision variable $\lambda(n)$ plotted in the top panel is defined as

$$\lambda(n) = \log\left[P(n|v = 10^\circ\,\mathrm{sec}^{-1})/P(n|v = -10^\circ\,\mathrm{sec}^{-1})\right].$$

If $\lambda > 0$, then the most likely stimulus is $v = 10^\circ\,\mathrm{sec}^{-1}$; if $\lambda < 0$ the most likely stimulus is $v = -10^\circ\,\mathrm{sec}^{-1}$.



have to decide whether the stimulus was + or −. To do this, we calculate the likelihood ratio, which is just the ratio of probabilities that this spike train was produced by either of these stimuli. It is convenient to take the logarithm of this quantity, and this "log-likelihood ratio" $\lambda(t_1, t_2, \cdots, t_N)$,

$$\lambda(t_1, t_2, \cdots, t_N) = \ln\left(\frac{P[t_1, t_2, \cdots, t_N|+]}{P[t_1, t_2, \cdots, t_N|-]}\right). \quad (4.3)$$

provides a convenient decision variable: maximum likelihood is the statement that if $\lambda > 0$ we guess that the spike train was produced by stimulus +, and if $\lambda < 0$ we guess −. Again, this procedure is optimal in that it will maximize the probability of a correct decision.

In the Poisson model, it is easy to find the form of $\lambda$, substituting Eq. (4.2) (and its analog for stimulus −) into Eq. (4.3). The result is that

$$\lambda(t_1, t_2, \cdots, t_N) = \ln\left[\frac{r_+(t_1)}{r_-(t_1)}\right] + \ln\left[\frac{r_+(t_2)}{r_-(t_2)}\right] + \cdots + \ln\left[\frac{r_+(t_N)}{r_-(t_N)}\right] - \int_0^T dt[r_+(t) - r_-(t)]. \quad (4.4)$$

Thus we see that optimal discrimination involves adding up contributions from each individual spike, then subtracting off a constant. The constant just insures that we can put our discrimination threshold at $\lambda = 0$. If we carry through this optimal discrimination procedure—observe the spikes, compute $\lambda$, make a decision depending on whether $\lambda$ comes out positive or negative—how often will we get the right answer?

Our discrimination procedure works because, as illustrated in Fig. 4.8, the stimulus + generates, on average, spike trains that map to positive values of $\lambda$, and the converse is true for stimulus −. So let's start by computing these average values. The average of $\lambda$, given that stimulus + is being presented, is, by definition,

$$\langle\lambda(t_1, t_2, \cdots, t_N)\rangle_+ = \sum_{N=0}^{\infty} \int_0^T dt_1 \int_0^T dt_2 \cdots \int_0^T dt_N\, \lambda(t_1, t_2, \cdots, t_N) \times P[t_1, t_2, \cdots, t_N|+]. \quad (4.5)$$

We can now substitute Eq. (4.4) into this expression and go to work reducing things down to something manageable. The key to the calculation is recognizing that $\lambda$ is a sum of terms that refer to the occurrence of individual spikes, whereas $P[t_1, t_2, \cdots, t_N|+]$ is a product of such terms. This means that when

we integrate over all the spike arrival times, we can break things up into sums and products over integrals that involve just one arrival time. Then we have to count how many of these terms occur, and add them up, being careful to identify the infinite series which adds up to cancel the exponential in front of $P[t_1, t_2, \ldots, t_N|+]$. All these same steps can be used to compute the average of the sum $\sum_{i=1}^{N} f(t_i)$, where $f$ is an arbitrary function; in the case of interest, $f(t) = \ln[r_+(t)/r_-(t)]$. Mathematical details are in section A.17, and the general result is

$$\left\langle \sum_{i=1}^{N} f(t_i) \right\rangle_+ = \int_0^T dt r_+(t) f(t). \tag{4.6}$$

Now we can use this general result to compute the average value of $\lambda$ given that the stimulus was "+":

$$\langle \lambda(t_1, t_2, \cdots, t_N) \rangle_+ = \left\langle \sum_{i=1}^{N} \ln \left[ \frac{r_+(t_i)}{r_-(t_i)} \right] \right\rangle_+ - \int_0^T dt [r_+(t) - r_-(t)] \tag{4.7}$$

$$= \int_0^T dt r_+(t) \ln \left[ \frac{r_+(t_i)}{r_-(t_i)} \right] - \int_0^T dt [r_+(t) - r_-(t)], \tag{4.8}$$

and we can do the same calculation given that the stimulus was "−":

$$\langle \lambda(t_1, t_2, \cdots, t_N) \rangle_- = \int_0^T dt r_-(t) \ln \left[ \frac{r_+(t_i)}{r_-(t_i)} \right] - \int_0^T dt [r_+(t) - r_-(t)]. \tag{4.9}$$

Thus we see, as we hoped, that our decision variable is different, on average, in response to the two signals. The magnitude of this difference is the "signal" for making the discrimination. The fact that $\lambda$ fluctuates around these average values provides the "noise" against which we must fight.

The magnitude of the discrimination signal is

$$\Delta M = \langle \lambda(t_1, t_2, \ldots, t_N) \rangle_+ - \langle \lambda(t_1, t_2, \ldots, t_N) \rangle_-$$

$$= \int_0^T dt \left[ r_+(t) - r_-(t) \right] \ln \left[ \frac{r_+(t)}{r_-(t)} \right]. \tag{4.10}$$

We are especially interested in cases where different stimuli are just barely discriminable, so that the difference in rates $\Delta r(t) = r_+(t) - r_-(t)$ is small. Clearly if $\Delta r(t)$ is zero, then there is no signal for discrimination and $\Delta M$ is also zero. We would like to find the value of $\Delta M$ at small but nonzero $\Delta r(t)$, and to do this we use a Taylor series expansion and systematically discard higher powers of $\Delta r(t)$ which are progressively less important at smaller



$\Delta r(t)$. Details are given in section A.17, and we find

$$\Delta M \approx \int_0^T dt \frac{[\Delta r(t)]^2}{r(t)}, \tag{4.11}$$

where $r(t) = (1/2)[r_+(t) + r_-(t)]$ is the average of the two rates.

The reliability of discrimination is determined by comparing the mean signal with its variance; higher moments turn out not to be important. By the same arguments used to arrive at Eq. (4.6), one can compute the variance of any sum $\sum_{i=1}^{N} f(t_i)$. The result has the simple form

$$\left\langle \left[ \delta \sum_{i=1}^{N} f(t_i) \right]^2 \right\rangle_+ = \left\langle \left[ \sum_{i=1}^{N} f(t_i) \right]^2 \right\rangle_+ - \left[ \left\langle \sum_{i=1}^{N} f(t_i) \right\rangle_+ \right]^2$$

$$= \int_0^T dt r_+(t) [f(t)]^2. \tag{4.12}$$

In our case, with $f(t) = \ln[r_+(t)/r_-(t)]$, we find the variance of $\lambda$ to be

$$\langle \left[ \delta\lambda(t_1, t_2, \cdots, t_N) \right]^2 \rangle_+ = \int_0^T dt r_+(t) \left( \ln \left[ \frac{r_+(t)}{r_-(t)} \right] \right)^2$$

$$\approx \int_0^T dt \frac{[\Delta r(t)]^2}{r(t)}, \tag{4.13}$$

where in the last line we make the approximation once again that $\Delta r$ is small. Note that in this limit the variance of $\lambda$ is the same in response to either the + or the − stimulus. Thus it makes sense to think of the discrimination task with a signal corresponding to the difference in means $\Delta M$ and a noise with variance

$$\sigma^2 = \int_0^T dt \frac{[\Delta r(t)]^2}{r(t)}. \tag{4.14}$$

Then we have a signal-to-noise ratio

$$SNR = (\Delta M)^2/\sigma^2$$

$$= \left( \int_0^T dt \frac{[\Delta r(t)]^2}{r(t)} \right)^2 \left( \int_0^T dt \frac{[\Delta r(t)]^2}{r(t)} \right)^{-1}$$

$$= \int_0^T dt \frac{[\Delta r(t)]^2}{r(t)}. \tag{4.15}$$

This gives us the signal to noise ratio for discrimination between *any* two signals using the output of a neuron with firing statistics corresponding to a Poisson process, at least in the limit where the stimuli are very similar.

To understand the result of Eq. (4.15) it is convenient to think about a special case. Suppose that the rates are not time dependent. Then it doesn't matter exactly when the spikes occur, and all we have to do is count the number of spikes. We know that the two stimuli, on average, will give spike counts of $N_+ = r_+T$ and $N_- = r_-T$, respectively, so the difference in spike count is $\Delta N = \Delta r T$. We also know that for a Poisson process the variance of the spike count is equal to the mean, so that the signal to noise ratio becomes

$$\begin{aligned} SNR &= (\Delta N)^2/N \\ &= (\Delta r T)^2/(rT) = T\frac{(\Delta r)^2}{r}, \end{aligned} \tag{4.16}$$

which is the same result as we get from doing the integral in Eq. (4.15), as it must be. But, more importantly, we can imagine applying our simple counting result to very small time windows of size $dt$ and then adding up the resulting SNRs; we are allowed just to add them because, for a Poisson process, the spikes in different time bins occur independently. Then the total SNR is exactly the integral in Eq. (4.15).

To summarize, Eq. (4.15) gives us the signal to noise ratio for discrimination based on the output of a Poisson neuron, and this is just the generalization to time dependent rates of the simple idea that Poisson processes have $\sqrt{N}$ fluctuations. Now we can get back to Siebert's problem, the limit to frequency discrimination.

Firing rates of auditory neurons have a time dependence locked to the frequency and phase of the stimulus, as described in section 2.1.2. This means that

$$r(t) = R(\omega)g(\omega t), \tag{4.17}$$

where $R$ is the average firing rate in response to a tone of frequency $\omega$, and $g$ describes the shape of the "phase histogram," that is, the probability of spike occurring at different phases relative to the sine wave stimulus as in Fig. 2.4. Both $R$ and $g$ also depend on the intensity of the sound, but we don't write this out explicitly. Now, if we make a small change in the frequency $\omega$, we get a change in $r(t)$ that comes both from the change in average rate—that is, from a change in $R$—and from the change in $g$. Siebert showed that when we compute the signal to noise ratio for frequency discrimination, these terms contribute independently.



When we add up contributions from the many auditory neurons, this separation into two terms quantifies our intuition that there are two very different sources of information about frequency (Siebert 1970; de Boer 1976): place information, corresponding to the fact that cells emerging from different locations in the cochlea are tuned to different frequencies; and timing information, corresponding to the fact that interspike intervals tend to cluster around multiples of the period of the sound. These intuitive "sources of information" correspond mathematically to the contributions of $R$ and $g$, respectively.

Siebert found that there is more information available in the timing cues, and certainly for one cell we can arrange this to be true by studying frequencies near the characteristic frequency of that cell, that is, the frequency where $R$ is maximal. This maximum is in fact quite broad for loud tones, which is part of the reason timing is so much more informative. So, to simplify our discussion, let's assume that $R$ doesn't change when we change frequency. Then the change in time dependent rate is just

$$\begin{aligned} \Delta r(t) &\approx \frac{\partial r(t)}{\partial \omega}\Delta\omega \\ &= R\frac{\partial g(\omega t)}{\partial \omega}\Delta\omega \\ &= Rtg'(\omega t)\Delta\omega, \end{aligned} \tag{4.18}$$

where, as usual, $g'$ denotes the derivative of the function $g$. Now we can compute the signal to noise ratio for discrimination, substituting into Eq. (4.15):

$$\begin{aligned} SNR &= \int_0^T dt \frac{[\Delta r(t)]^2}{r(t)} \\ &= \int_0^T dt \frac{(Rt\Delta\omega)^2}{R}\cdot\frac{[g'(\omega t)]^2}{g(\omega t)}. \end{aligned} \tag{4.19}$$

The function $g$ is by definition a periodic function of $\omega t = \phi$, since $\phi$ is the phase of firing relative to the sine wave. If we can assume that the interval $0 < t < T$ contains many of these periods, which just means that we listen to many cycles of the pure tone, then we can replace the terms that depend on $g$ by an average over one cycle $0 < \phi < 2\pi$, and then do the remaining integral of $t^2$. The result is

$$SNR = \frac{2\pi}{3}R(\Delta\omega)^2T^3\int_0^{2\pi}\frac{d\phi}{2\pi}\frac{[g'(\phi)]^2}{g(\phi)}. \tag{4.20}$$

This equation tells us that an observer of a single auditory neuron could, by processing the spike train optimally, make frequency discrimination with this $SNR$. With a typical firing rate of $R \sim 50\ s^{-1}$ and integration time of $T \sim 100$ ms, the signal to noise ratio for discriminating a frequency difference of $\Delta f$ Hz ($\Delta\omega = 2\pi\Delta f$) is then given by

$$SNR \sim 4\left(\frac{R}{50\,s^{-1}}\right)\left(\frac{\Delta f}{1\,\mathrm{Hz}}\right)^2\left(\frac{T}{100\,\mathrm{ms}}\right)^3\int_0^{2\pi}\frac{d\phi}{2\pi}\frac{[g'(\phi)]^2}{g(\phi)}. \tag{4.21}$$

The function $g$ is something that has been *measured* for thousands of auditory neurons. Furthermore, we have defined it so that it is normalized,

$$\int_0^{2\pi}\frac{d\phi}{2\pi}g(\phi) = 1, \tag{4.22}$$

so unless something very strange happens, the integral we need to evaluate in Eq. (4.21) can't be a very large or small number, but is most likely something of order one. In any case we know that our model of neurons as a Poisson process isn't exactly right, so maybe we shouldn't work too hard chasing factors of two. Then we can summarize all the work of this section by the simple conclusion:

$$SNR \sim \left(\frac{R}{50\,s^{-1}}\right)\left(\frac{T}{100\,\mathrm{ms}}\right)^3\left(\frac{\Delta f}{1\,\mathrm{Hz}}\right)^2. \tag{4.23}$$

This means that we should achieve reliable discrimination, that is $SNR = 1$, when the frequency difference is $\Delta f \sim 1$ Hz, and this results from "listening" to just one neuron for 100 ms.

Although it seems to be a nice, simple result, Eq. (4.23) is a disaster. Indeed, the whole message of Siebert's original work was that this simple picture of optimal discrimination based on spike trains is completely inconsistent with the phenomenology of human frequency discrimination (Siebert 1970).

The first problem is one of scale. With reasonably loud tones, many neurons will exhibit strong phase locking, and so the signal to noise ratio for an observer of the entire cochlea must be much larger than in Eq. (4.23). Correspondingly, the threshold for reliable frequency discrimination should be much smaller than 1 Hz, and it is not; Siebert estimated that the discrepancy is at least two orders of magnitude. The second problem is the dependence on time—human frequency discrimination thresholds improve approximately as $T^{-1/2}$, not $T^{-3/2}$ as predicted from Eq. (4.23). Again this discrepancy is far outside reasonable experimental error.



In our discussion of the homunculus in the introduction, we emphasized that the strategies adopted by the homunculus depend on the nature of the sensory world. If we can really assume that the world consists of one out of two possible pure tones, then all of the cells in the auditory nerve phase lock to one tone or the other, and hence all these cells provide some information about the frequency of the tone; frequency discrimination performance should be, therefore, much better than in Eq. (4.23). This means that (for example) cells tuned to frequencies near 10 kHz are giving information relevant to discrimination in the neighborhood of 1 kHz. For natural stimuli this is seldom if ever the case—events in very distant frequency bands are uncorrelated unless they are in near-harmonic relation.

Similarly, the world of pure tones is also a world of infinite correlation times, so that the frequency at one instant of time and the frequency 100 ms later are identical. Again this does not happen in natural stimuli, where frequencies are modulated and phase coherence is lost. The assumption that signals are coherent across the entire interval of stimulus presentation is what allows for the $T^{-3/2}$ improvement of discrimination performance.

Goldstein and Srulovicz have presented a model for optimal processing of interspike intervals that addresses some of these discrepancies between the strict frequency discrimination task and the more natural problem of pitch estimation in dynamic stimuli. (Goldstein and Srulovicz 1977; Srulovicz and Goldstein 1983). To begin, by focusing on interspike intervals one throws away information that could be carried by the long term phase coherence of the spike train, and hence one immediately recovers $T^{-1/2}$ rather than $T^{-3/2}$ improvements with integration time. This strategy is optimal for estimation of signals that are modulated on time scales comparable to the typical interspike intervals.

Like Siebert (1970), Srulovicz and Goldstein (1983) focus on the task of frequency discrimination, but they insist on a model that generates an estimate of the stimulus frequency without knowing in advance that one of two frequencies will be presented. This implies that information about discrimination in the neighborhood of frequency $f$ is indeed dominated by neurons tuned near $f$, as one might intuitively expect. The effective number of cells contributing to the estimate is therefore smaller, and this restriction of the frequency band plays a crucial role in generating the correct predictions for the frequency discrimination threshold.

In earlier work, Goldstein and colleagues (Goldstein 1973; Goldstein et al. 1978) had shown how the problem of frequency discrimination for pure tones could be linked to the more natural task of pitch discrimination for complex

sounds. The central idea of this work is that frequency discrimination experiments measure the precision with which individual frequency components are represented, and from these data one can work out the optimal strategies for identification of harmonic complexes. Applying these optimal strategies, one generates predictions for the accuracy of pitch discrimination in terms of the already measured accuracy of frequency discrimination. In principle, these predictions are parameter free, and they are indeed reasonably successful. The optimal processor theory also makes correct predictions for the perceived pitch of sounds which are not quite harmonic complexes, including the conditions for ambiguous percepts.

We want to emphasize that, although there are many open questions, the scope of the problem addressed in this work is extremely broad. The program is to go from the statistics of auditory nerve responses to the accuracy of frequency discrimination and finally to pitch perception, at each stage guided by the principle of optimal processing. We have learned that to make this program plausible one must be careful to define reasonably natural tasks, closer to the general estimation problem than to the more classical discrimination problem, and that one must think carefully about the dynamics of naturally occurring signals.

### 4.1.4 Motion discrimination in monkey vision

The spirit of the Barlow-Levick experiment is to explore the reliability of neurons in a discrimination task for which we know the performance of the organism as a whole. In the case of photon counting, the performance of the organism is equal to the physical limit imposed by the quality of the input signal, so that comparing a single neuron to the output behavior is the same as comparing it to the input receptor cells. In the case studied by Barlow and Levick, there is no doubt that all the information the animal uses in making perceptual decisions passes through the optic nerve, so one needn't demonstrate that the animal is "listening" to the particular cell studied in the experiment. A similar situation exists in invertebrates, where we know that the activity of single identified neurons can influence behavior. How should we think about the mammalian cortex, where aspects of visual information are shared among billions of neurons?

Perhaps the most direct attack on the problem of reliability in cortical neurons has come from Newsome and coworkers (see Newsome et al. 1990, for a summary), studying cells in area MT of the monkey visual cortex. A number of experiments indicate that area MT plays a vital role in visual motion estimation. Cells in this area give direction selective responses to moving patterns;



the response to motion is relatively robust to changes in the spatial pattern, so long as something is moving in the cell's receptive field. Damage to MT produces an immediate impairment of the monkey's performance in tasks requiring discrimination of the direction of motion, but not in tasks requiring only the detection of contrast (Newsome and Paré, 1988), and direct electrical stimulation of small regions in MT can bias the monkey's decisions about motion direction during performance of a discrimination task (Salzman, Britten, and Newsome 1990; Salzman et al. 1992). Starting from this foundation, Britten et al. (1992) set out to compare the performance of the monkey with that of individual MT neurons on the same motion discrimination task.

The key point about the Britten et al. experiments is that the monkey is doing the discrimination task *at the same time* that one is recording the activity of single neurons. This has three implications. First, one can compare neural and behavioral performance under identical conditions in one animal, rather than relying on population averages. Second, one can choose the parameters of the behavioral task to match the selectivity properties of the neuron being studied, thus maximizing the chances that the activity of this particular neuron is relevant to the behavior. Finally, one can compare neural activity and behavioral decisions to see if they are correlated on a trial by trial basis—does the monkey actually "say yes" more often when this one cell fires more spikes?

The task used in the MT experiments involves discrimination of the direction of motion in random dot patterns. Dots appear randomly on the screen and persist at fixed locations for a short time, less than the time resolution of the visual system. A small fraction of the dots are refreshed at new locations, displaced in space and time to simulate motion at a fixed velocity, and the remaining dots are refreshed at random locations. If all of the dots are replaced with a fixed spatiotemporal displacement, one sees completely coherent motion, and it is trivial to discriminate between motions in opposite directions. As the fraction of dots participating in coherent motion declines, the appearance of the display becomes more random and the reliability of discrimination declines, as shown schematically in Fig. 4.9a. The result of a psychophysical experiment is the probability of correct discrimination as a function of the fraction of dots participating in coherent motion, as in Fig. 4.9c. To make meaningful comparisons with neural data, dots are displayed in a region of space that matches the receptive field of the recorded neuron, and the axis of motion is along the axis to which the cell is most selective. In most of the experiments, the stimulus is displayed for two seconds, and the monkey must indicate a response by making an eye movement in the estimated direction of motion.

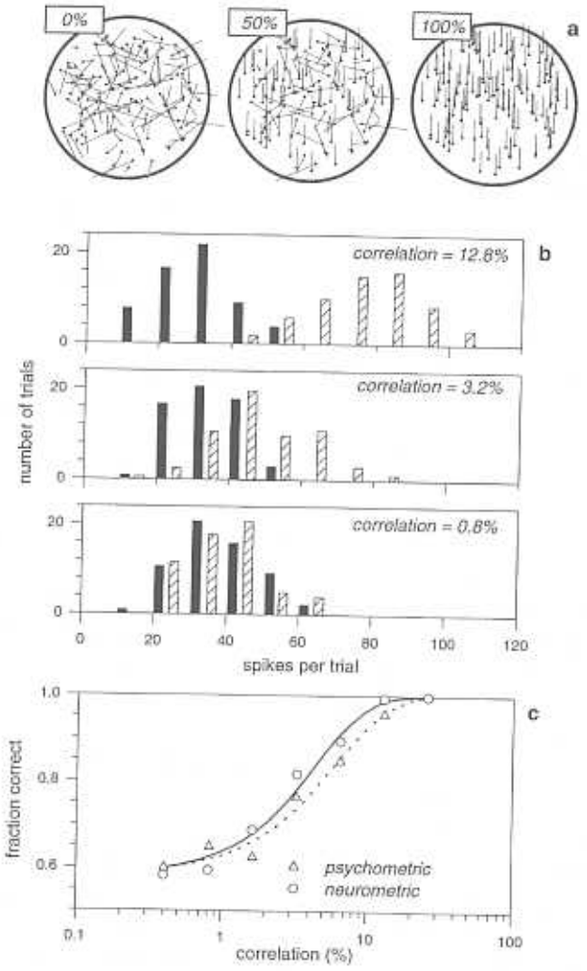



**Figure 4.9**
Discriminability of random motion stimuli. The stimulus in these experiments is a field of dots that are generated at random locations and refreshed either at random or at locations that simulate constant velocity motion; the fraction of dots with 'directed' refresh positions is called the correlation of the motion stimulus. The panels in (a) show three cases with correlations of 0%, 50%, and 100%. The panels in (b) show spike count distributions from an MT neuron at different correlation levels. The hashed bars represent responses to motion in the preferred direction of the neuron, and the dark bars represent responses to motion in the null direction. (c) Summary of measurements such as those in (b)—when more than 5% of the dots participate in coherent motion, the discriminability of motion in the preferred vs. null direction based on response from this single cell is quite reliable. Adapted from Newsome et al. (1990).



Britten et al. chose to quantify the neural response by counting spikes throughout the two seconds of stimulus presentation, so this aspect of the analysis is a direct application of the ideas developed by Barlow and Levick (1969) for retinal ganglion cells. For most cells, the probability of correct discrimination based on spike count is very similar to the probability of correct discrimination by the monkey; in some cases the monkey is more sensitive than the individual cell, in some cases the cell is more sensitive, and in most cases the dependence on the fraction of coherent dots is similar. The best performance in either case corresponds to 82% correct discrimination when just 2–3% of the dots are moving coherently. Although neural and behavioral performance are, on average, the same, since each task is tuned to the particular neuron being recorded one can ask whether there is a cell-by-cell correlation in the variability of the threshold, and this does not appear to be the case. Nonetheless, when the fraction of coherently moving dots falls to zero, so that the monkey is forced to guess, this guess is (weakly) correlated with the spike count on a trial by trial basis (Newsome et al. 1995; Britten et al. 1996). These results raise a number of interesting new questions.

The question that seems most obvious is why the monkey should not perform significantly better than an observer of one neuron. After all, there are many neurons in MT that are responding to the stimulus; why shouldn't the monkey average the responses of these cells to improve its performance? Zohary, Shadlen, and Newsome (1994) have given a simple answer, namely that the spike counts of different cells are correlated, so averaging does not significantly reduce the variance. More specifically, when they record from pairs of neurons in MT they find that, so long as both cells respond to the stimulus, the fluctuations in spike count are correlated. Thus it is not the case that one cell is correlated only with its nearest neighbor on some lattice, but rather that essentially all cells that have a chance of contributing to the behavioral decision exhibit small but significant pairwise correlations.

To appreciate the implications of this observation, imagine that we have $N$ identical neurons, each of which responds with a change in the mean spike count by an amount $\Delta m$ spikes, and each of which has a variance $\sigma^2$ in the spike count distribution. If we add up all the spikes, the total spike count will change by $N\Delta m$, and, because the cells are independent, the variance of the total spike count is just $N\sigma^2$. Thus the signal to noise ratio,

$$SNR = (\text{change in mean})^2/\text{variance} = N(\Delta m)^2/\sigma^2, \tag{4.24}$$

is improved by a factor of $N$. If threshold corresponds to a fixed signal to noise ratio, averaging over $N$ cells means that we can detect a change in spike count that is $\sqrt{N}$ times smaller, as expected. How is this changed by correlations?

For simplicity let us consider a model in which all cells that contribute to the discrimination have equal pairwise correlations, with correlation coefficient $\rho$. Each cell produces $m_i$ spikes, and the correlations are summarized by

$$\begin{aligned} \langle \delta m_i \delta m_j \rangle &= \sigma^2 \qquad i = j \\ \langle \delta m_i \delta m_j \rangle &= \rho\sigma^2 \qquad i \neq j, \end{aligned} \tag{4.25}$$

Now the change in the total number of spikes is still $N\Delta m$, but the variance of the total spike count is

$$\begin{aligned} \left\langle \left( \sum_{i=1}^{N} \delta m_i \right)^2 \right\rangle &= \sum_{i=1}^{N} \sum_{j=1}^{N} \langle \delta m_i \delta m_j \rangle \\ &= \sum_{i=1}^{N} \langle (\delta m_i)^2 \rangle + \sum_{i=1}^{N} \sum_{j \neq i} \langle \delta m_i \delta m_j \rangle \\ &= N\sigma^2 + N(N-1)\rho\sigma^2. \end{aligned} \tag{4.26}$$

Then we see that the signal to noise ratio becomes

$$SNR = \frac{(N\Delta m)^2}{N\sigma^2 + N(N-1)\rho\sigma^2} \to \frac{1}{\rho}(\Delta m)^2/\sigma^2, \tag{4.27}$$

where we indicate the limiting behavior for large numbers of cells, $N \to \infty$. We see that, in the presence of correlations, the signal to noise ratio never improves by more than a factor of $1/\rho$, no matter how many cells we include in our average. Correspondingly the threshold for the organism can never be more than a factor of $1/\sqrt{\rho}$ smaller than the threshold of a single cell. Zohary, Shadlen, and Newsome (1994) report an average $\rho \sim 0.12$, so we expect that behavioral thresholds shouldn't be more than a factor of three better than neural thresholds, and this is about right.

A further interesting feature of the MT data is that neural performance approximates behavioral performance even when the response of the neuron is summarized by the total spike count, without regard for the temporal structure of the response. Does this mean that the monkey is unable to "listen" to the timing of cortical spikes? Or, perhaps, that the temporal structure, even in principle, provides no information of relevance to the behavioral task? Such conclusions certainly seem in opposition to the view developed in the preceding chapters, namely that individual spikes can carry detailed information about the time variation of the sensory stimulus. It is, of course, possible that



the mammalian cortex is just different from systems we have described so far, but the available data suggest a more interesting possibility.

Most of the MT experiments focus on a discrimination task in which the moving random dot stimulus is presented continuously for two seconds; the magnitude and direction of motion are constant over this period, but of course the display fluctuates wildly when the fraction of coherently moving dots is small. If the task is changed so that stimuli are presented for only 100 ms, the behavioral thresholds for both humans and monkeys increase, but typically by a small factor (less than three). On the other hand, if one counts spikes for only 100 ms, the neural threshold for direction discrimination rises dramatically, and none of the individual cells described by Britten et al. (1992) approaches the performance of the most sensitive human or animal observers in the 100 ms window. Clearly, there is something different about the connection between neural and behavioral discrimination on these shorter time scales.

We have emphasized that questions about spike count versus spike timing codes must be phrased in the context of the stimulus dynamics. In several cases, there are interesting stimulus variations on time scales comparable to typical interspike intervals, so these variations must be represented by roughly one spike. In this view, "timing codes" are not mysterious—the timing of individual spikes signals the timing of particular events in the sensory world, as is made especially clear when we reconstruct the stimulus waveform. If we hold a stimulus constant for two seconds, we have defined a world in which there are no events to be timed, and we might therefore expect that the timing of individual spikes is irrelevant.

For the random dot displays used in the Britten et al. experiments, 100 ms of viewing in a 10° diameter receptive field corresponds to seeing roughly 130 dots, of which only 8% participate in coherent motion at threshold for the best human observers. In this limit the correct motion percept is thus carried by ~5 pairs of dots. With these stimulus parameters, the average spike count in a 100 ms window is no more than 5 spikes, and most likely is of order 1 to 3. This raises the possibility that hidden in the "constant" stimuli of these experiments there are indeed interesting sensory events—the occurrence of individual dot pairs or clusters—whose timing could be indicated by the timing of individual spikes. It is even possible that under normal conditions our perceptions are dominated by these individual events. Preliminary psychophysical experiments (Bair 1995) indicate that human observers can report fluctuating estimates of motion direction at various times in the presentation of the random dot stimulus, but that the pattern of these fluctuations is reproducible and tied to the details of the particular dot pattern being presented.

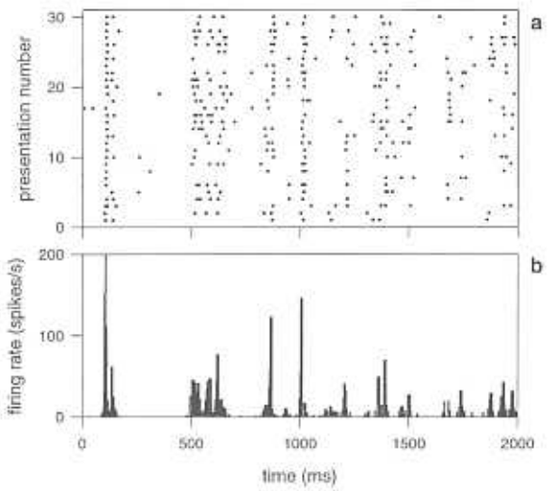



**Figure 4.10**
Raster plots of individual spike times and time dependent firing rate from experiments on MT neurons such as those in Fig. 4.9. The stimulus was presented from 0 to 2 s. In this case, there is no coherent motion, and dots appear at random on the screen and disappear after a short time. Repetition of the stimulus corresponds to a precise repetition of the location and arrival time of each dot relative to the $t = 0$ marker. Adapted from Bair and Koch (1996).

Bair and Koch (1996) have reanalyzed some of the experiments from Newsome and coworkers, focusing on those experiments where precisely the same random dot display was presented repeatedly on several trials. As shown in Fig. 4.10, repeated presentation of the same dots results in rather reproducible spike trains, a result that is certainly consistent with the hypothesis that the timing of particular sensory events is represented faithfully in the timing of individual spikes.

One of the most appealing aspects of the MT experiments is the trial by trial correlation of psychophysical and neural response. It seems possible that one could perform similar experiments, but force the monkey to respond to very brief stimuli, and in such a context one could search for correlations between behavior and the occurrence and timing of individual spikes. Not so long ago it might have seemed far fetched to suggest that the behavior of a monkey was correlated with the number of spikes in just one cell out of the billions in the cortex, although this is what Barlow (1972) proposed explicitly in his "neuron



doctrine for perception," and it is now clear that these correlations exist. We will return to the significance of individual spikes.

## 4.2 HYPERACUITY

Perhaps the most remarkable examples of reliability in neural computation are associated with situations where the absolute numerical data—thresholds for reliable discrimination, or equivalent noise levels at the output of a computation—surpass our intuitive estimates by orders of magnitude. In the extreme, echolocating bats can apparently resolve jitter in the arrival time of their echoes with a precision of 10 *nano*seconds (Simmons 1989). The weakly electric fish are not far behind, adjusting their own electrical signals in response to several hundred nanosecond shifts in signals from neighboring fish (Carr, Heiligenberg, and Rose 1986; Carr 1993). These results are surprising because the natural scale of neural activity seems to be milliseconds, not microseconds, and certainly not nanoseconds.

### 4.2.1 Where is the limit?

The experiments on extreme temporal acuity have their antecedents in experiments on spatial acuity in the visual system, experiments that date back well into the nineteenth century. What is our intuitive expectation about the scale of spatial precision? We know that the eye samples the world with a discrete lattice of photoreceptors, and in the human fovea this lattice spacing on the retina corresponds to an angular spacing of $\sim 0.01$ degrees in the visual world. Independent of the receptor lattice, diffraction through the pupil and lens will blur our image of the world, and this also washes out details smaller than $\sim 0.01$ degrees. These physical and geometrical considerations strongly suggest that the acuity of our foveal vision should correspond to angular displacements of about 0.01 degree, or (roughly) a displacement of one foot seen from a mile away, which gives a good feeling for how some of the early experiments were done.

Suppose that we ask a human observer to tell us whether the visual field contains two dots or a single dot of twice the brightness. Reliable discrimination occurs when the dots are separated by about 0.01 degrees, in agreement with our estimates of visual acuity. But something new happens if we ask the observer to discriminate the displacements of vernier patterns—now the threshold for reliable discrimination is roughly five times smaller, $\sim 0.002$ degrees. The smallest displacement thresholds reported for human observers are

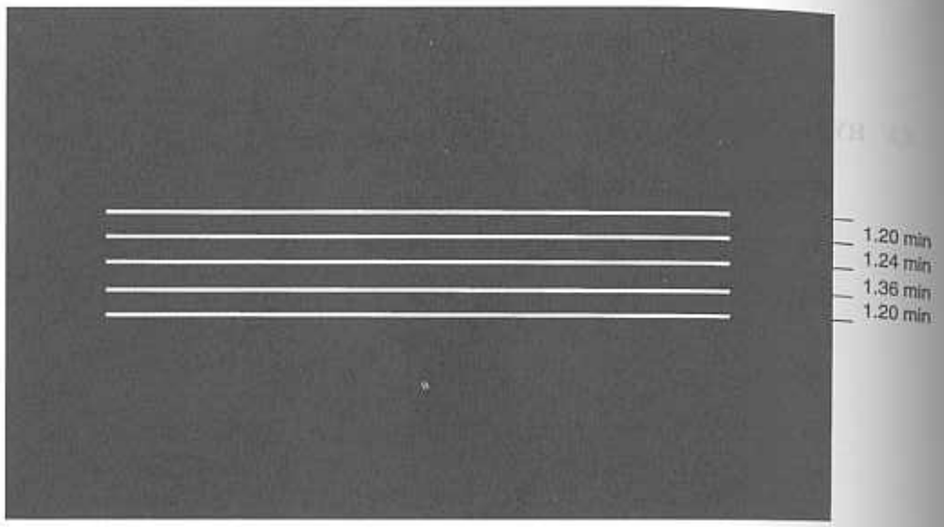



**Figure 4.11**
Hyperacuity stimuli studied by Klein and Levi (1985). The task is to decide if the middle line is displaced up or down from the center defined by the other four lines. If this figure is viewed from a distance of roughly 10 meters, then the spacings between the lines correspond to the angular displacements shown at right, and the displacement of the middle line is 0.06 minutes of arc, or 3.6 seconds of arc, up from the center. This is four times larger than the best thresholds found by Klein and Levi, which were 0.9 second of arc, or 0.00025 degrees. The angular separation of receptor cells in the fovea is ~ 30 seconds of arc, so hyperacuity in this task corresponds to the discrimination of displacements 30 times smaller than the nominal limit set by receptor sampling.

~ 0.0003 degrees (1 second of arc) for the task of centering a line between two other lines, as in Fig. 4.11 (Klein and Levi 1985).

One might imagine that there is something special about vernier patterns, since displacement information can be collected from many locations. In fact, the discriminability of these small displacements is quite robust, and can be exhibited by asking observers to detect changes in the separation of two already distinguishable dots. This whole collection of phenomena, in which we resolve displacements smaller than the receptor spacing on the retina, is termed *hyperacuity*. Many of the experiments that established hyperacuity as a general phenomenon were done by Westheimer and colleagues, and Westheimer (1981) has written a review that describes not only these experiments



but also the long historical background. At some point in the history, the idea that our brains could resolve displacements smaller than the nominal "diffraction limit" seems to have caused considerable consternation. More recently we have noticed that theoretical papers purporting to explain hyperacuity have appeared almost without reference to the fundamental physical limitations. In some sense we have gone from a time when hyperacuity seemed impossible to a time in which it seems obvious. Neither extreme view is correct.

Hyperacuity is surprising in two respects. First, the detected displacements are smaller than the spacing between photoreceptors. Second, the displacements are smaller than the "resolution" of the eye's optics. It is crucial to understand that these are only *apparent* limits, and that the real limit to spatial discrimination is set by these effects together with noise. To get the basic idea, we can think about trying to measure the position of a single dot using a linear array of photodetectors, as in Fig. 4.12. Each photoreceptor averages over some region of the visual field determined by the eye's optics and by the receptor's own geometry. In fact, photoreceptors are usually small enough that they act as optical waveguides, and this effect can give a nontrivial (and wavelength dependent) structure to the photoreceptor's integration region. All of these effects can be summarized, as in Fig. 4.12, by an empirical angular sensitivity profile $M(\phi)$—if cell number $n$ in the array is stimulated by a point light source at angular position $\phi$, then the photocurrent generated by the receptor is proportional to $M(\phi - \phi_n)$, where $\phi_n$ is the direction that this cell is "looking" in the visual field. If we know these angular sensitivity profiles, then we can construct the response of an array of cells to a single spot stimulus and ask what happens when that stimulus moves by a small amount. What we see is that, even for arbitrarily small displacements, something always changes—photoreceptors are not on/off switches, so they can respond by giving fractional changes in output when the stimulus is moved by fractions of a receptor spacing.

To tie these ideas down to concrete experiments, we show (Fig. 4.12) raw data used to measure the angular sensitivity profile of a photoreceptor in the fly's eye. Note that the cell produces graded voltage changes as the spot is moved by tiny fractions of the spacing between photoreceptors. Obviously, for very small displacements the resulting voltage changes are very small, and the question is whether these changes can be resolved above the background of voltage noise. If the signal to noise ratio is sufficient, however, it is clear that

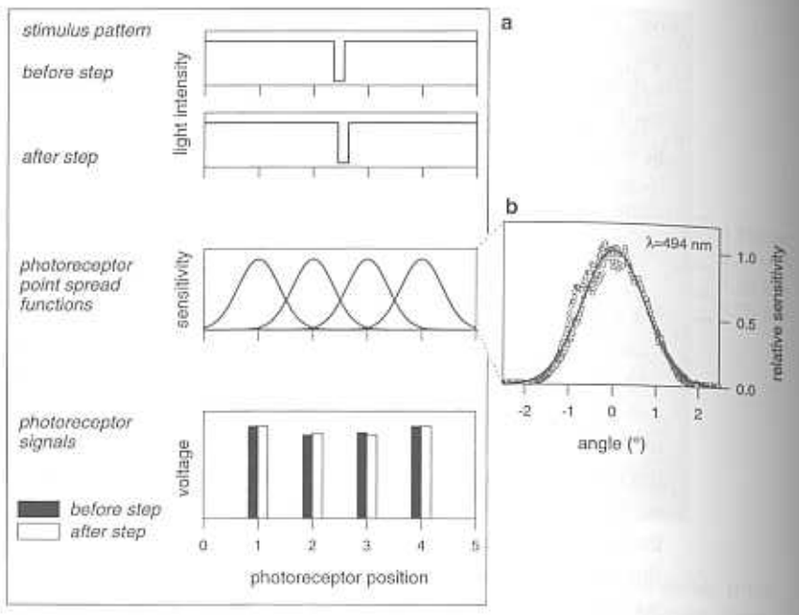



**Figure 4.12**
Spatial sampling in the fly visual system. (a) Schematic of responses of fly photoreceptor array to displacement of a point stimulus. The top two panels show the intensity pattern before and after the step. These intensity patterns are filtered by the photoreceptor point spread function to determine the photoreceptor voltages before and after the step, where for simplicity we assume the voltage to be directly proportional to the light intensity. (b) Measurement of the point spread function of fly photoreceptors, redrawn from Smakman, van Hateren, and Stavenga (1984).

one can, in principle, resolve displacements much smaller than the receptor lattice spacing.

What about the diffraction limit? Diffraction sets the width of the angular sensitivity profile, and the changes in receptor response to small displacements become much smaller if this profile is very wide. But again there is some small response even to displacements that are much smaller than the width of the sensitivity profile, which is the nominal resolution of the optical system. With sufficient signal to noise ratio in the images, we can detect these small changes, and we can improve our signal to noise by appropriately adding up signal from all the pixels at the "edge" of the object we are looking at. This is not just a matter of principle. Imaging of small displacements—much smaller



than the roughly one micron resolution of the light microscope—are a routine technique in many fields, including cell biology. Similarly, the entire image can be enhanced by "deconvolution" of the blur introduced by diffraction. Again the key question is whether one has sufficient signal relative to the noise.

What does all of this teach us about the visual system? First, we learn that the observation of hyperacuity is not fundamentally mysterious. Our friends with microscopes do essentially the same task on a daily basis, and the astronomers do the same (and more) with their telescopes. Second, diffraction and photoreceptor lattice structure by themselves do not set a limit to displacement discrimination, but these effects *together with noise* set a limit that cannot be beaten down by the brain, no matter how clever. This physical limit to displacement discrimination can be calculated quantitatively from a realistic model of the imaging system and the noise in the receptor cells.

When we first learned about hyperacuity, we were impressed that the absolute performance of the visual system should be so good. Now we know that by comparing the performance to the naive diffraction limit we were making a mistake—the diffraction limit is not a real limit, so we shouldn't really be surprised that the system surpasses it. But we have also learned that there *is* a real limit imposed by diffraction and noise together. If the system reaches this limit, then it is, in a sense, the perfect processor, making use of all the available spatial information and computing displacement with the maximum possible reliability. Hyperacuity thus gives us an opportunity to probe the reliability of neural computation: Can the visual system discriminate displacements with a reliability that reaches the fundamental physical limits imposed by diffraction and noise in the photoreceptor array, or does the central nervous system add significant amounts of noise in the process of computation? Geisler (1984) addressed this issue and gave a simple and general argument about the relation of acuity and hyperacuity.

In an acuity task we are asked to distinguish one point source from two point sources of the same total intensity. When projected through the eye's optics, the single point source produces a pattern of intensity on the retina $I_0(x)$. The two sources, separated by a distance $\ell$, produce a pattern that is just the sum of two $I_0(x)$ patterns, each of half the intensity:

$$I_2(x, \ell) = (1/2)[I_0(x - \ell/2) + I_0(x + \ell/2)]. \tag{4.28}$$

For small displacements $\ell$, the difference between the pattern from one point source and the pattern from two point sources can be found by expanding $I_2$ in a Taylor series and keeping only the leading term:

$$\Delta I(x) = I_2(x) - I_0(x) \tag{4.29}$$

$$= \frac{1}{2}\left[I_0(x - \ell/2) + I_0(x + \ell/2)\right] - I_0(x) \tag{4.30}$$

$$\approx \frac{1}{2}\left[I_0(x) - (\ell/2)\frac{\partial I_0(x)}{\partial x} + \frac{1}{2}(\ell/2)^2\frac{\partial^2 I_0(x)}{\partial x^2} + \ldots\right]$$

$$+ \frac{1}{2}\left[+I_0(x) + (\ell/2)\frac{\partial I_0(x)}{\partial x} + \frac{1}{2}(\ell/2)^2\frac{\partial^2 I_0(x)}{\partial x^2} + \ldots\right]$$

$$- I_0(x) \tag{4.31}$$

$$\approx (1/8)\ell^2\frac{\partial^2 I_0(x)}{\partial x^2}. \tag{4.32}$$

We see that the intensity changes are proportional to $\ell^2$—the signal is *second order* in the displacement. If the two sources are already separated by a distance $\ell_0$, however, a change in separation $\Delta\ell$ produces a change in the pattern of light on the retina $\Delta I \propto \Delta\ell$, so that the signal is of *first order* in the displacement. If displacements are very small, the separated sources thus give a much larger signal for the same displacement. Since the noise is always the same—random arrival of photons at the retina or other noise sources in the photoreceptor themselves—the displacement threshold should be smaller when the points whose position we are judging are already separated, and this is precisely the condition for observing hyperacuity. The crossover from acuity to hyperacuity should occur as the points (or lines) are displaced by an amount related to the point-spread function of the eye's optics, and this is in rough agreement with experiment.

Let us imagine that the effects of photon shot noise and other noise in the photoreceptors can be summarized by an effective spatial white noise added to the image. Then the signal to noise ratio for discrimination between two signals is, as explained in section A.18, an integral of the (squared) intensity differences over the entire image,

$$SNR = \frac{1}{N_0}\int d^2x[\Delta I(x)]^2, \tag{4.33}$$

where $N_0$ is the effective noise level. We see that, for acuity tasks,

$$SNR \propto \ell^4 I_0^2/N_0, \tag{4.34}$$

whereas for hyperacuity tasks,

$$SNR \propto (\Delta\ell)^2 I_0^2/N_0. \tag{4.35}$$



If photon shot noise is dominant, then the noise level is proportional to the mean intensity itself, $N_0 \propto I_0$, so at some fixed $SNR$ corresponding to the threshold for reliable discrimination we have

$$\ell_{\text{acuity}} \propto (I_0)^{-1/4},$$

$$\Delta\ell_{\text{hyperacuity}} \propto (I_0)^{-1/2}. \tag{4.36}$$

The intensity dependence of acuity and hyperacuity are thus predicted to be very different if the visual system makes optimal use of the information in the receptor array. The qualitative statement is correct even if photon shot noise is not dominant, as long as the signal to noise ratio in individual receptors is a monotonically increasing function of light intensity.

Unfortunately, the predictions of Eq. (4.36) do not provide a very good account of the data on intensity dependence of acuity and hyperacuity. On the other hand, these predictions suggest the robust qualitative effect that at *small* intensities one should find that acuity is better than "hyperacuity," essentially because $(I_0)^{1/4}$ is larger than $(I_0)^{1/2}$ at small $I_0$. Geisler and Davila (1985) set out to search for this effect and found it. McKee (1991) has emphasized that it is very difficult to account for the intensity dependence of hyperacuity in any theory, whether one focuses on physical limits, as in Geisler's discussion, or on the properties of feature detectors or "channels" in the central visual pathways that might provide additional noise sources. Clearly, it would help to know more about the signal and noise characteristics of cones (Schnapf et al. 1990), but McKee also suggests that, even when we ask a subject to perform a simple hyperacuity task, the brain may be working on the more complex problem of describing the shapes of objects in the visual environment. Theoretically, then, we should try to understand the physical limits to these apparently more cognitive tasks.

We close this section by drawing attention to an analog of hyperacuity so familiar that it passes almost without comment—color discrimination. We have three types of cone, each of which has a very broad absorption profile. This collection of three receptors samples the range of wavelengths in much the same way that the spatial array of photodetectors samples the visual field. In the case of angular displacements, the basic scales are set by diffraction, while here the analog of the "diffraction limit" is set by the width of the absorption profiles, as shown in Fig. 4.13. If we change the wavelength of a monochromatic light source, we can distinguish changes much smaller than the spacing between peaks of the absorption bands or the widths of the individual bands themselves. Even without the quantitative discrimination experiments, the fact

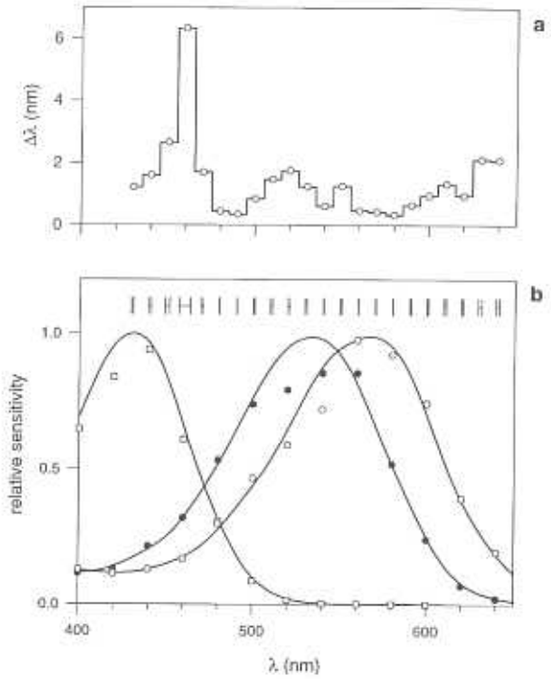



**Figure 4.13**
Wavelength discrimination (a) and photosensitivity of cone photoreceptors (b). Humans are able to match the wavelengths of two light stimuli to about 1 nm in color matching experiments, as shown in (a). However, cone photoreceptors sample different wavelengths with the rather coarse tuning curves shown in panel (b). These measurements are from red, green, and blue cones from the macaque retina, and they are fitted with standard Dartnall curves. To facilitate comparison, the trace of error bars near the top of (b) replots the $\Delta\lambda$ values in (a) on the same scale as the abscissa in (b). (a) redrawn from Mollon, Estévez, and Cavonius (1990), (b) Redrawn from Nunn, Schnapf, and Baylor (1984).

that our language provides more than three distinct color names strongly suggests that we can make discriminations beyond the limit of receptor sampling.

### 4.2.2 Experiments with single neurons

The difficulties of comparing human behavior with the physical limits suggest that we look for systems where hyperacuity is observed in the responses of individual neurons. Until the mid–1980s there were no such systems. In 1984, preliminary results were reported from an experiment in the fly visual system



(de Ruyter van Steveninck, Bialek and Zaagman 1984), demonstrating that observing just a few spikes from a single motion sensitive neuron is sufficient to discriminate reliably between displacements that differ by roughly one tenth of the spacing between photoreceptors in the compound eye. The performance of this system will be explored in detail in the following sections. At about the same time, three groups reported performance in the hyperacuity regime for single neurons in the mammalian visual system (Parker and Hawken 1985; Shapley and Victor 1986; Swindale and Cynader 1986).

Working in the primary visual cortex of monkeys, Parker and Hawken (1985) studied displacement discrimination by measuring the probability that a cell fires in response to presentation of a pattern in a given location. They found that this probability varies extremely rapidly with location, so that a simple spike/no spike criterion is sufficient to discriminate positions that differ by much less than the nominal width of the cell's receptive field. In the best cases, the threshold for reliable discrimination corresponds to a displacement of 11 seconds of arc, which is essentially equal to the human threshold for this task at comparable retinal eccentricities.

The results of Parker and Hawken provide one more example of a system where the reliability of signaling by one neuron, in this case one spike from one neuron, approaches the reliability of the whole organism. In addition, we can give this experiment an information theoretic interpretation. When we have enough information to choose correctly between two alternatives, we have exactly one bit. At the threshold for reliable discrimination, one makes errors 25% of the time (by convention), and this noise reduces the information transmission to 0.19 bits. But this information is carried by on average, half a spike, so the information transmission rate is 0.38 bits per spike.

Recently Lee et al. (1993) have returned to the problem of hyperacuity in monkey vision, this time recording from retinal ganglion cells, as Shapley and Victor (1986) had done in cats. Lee et al. studied ganglion cell responses to the sudden displacement of an edge, with displacements in the range of 0.5–4 minutes of arc. Psychophysical thresholds at the relevant retinal eccentricities are 1 arc min or less for edges with contrast greater than 20%. For magnocellular ganglion cells, analysis of post-stimulus time histograms demonstrates that the step motion produces a transient response of the cell, with *peak* changes in firing rate of 23 spikes per second per min of arc at 20% contrast. But these rate changes decay with a time constant of less than 40 ms, so that detectable displacements produce somewhat less than one extra spike in one ganglion cell. Nonetheless a single extra impulse would be detectable, because the variance of the maintained discharge (in the absence of a step displacement) is less than one spike in a 40 ms window. Parvocellular ganglion cells are more numerous

but also more than an order of magnitude less sensitive for this displacement detection task, and Lee et al. argue that performance at this task is therefore determined by the magnocellular pathway.

The comparison of human hyperacuity performance with the physical limit may be confounded by cognitive factors, but this is not the case for ganglion cells. It should be possible to compare ganglion cell performance in the Lee et al. experiment to cone noise levels as measured by Schnapf et al. (1990). More generally, we know from Barlow, Levick, and Yoon (1971) and from Aho et al. (1988) that, under dark adapted conditions, the "noise" at the ganglion cell reflects random photon arrivals and dark noise in the rod photoreceptors, not some excess noise contributed by the retinal circuitry. One would like to make a similar comparison under daylight, cone dominated conditions.

The notion that reliable responses in the hyperacuity regime are carried by single spikes is, at first sight, surprising. We emphasize, however, that this same result emerges from many different experiments: in the fly motion sensitive neurons (de Ruyter van Steveninck, Bialek, and Zaagman 1984; de Ruyter van Steveninck and Bialek 1992, 1995), as described in detail below, in the monkey primary visual cortex (Parker and Hawken 1985) and retinal ganglion cells (Lee et al. 1993), and, most recently, in the directionally selective ganglion cells of the rabbit retina (Grzywacz, Amthor, and Merwine 1994).

As Shapley and Victor (1986) emphasized, some combination of retinal ganglion cells *must* provide the information that makes possible performance in the hyperacuity regime. For the particular task they consider, Lee and coworkers argue from the anatomy of the monkey retina and the sizes of receptive fields that at most ten ganglion cells are really "tuned" to the task, giving near maximal responses to the step. This certainly makes it plausible that thresholds for discrimination based on spike trains of single cells will be close to thresholds for the whole organism, and we could have guessed that this would be true even before the experiment of Lee et al.. The surprise is that the total number of spikes involved is so small. It would seem that the total number of extra spikes at psychophysical threshold is of order ten or less, and Lee et al. emphasize that either correlation among neighboring neurons' spike counts or uncertainty on the part of the observer about which cells to "listen" to in making the decision (and when to listen to them) could reduce still further the effective number of neurons contributing to the perceptual decision. One might be tempted to draw the analogy with photon counting, where Lorentz' original calculation suggested that the visual system gives reliable responses to 10–100 photons, and several generations of experiments pushed this number down to one photon. The results of Lee et al. should inspire the search for



stimulus conditions in which human observers detect not ten, but perhaps just one extra spike in the entire optic nerve.

### 4.2.3 Temporal hyperacuity

It is perhaps natural that, as highly visual creatures, we are drawn to the investigation of the visual system in humans and in other animals. But we must remember that many creatures view the world through rather different senses. Many animals navigate guided primarily by odors, and we know of visual systems, such as the bee's, that are sensitive to different regions of the electromagnetic spectrum. But in some way these are extensions or extrapolations of our own sensory experiences. There are at least two systems for imaging the environment that have no counterpart among our own senses: echolocation and electrolocation. In both cases, the synthesis of a spatial image depends in a clear and obvious way on the processing of information in the time domain, and the evolution of precise imaging in these systems has given us spectacular evidence concerning the temporal precision of the nervous system. In particular, the echolocating bats *Eptesicus fuscus* can discriminate echo delay differences as small as 10–50 nanoseconds (Simmons 1989; Simmons et al. 1990), and the weakly electric fish *Eigenmannia* responds to 100 nanosecond shifts of phase in an oscillating electric field (Rose and Heiligenberg 1985). We focus here on the echolocation problem.

A schematic of the bat behavioral experiments is shown in Fig. 4.14. Briefly, bats stand at the foot of a **Y** and will have to decide which arm of the **Y** carries the signal. Each arm has a microphone and loudspeaker to produce synthetic echoes of the bat's ultrasonic pulses. On one side, the echo is always given with fixed delay, and on the other side the delay changes from pulse to pulse by $\pm\delta\tau$. The signal is this jitter of the synthetic target. The bat's ultrasonic pulse has a width of several milliseconds, and one might naively suppose that this sets a basic time scale for discrimination of arrival times. But sound travels roughly one foot in a millisecond, and so coarse an image of the world would seem of little use to the bat.

One can do a rather elegant, if elementary, experiment to get a feeling for the precision of bat echolocation (Trappe 1982; reviewed in Simmons 1989). If one tosses a small mealworm into the air, a bat will catch the worm and eat it, bringing its wings under the target like a scoop. Stroboscopic films of this maneuver demonstrate that the bat has its head locked onto the target very rapidly, perhaps in response to just one echolocation pulse. But if one dips the mealworm in flour one can find the dust mark on the wings, so that we know where the target was intercepted. Repeating the experiment many times

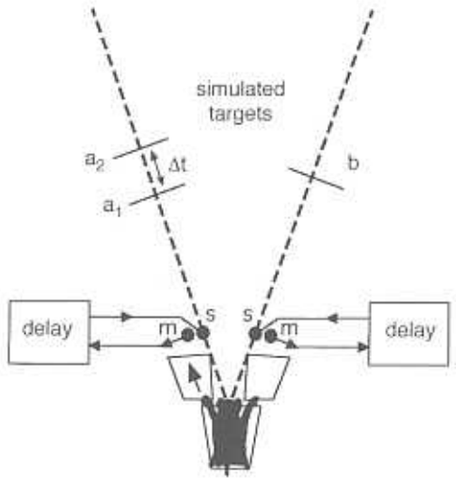



**Figure 4.14**
Experimental setup for bat behavioral studies. The bat is trained to land on a Y-shaped platform. Along each arm of the y are a microphone and a speaker. The bat's calls are picked up by the microphone, and then delayed and played back through the speaker. In one direction the delay is fixed. In the other direction the delay alternates between each call. The bat's task is to step along the platform in the direction of the moving target. Redrawn from Simmons (1989).

should leave a dust mark whose size measures the variance in the location of the interception point and hence (roughly!) the variance of the bat's estimate of the target position. The result of this experiment is a dust mark of order one centimeter across, which, given the speed of sound, corresponds to $\sim 35$ microsecond accuracy in the measurement of pulse arrival times. It isn't clear that the bat is interested in maximizing the precision with which the worm hits its wing, but the fact that the bat *can* reach this precision tells us that its measurements must be at least this accurate. The 35 microsecond time scale matches not the width of the pulse but rather the fundamental period of the underlying sound pressure waveform, which is a combination of two or three harmonics with a fundamental frequency that is swept downward through the pulse.

The experiments by Simmons (1979) demonstrate that bats reliably discriminate echo jitters *below* 30 $\mu$s, down to 1 $\mu$s or perhaps 500 nanoseconds. More surprisingly, bats that achieve essentially perfect discrimination of 10 $\mu$s jitter become confused (errors increase) when the jitter is increased, and



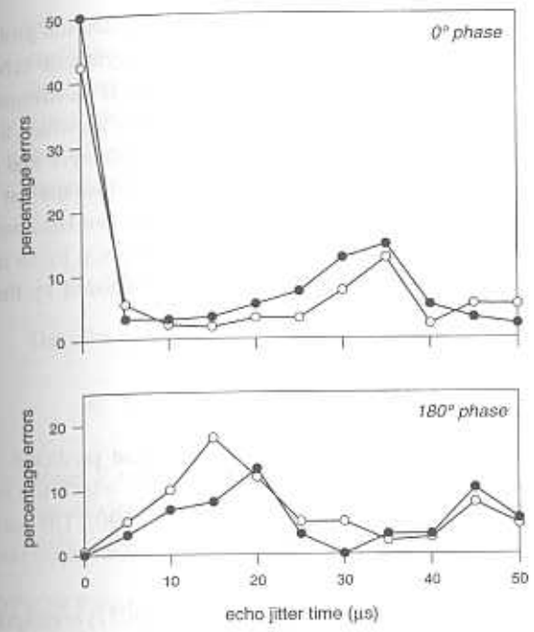



**Figure 4.15**
Performance of two bats (open and closed circles) detecting jittered echoes. In the top panel, synthetic echoes are returned to the bat with no phase shift (the 0° condition) and discrimination must be based on the temporal displacement or jitter alone. In the bottom panel, the bat discriminates between echoes that differ both by a temporal jitter and by a 180° phase shift. In the 0° phase condition the bats make errors at zero jitter, as they must, but also around 35 $\mu$s, close to the period of the echolocation waveform. The 180° condition data indicate that the bat can discriminate a phase reversal alone (at zero jitter), and that confusion arises at delays of 15–20 $\mu$s and also around 45 $\mu$s; again these regions of confusion are separated (roughly) by the period of the waveform. Adapted from Simmons (1989).

the confusion peaks at a jitter $\delta\tau$ that brings the sound pressure waveform of the pulse back into phase with itself, as shown in Fig. 4.15. Put differently, errors are maximal at delay differences that correspond to peaks in the autocorrelation function of the sound signal, as if the bat makes measurements with reference to a perfect copy of the signal waveform—including details at frequencies of order 30 kHz.

Suppose the bat really does have a perfect reference signal, so it is listening for a *known* pulse shape $s_0(t-\tau)$ in a background of noise $\eta(t)$. The noise

might arise within the auditory system itself, but, to make the problem well posed, one can add a background of white noise to the synthetic echoes in the behavioral experiment, and then manipulate this noise level (Simmons et al. 1990). This noise sets a physical limit to the reliability with which the bat can detect the difference between a reference value of the delay $\tau$ and a slightly different value $\tau + \delta\tau$, and we would like to see how close the bat comes to this limit. As explained in section A.18 (see also Menne and Hackbarth 1986), the signal to noise ratio for this discrimination task is given by an integral of the (squared) time derivative of the pulse shape, normalized by the spectral density $N_0$ of the added noise,

$$SNR = (\delta\tau)^2 \frac{1}{N_0} \int dt \left[ \frac{ds_0(t-\tau)}{dt} \right]^2 . \tag{4.37}$$

The performance of the bat is essentially equal to that predicted from this optimal signal to noise ratio, down to noise levels for which the limit (corresponding to $SNR = 1$) is 40–50 ns (Simmons et al. 1990). The performance of the bat at still smaller noise levels plateaus at a discrimination threshold of about 10 ns, as shown in Fig. 4.16.

In what sense does performance at the level of Eq. (4.37) correspond to hyperacuity (Altes 1989)? In vision, when we view a point source of light, the

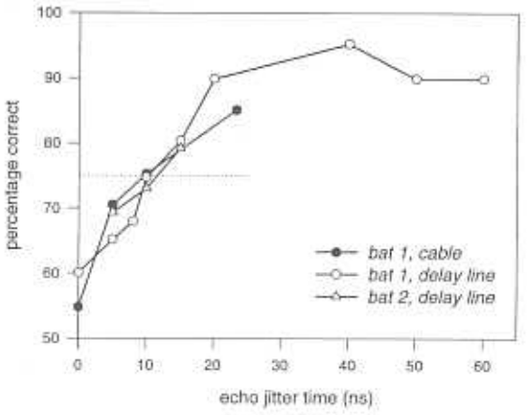



**Figure 4.16**
Performance of two bats in the task described in Fig. 4.14. The percentage of correct decisions is plotted against the time jitter $\Delta t$. Adopting a criterion of 75% correct, bats are able to discriminate a time jitter of 10–12 ns. Adapted from Simmons (1989).



image on our retina has a width determined by the diffraction pattern of the pupil–lens system; hyperacuity corresponds to seeing beyond this diffraction limit, and is possible only at sufficiently large $SNR$. In the time domain relevant to the bat, the basic scale of the "image" is set by the time dependence of the pulse waveform $s_0(t)$ itself. We can define this time scale $\Delta T$ in terms of the time derivatives of the echo pulse,

$$\frac{1}{(\Delta T)^2} = \left\{ \int dt \left[ \frac{ds_0(t)}{dt} \right]^2 \right\} \left\{ \int dt \left[ s_0(t) \right]^2 \right\}^{-1} . \tag{4.38}$$

Then the threshold for reliable discrimination of target jitter is

$$\delta\tau \sim \frac{\Delta T}{\sqrt{E/N_0}}, \tag{4.39}$$

where $E$ is the total "energy" of the pulse $E = \int dt \left[ s_0(t) \right]^2$. As in the case of spatial vision, the limit to discrimination $\delta\tau$ is smaller than the basic physical scale $\Delta T$ if the signal to noise ratio of the pulse, $E/N_0$, is sufficiently large.

## 4.3 MOTION PROCESSING IN THE FLY VISUAL SYSTEM

In a series of experiments on the blowfly *Calliphora vicina*, it has been possible to demonstrate that an identified motion sensitive neuron in the visual system encodes the trajectory of rigid motions with a precision well within the hyperacuity regime. Indeed, this precision approaches the limits imposed by noise in the photoreceptor array, and this noise is in turn dominated by photon shot noise. In a very direct sense, the precision of movement computation in this system is thus limited by physical considerations.

### 4.3.1 Limits to discrimination

We begin the discussion of reliability in fly vision with experiments that emulate the classic discrimination paradigm of human psychophysics. These experiments are generalizations of the Barlow–Levick (1969) experiments described in section 4.1.2, and they lead to a precise description of neural reliability in a highly restricted task (de Ruyter van Steveninck, Bialek, and Zaagman 1984; de Ruyter van Steveninck, 1986; de Ruyter van Steveninck and Bialek 1992, 1995). Then we will see how stimulus reconstruction experiments demonstrate the same precision in a real-time estimation task that is much closer to the problem the fly is solving in natural flight (Bialek et al.

1991; Rieke et al. 1996). Before embarking on the study of neural responses, however, we would like to understand the limits to reliability imposed by the properties of the photoreceptors. To do this we will put ourselves in the position of the fly, looking only at the receptor voltages and trying to decide among different possible stimuli.

The compound eye of *Calliphora* samples the world quite coarsely; the photoreceptors form a hexagonal lattice with a horizontal spacing of 1.35°. What does the fly actually "see" when we displace a pattern by a small fraction of the photoreceptor spacing? From experience in the laboratory we know that the effective noise level for an image (or any measurement) is determined by the available averaging time. Clearly, if the fly can average forever, it could in principle resolve arbitrarily small displacements. But for an insect flying at several meters per second, forever is a very long time. Even one second is too long a time to wait before adjusting one's flight path to stabilize straight flight or to turn in response to interesting stimuli in the environment. As we have discussed in section 2.2.1, flies can produce flight torques within 30 ms of a visual stimulus.

Let us recall that 30 ms is the duration of one video frame—in a time where we cannot see any variations at all, the fly can complete the computation of motion and begin its motor response. This time is short even for the fly, however, since it takes roughly 10 ms for the photoreceptors to reach their peak response, and the motion sensitive neurons are several synapses removed from the sensory periphery. Even if we deliver a very large motion stimulus in a high contrast environment, the motion sensitive cells have a roughly 15 ms latency. The fly must thus be able to compute motion based on a 5–15 ms average of the photoreceptor outputs. But the noise in the photoreceptor voltage has a correlation time of 8–10 ms, so that averaging over 10 ms does not "average away" the noise at all. In effect, the fly is capable of judging motion based on two successive snapshots of the receptor voltage array.

Under the conditions of the experiments on the fly's H1 neuron, the standard deviation of the voltage noise in the photoreceptors is 0.49 mV. The sensitivity of the cell is ~ 3 mV/unit contrast, but the experimental stimuli have a contrast of 0.16 when viewed through the photoreceptor aperture. Thus the signal to noise ratio is about 1 in each cell. In Fig. 4.17 we show simulated snapshots of the photoreceptor voltage array at this signal to noise ratio. We emphasize that these images are generated using parameters actually measured for the fly's photoreceptor sensitivity, noise, and blur under exactly the same conditions as the H1 experiments (de Ruyter van Steveninck 1986).



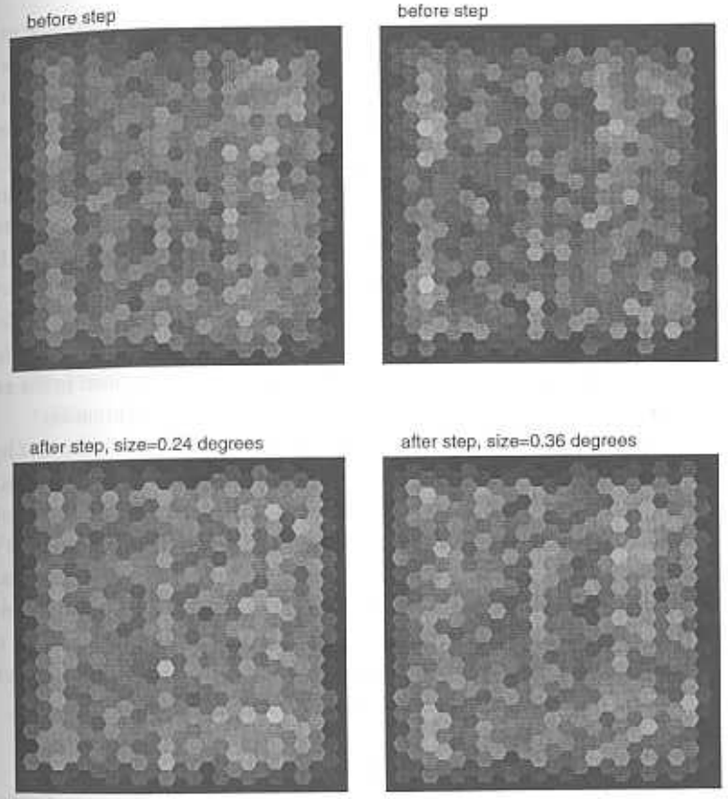



**Figure 4.17**
Schematic of response of the fly photoreceptor array to a displacement step of a random bar pattern. The voltage in each photoreceptor, represented by a gray level, is the sum of two terms—the average voltage, related to the stimulus, and a random deviation or noise. The average photoreceptor voltage at each position is computed by filtering the bar pattern through the measured point spread function of the photoreceptors (see Fig. 4.12b). The statistics of voltage noise in the photoreceptor have been characterized experimentally (see Fig. 3.12); for the conditions relevant here the noise has a 0.72 mV standard deviation, and so the random voltage deviations in each pixel are drawn independently from a Gaussian distribution with a 0.72 mV standard deviation. The two images on the left are two snapshots of the photoreceptor outputs before and after a 0.24° step (to the right) of the visual pattern; on the right are the responses to a 0.36° step. We assume the snapshots are separated by more than 5 ms so that the noise voltages in different panels are uncorrelated. Forced choice discrimination experiments on H1 show that this single neuron is able to discriminate between these two steps with a reliability of 66%.

Figure 4.17 makes it clear that *on the time scales of relevance to fly behavior, the sensory input is very noisy.* The noisiness becomes even more apparent when we look at motion, that is, at two successive snapshots that differ by a small displacement of the stimulus pattern. The experiments on step discrimination in H1 are in the hyperacuity regime, where we ask if the output of H1 is sufficient to distinguish between displacements of 0.18 and 0.27 receptor spacings (0.24° and 0.36°, respectively). Voltage arrays in response to these displacements are shown in the different panels of Fig. 4.17. It is clearly very hard to distinguish which is the larger displacement, yet we will see that the fly extracts enough information from these noisy signals to allow reliable discrimination. The key to doing this, of course, is that H1 can integrate motion signals from the 2,500 photoreceptors being stimulated in the experiment. How well could the fly do by integrating all of this information?

The images in Fig. 4.17 provide strong qualitative evidence that hyperacuity in motion estimation will be hard. To quantify this impression and calculate the real limit to movement discrimination requires a precise mathematical formulation of the problem (de Ruyter van Steveninck and Bialek 1992, 1995; Bialek 1992). The essential conclusion is that, on the 30 ms time scale relevant to behavior, movement discrimination should reach a signal to noise ratio of unity for a displacement difference of 0.06°. To see whether H1 can in fact approach this limit, we need to measure discrimination performance based on single examples of the spike train in response to movement steps.

### 4.3.2 Discrimination experiments with H1

To analyze the reliability of discrimination based on the spike train of a single neuron we generalize the approach of Barlow and Levick (1969). Rather than counting the spikes produced by the cell, we try to give a complete and model independent description of the spike train, keeping track of the arrival times of individual spikes. This is feasible because we know that the fly responds quickly, so that small numbers of spikes are relevant, and because the fly preparation is sufficiently stable that one can present each stimulus more than ten thousand times. These long recording times imply that one can collect very complete statistical information about the spikes in response to particular stimuli (de Ruyter van Steveninck 1986; de Ruyter van Steveninck and Bialek 1992, 1995).

The most general approach is to divide time into discrete bins, then describe each spike train as a sequence of 1's and 0's representing the presence or absence of a spike in each bin. As explained in the discussion of spike train entropy (section 3.1.2), this associates each spike train with a binary number



or "word" $Q$. If we could make very small bins and very long binary words, this would converge to an exact description of the neural response. In practice one is limited to rather short words, since the number of different possible words grows exponentially as we include more digits. In the fly, the first 15 ms following a step displacement is just latency, and after 40 ms the fly will have made a decision. There are enough data ($\sim 10^4$ presentations) to analyze up to 13 digit binary words, so one possibility is to choose a 2 ms bin size and cover the time window from 15 to 41 ms. Interval distributions are smooth on the 2 ms time scale, so this bin size seems reasonable, and one can check that (over a smaller time window) smaller bins give identical results. Thus it is possible to give a complete description of the firing patterns over the full window for behavioral decision making, limited only by the 2 ms bin size.

The step discrimination task (Fig. 4.18) is to decide which of two possible motions occurred. To quantify the performance of the fly, we make this decision using only a single example of the spike train. To be precise, we observe a particular spike train, described by the binary number $Q$, and we must decide whether this arose from the step of size $\alpha_1$ or the step of size $\alpha_2$. As we have mentioned, and as explained in detail in section A.16, the strategy that will give the maximum fraction of correct decisions is maximum likelihood. In the case that the two stimuli are shown with equal probability, the decision rule is that, having observed $Q$, we choose $\alpha_1$ if $P(Q|\alpha_1) > P(Q|\alpha_2)$, and vice versa. On average, the probability of correctly identifying step $\alpha_1$ is then

$$P_c(\alpha_1) = \sum_{\{Q\}} P(Q|\alpha_1) \cdot H[P(Q|\alpha_1) - P(Q|\alpha_2)], \tag{4.40}$$

where the summation is over the set of all possible neural firing patterns $\{Q\}$. The Heaviside step function $H[x]$ is defined by

$$H[x < 0] = 0 \tag{4.41}$$

$$H[x > 0] = 1. \tag{4.42}$$

An interchange of indices 1 and 2 in Eq. (4.40) yields the formula for correct identification of $\alpha_2$. The proportion of correct judgments in the entire experiment is then simply $P_c(\alpha_1, \alpha_2) = [P_c(\alpha_1) + P_c(\alpha_2)]/2$, which from now on will be referred to as $P_c$.

It is convenient to think about the fraction of correct discriminations, $P_c$, in terms of an equivalent "discriminability parameter" $d'$.[2] The basic idea (Green

2. Note that what we have called the signal to noise ratio in previous sections is the square of the discriminability parameter, that is $SNR = (d')^2$.

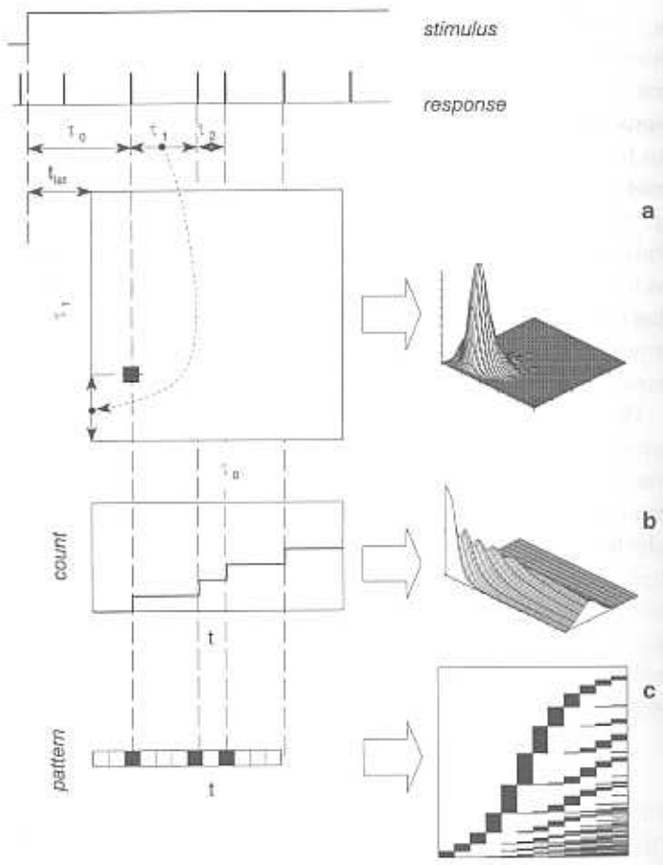



**Figure 4.18**
Schematic of step discrimination task. The fly observes a random spatial pattern. At $t = 0$ the pattern is stepped either 0.24° or 0.36° while the response of H1 is monitored. We can ask, as described in the text, with what reliability an observer of the H1 spike trains can discriminate between these two steps. This requires setting up distributions of response. Three examples of such distributions are shown. In all of these we exclude a latency time $t_{lat}$=15 ms from analysis. In (a) a histogram is formed of the timing of the first two spikes following the stimulus, and in (b) we construct histograms of the total count in increasing time windows following the stimulus. In (c) we regard the spike train as a binary sequence at a certain time resolution (2 ms in this case), and we characterize the probability of finding each possible binary string (Fig. 4.20). Of course H1 uses the signals contained in the photoreceptor responses (see Fig. 4.17) to report wide field movement of the visual scene. From the characteristics of those responses we can compute the reliability of movement discrimination of an ideal observer of the photoreceptor signals. This can then be compared to H1's measured performance.



and Swets 1966) is to consider the idealized problem of detecting a fixed signal of amplitude $A$ in a background of Gaussian noise with standard deviation $\sigma$. The probability of correct detection must be a function only of the ratio $d' = A/\sigma$, and from Fig. 4.19 we see that this probability is simply related to the area under the Gaussians. Thus, if we define a function

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} dz \exp(-z^2/2), \tag{4.43}$$

the probability of correct discrimination is

$$P_c(d') = \Phi(d'), \tag{4.44}$$

and we note that $P_c(d' = 0) = 1/2$ and $P_c(d' \to \infty) = 1$, as we expect. We can use Eq. (4.44) to calculate the probability of correct discrimination, but we can also use it to translate an observed $P_c$ into an equivalent signal to noise ratio. In what follows, we describe the performance of H1 in terms of $d'$, which is directly comparable to the signal to noise ratio in the optimal discriminator.

How do we think about the conditional probabilities $P(Q|\alpha)$? Experimentally, the conditional probability of firing patterns $P(Q|\alpha)$ is determined by counting the number of occurrences of each possible $Q$ for a large number of presentations of stimulus $\alpha$. The resulting family of probabilities is described conveniently by a tree branching at each time bin, as in Fig. 4.20. Such a tree is constructed by ordering all recorded firing patterns according to the binary number represented by $Q$. At each node, the probability splits into two parts. Each of these parts describes the probability of occurrence or nonoccurrence of a spike in time bin $k$, given a history of spike occurrences specified by all the ancestral nodes up to the root. Notice that we can stop this procedure at any point, having characterized the spike trains only up to some finite time after the step. If we analyze discriminability under these conditions, we can see how discrimination power improves with time.

Firing pattern distributions for three step sizes are represented by the trees shown in Fig. 4.21. Given these probability trees, we can calculate the probability of correct discrimination according to Eq. (4.40), and, as mentioned earlier, we can do this for each possible value of the time window following the step. The results are summarized in Fig. 4.22 as a plot of $d'$ versus time for discrimination between various pairs of steps. As a check, the performance was computed for 0.5 ms and for 1 ms bins as well, and the results for the first 6.5 ms and 13 ms, respectively, were essentially the same as for the case with 2 ms time bins, justifying the choice of bin size.

At first sight it may seem contradictory to construct a continuous measure of discriminability from the spike train, in which information is encoded in the

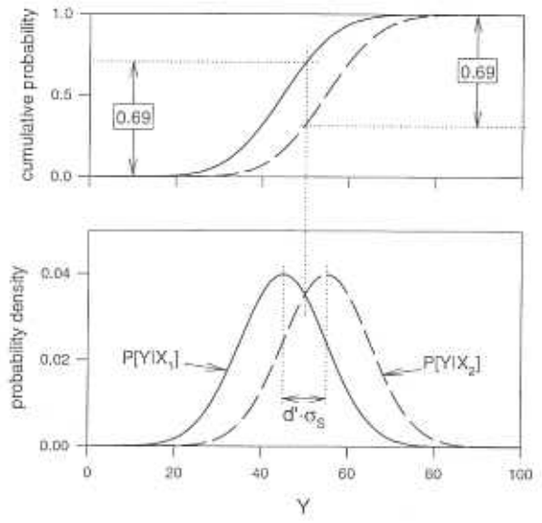



**Figure 4.19**
Definition of $d'$. Suppose we are asked to infer the value of an input variable $X$ from observation of an output variable $Y$. In the case we consider here, $X$ can have either one of two values, $X_1$, or $X_2$. If the probability distribution of $Y$ given $X_1$ differs from that of $Y$ given $X_2$, then we can in principle make an inference better than chance. A simple case is illustrated in the figure. Here, $P[Y|X_1]$ and $P[Y|X_2]$ are both Gaussian distributions, the first with mean 45, the second with mean 55, and both with standard deviation 10. If we observe a value of $Y < 50$ then we reason that our best guess is that this observation was caused by $X_1$, as can be seen from the figure. (Where we also assume that, a priori, the chances for $X_1$ and $X_2$ are equal). Conversely, for $Y > 50$ we would infer $X_2$ to be the input. We would like to know how well we do on average when we make repeated estimates. This can be read off from the cumulative distributions on top. If $X_1$ is the input, then in 69% of the cases $Y$ will be less than 50, in which case we draw the correct conclusion. In 31% of the cases will we be wrong, because of the actual value of $Y$ being greater than 50. Conversely, if $X_2$ was the input then $Y$ will be greater than 50 in 69% of the cases. Thus with the decision rule of choosing $X_1$ ($X_2$) for $Y < 50$ ($Y > 50$) we will be right 69% of the time and wrong 31% of the time. Here the distributions are made Gaussian, and the distance between their maxima is equal to their standard deviation. This corresponds to the conventional criterion for discriminability $d' = 1$. Strictly speaking, for nonGaussian distributions, $d'$ is not defined. However, from experimentally determined probability distributions and a decision rule one can always compute the probability of responding correctly. From this on can then get an equivalent value of $d'$.



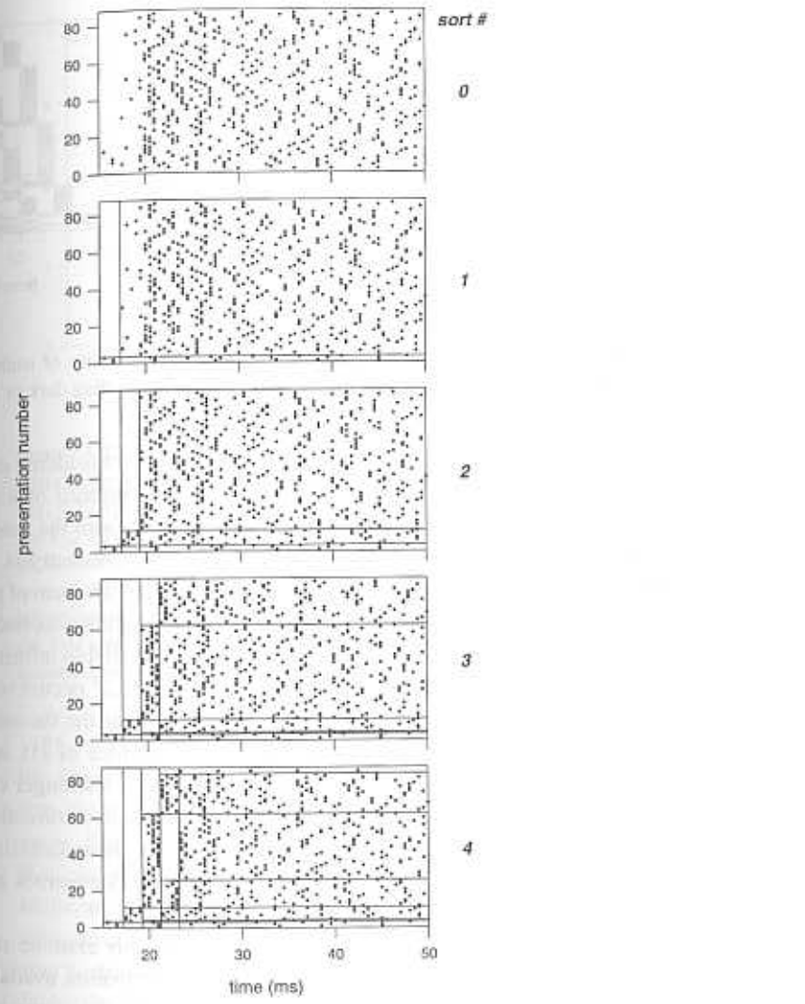



**Figure 4.20**
Construction of probability distributions of firing patterns, or trees, for discrimination experiments on H1. Raster plots of the response of H1 to repetitions of the same step stimulus are shown (panel 0). These raw data are partitioned in a subset with a spike in the first bin, and one without a spike in the first bin (panel 1). Each of these two subsets is partitioned separately, according to the presence or absence of a spike in the second bin (panel 2). This bipartitioning is shown here for four consecutive partitions. The probability of a particular pattern is given by the number of events in the corresponding partition, normalized by the total number of presentations.

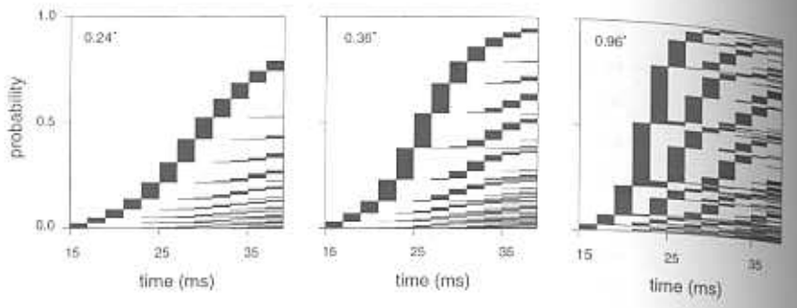



**Figure 4.21**
Distribution of firing patterns for three step sizes. The probability of each pattern of spikes after N time bins is given by the height of the corresponding dark or light bar.

form of discrete events. The neuron's response can be considered continuous because not only the occurrence, but also the nonoccurrence of a spike in a certain time window carries information, as was clear from the discussion of response-conditional ensembles in section 2.2.3. This symmetry is expressed in a natural way when the firing pattern is represented in the form of a tree. The discriminability $d'(t)$ is a monotonically nondecreasing function, because each time bin of the firing pattern distribution adds to the available information.

Discrimination between steps that differ by $\Delta\theta = 0.12^\circ$ occurs with $d' \sim 1$ within the 40 ms window of this analysis. We recall that the theoretical limit to discrimination is $\Delta\theta \sim 0.06^\circ$, so that the performance of H1 is within a factor of two of the physical limit. In fact, one can draw a stronger conclusion by analyzing the time dependence of $d'(t)$: If we include a correction for the latency of H1's response, the observed $d'(t)$ tracks the theoretical limit within a factor of two up to 30 ms after the step (de Ruyter van Steveninck and Bialek 1995), as seen in Fig. 4.25.

Which features of the spike train make possible this extreme reliability? We emphasize that this analysis collects *all* the information available in the spike sequence in a certain observation window, up to the approximation of measuring spike times in 2 ms bins. Starting from this complete description of the neural response, we can discard different aspects of the spike sequence and ask how well we can discriminate based on this reduced description of the neural response. In the extreme, we discard all timing information and count spikes in a window of size $T$ following the step. This brings us back to the Barlow–Levick experiment, but with a variable time window, and the results are shown in Fig. 4.22 together with the more complete analysis of firing patterns.



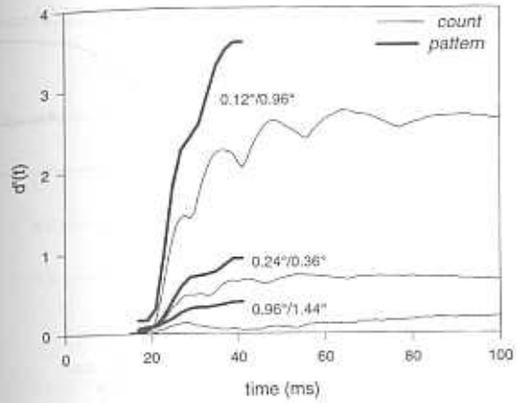



**Figure 4.22**
Discriminability versus time computed from firing pattern distributions like those in Fig. 4.21 (thick lines), or from time dependent spike-count distributions (thin lines), see Fig 4.18. Each trace shows the value of $d'(t)$ for a pair of step sizes as a function of the time window in which the spikes are observed. The 15 ms delay before $d'$ begins to increase is due to delays in signal transmission. The figure shows that discrimination based on counting spikes is roughly equivalent to that based on firing patterns, up to the first spike. After that, the spike count distribution is much less informative than the pattern distribution.

There are two key points about discrimination based on spike counts. First, this discrimination is always worse than if one keeps track of all the information in the spike train, as it must be, but the discrepancy is negligible at short times. Second, discrimination saturates near the physical limits with windows that contain on average one spike or less. This parallels the results of Zohary, Hillman, and Hochstein (1990) on the orientation discrimination of simple cells from area V1 of the monkey visual cortex. They found that discrimination performance based on spike count saturates rapidly with time windows larger than 60 ms, but that these windows contain an average of one spike.

If optimal discrimination occurs for windows that contain just one spike, it is misleading to say that discrimination is based on counting alone. "Counting" zero or one spike in a window of carefully chosen size is really equivalent to logging the arrival time of the first spike and deciding (in the context of step discrimination) that if a spike has not occurred by a particular time, then we must be looking at the smaller step size. This can be made precise by plotting probability distributions for the first spike arrival time, or the "zeroth interval,"

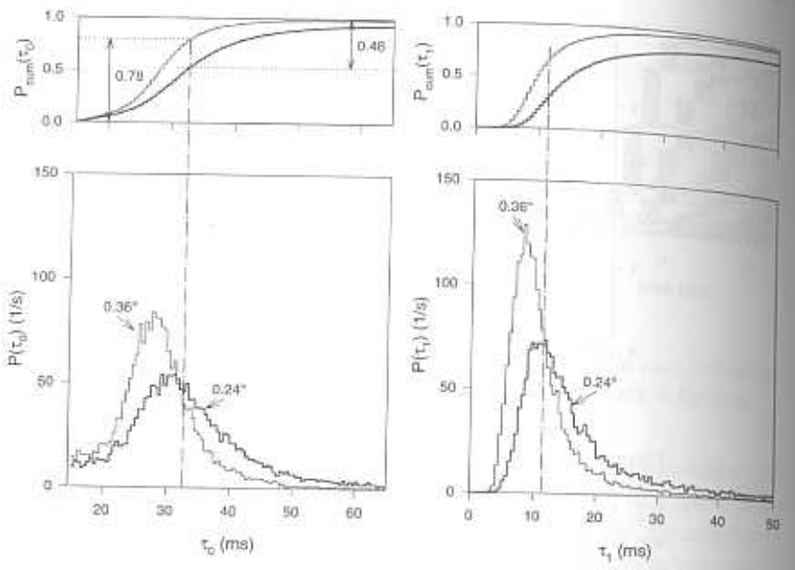



**Figure 4.23**
Distributions of spike intervals for 0.24° and 0.36° steps. $\tau_0$ is the interval from stimulus presentation to the first spike fired after 15 ms, and $\tau_1$ is the interval between that spike and the next one. The probability densities are shown below and the cumulative probabilities are shown on top. If one had to decide on a step size of either 0.24° or 0.36°, each being equally likely a priori, then based on the arrival time of the first spike, the best criterion (see text) would be to choose 0.24° when $\tau_0 > 32$ ms, and 0.36° otherwise. In that case the proportion of correct identifications can be read off from the cumulative distributions, as indicated by the dot-dash lines: If a 0.36° step was presented, correct identification will be made in 78% of the cases. For a 0.24° step, correct identification occurs in a fraction $(1 - 0.48) = 0.52\%$ of the cases, and the mean correct performance is 65%.

as shown in Fig. 4.23. Clearly, a large fraction of the available discrimination power resides in this one timing measurement. Discrimination based on the time from the first spike to the second spike (the first interspike interval) is almost as good, and discrimination with both intervals improves as if the two intervals carry independent information, as shown in Fig. 4.24.

We can make the point about the significance of the first spike more clearly by comparing neural discrimination performance with the physical limit. In Fig. 4.25 we see that discrimination based on the first spike arrival time tracks the discriminability based on the full spike pattern over the entire time interval in which neural performance tracks the physical limit. Thus the fly's visual



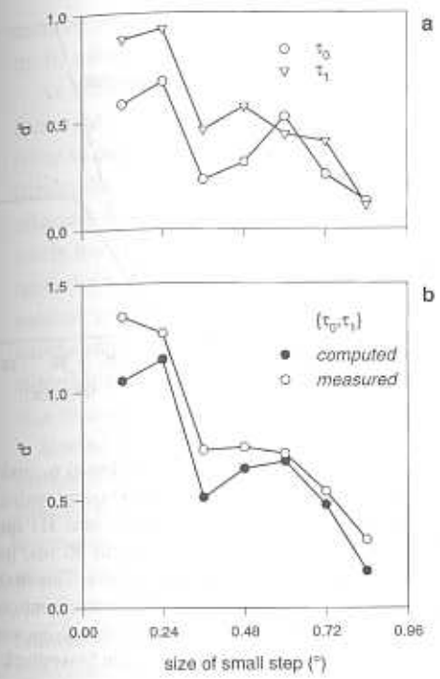



**Figure 4.24**
Discrimination performance for single and double intervals. (a) Values of $d'$ calculated from probability distributions such as those in Fig. 4.23. These values are for discrimination between a small step (size given by the abscissa) and a step that is 0.12° larger than the small one. (b) Values of $d'$ based on the combination $\{\tau_0, \tau_1\}$. These results are compared with the performance computed from $\tau_0$ and $\tau_1$ assuming these two intervals contribute independent information for the discrimination task. The agreement is reasonably good, indicating that the redundancy between successive intervals is minimal, at least under these conditions.

system extracts a motion signal with a reliability that approaches the limit imposed by noise in the photoreceptor cells, and, furthermore, this signal can be recovered by timing the occurrence of a single spike in one neuron.

### 4.3.3 Continuous estimation

The discussion of the step discrimination experiments gives us a clear measurement of neuronal reliability, but under very limited conditions analogous to human psychophysical experiments. In a natural setting, the fly must use its visual system to gather information about an angular trajectory $\theta(t)$, which is

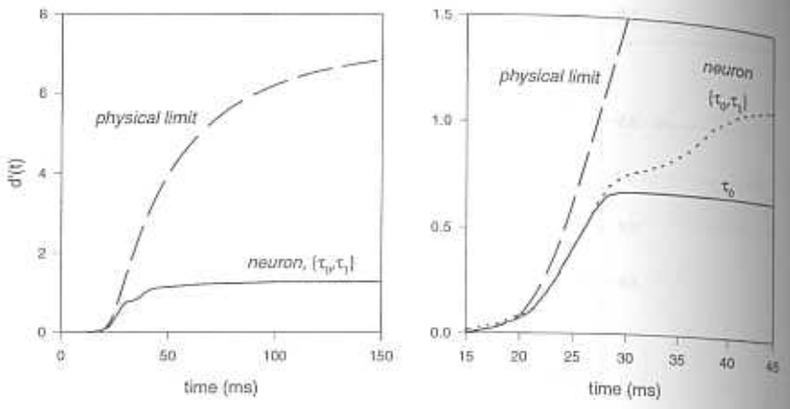



**Figure 4.25**
Comparison of discrimination performance using the single interval $\tau_0$, and the combination $\{\tau_0, \tau_1\}$, with the optimal discriminator (*physical limit*) using realistic photoreceptor signals as its input. For long times, the model outperforms H1 quite dramatically, as shown in the left panel. For short times, up to about 30 ms, however, the performance of H1 is only about a factor 2 less than the model's. This makes sense if we realize that 30 ms is a typical scale for flies to make course correction decisions. Also, we can reasonably assume that the fly is not interested in making very accurate discriminations at long time scales. For details see de Ruyter van Steveninck and Bialek (1995).

some unknown, continuous function of time. One can think of this trajectory as being random, drawn from some probability distribution $P[\theta(\tau)]$ which reflects the statistics of flight. The estimation process is limited in principle by the signal to noise ratio in the photoreceptor array, as in the case of the simpler discrimination task discussed above. Thus, another approach to determining the ability of H1 to signal movements is to ask how well an observer of the H1 responses can *estimate* a continuous, time varying movement signal.

To study the problem of continuous estimation, we again have the fly viewing a computer generated pattern, but now the pattern is moved continuously with a random angular velocity $\dot{\theta}(t)$. As described in chapter 2, it is possible to *decode* the spike train of H1 and recover estimates of the angular velocity; Fig. 2.20 shows an example of these reconstructions. These results provide evidence that the spike train of H1 contains enough information to infer—in real time, without averaging—details of the stimulus waveform on times comparable to the typical interspike interval, as suggested by the behavioral reaction times. The structure of the decoding algorithm confirms that the optimal esti-



mate of the waveform at one instant of time is controlled by the timing of, at most, a handful of spikes.

As in the analysis of Fig. 3.17, we separate the reconstructed velocity waveform into signal and noise components. We recall that the effective noise is the noise in our ability to estimate the stimulus waveform from real time observation of the spike train. If the noise in the reconstructions is Gaussian, then the effective noise spectrum, $N_{\rm eff}(\omega)$, is a complete characterization of how accurately the cell can encode the stimulus. Experimentally, the distribution of $n_{\rm eff}$ turns out to be essentially Gaussian, as shown in Fig. 4.26a. The reconstruction method maps the (discrete!) spike train into a continuous signal that bears a simple statistical relation to the original stimulus—it is a filtered version of the stimulus with added Gaussian noise, and hence the effective noise spectrum $N_{\rm eff}$ provides the meaningful description of the precision of the neural code.

Instead of thinking about the effective velocity noise, we can imagine that we use the ouput of H1 to estimate angular displacement. Of course, our estimates will be very bad at low frequencies, since the cell is not sensitive to a constant displacement. This should be revealed as a large effective displacement noise at low frequencies. Since the Fourier components of velocity are just those of the displacement multiplied by a factor of the frequency, the spectral density of displacement noise is obtained by dividing the velocity noise spectrum by the square of the frequency. The resulting spectral density of displacement noise, shown in Fig. 4.26b, measures the ability of an observer of the H1 spike train to judge the amplitude of a horizontal dither or oscillation as a function of frequency.

At low frequencies, the displacement noise rises rapidly, as expected. For frequencies in the range of 10–25 Hz, however, the noise is relatively flat at a level of $10^{-4}$ deg$^2$/Hz, or $\sim 0.01^\circ/\sqrt{\rm Hz}$. What does this noise level mean? If we, as observers of the H1 spike train, could afford to integrate for one second, the reconstructions would have an accuracy of $\sim 0.01^\circ$, much smaller than either the spacing of receptors on the retina or the nominal diffraction limit from the facet lenses. While it is unreasonable to suppose that the fly integrates for one second—we know that it can respond in 30 ms—any sensible integration time will lead to motion estimation in the hyperacuity regime, in agreement with the discrimination experiments discussed in the previous section. The first conclusion from Fig. 4.26 is thus that the fly visual system is capable of hyperacuity not just in discrimination tasks, but also in the much harder problem of *estimation*.

What computations must the fly carry out—what processing of the photoreceptor voltages—in order to achieve this level of precision in motion estimation? Information about movement across the visual field is carried in the

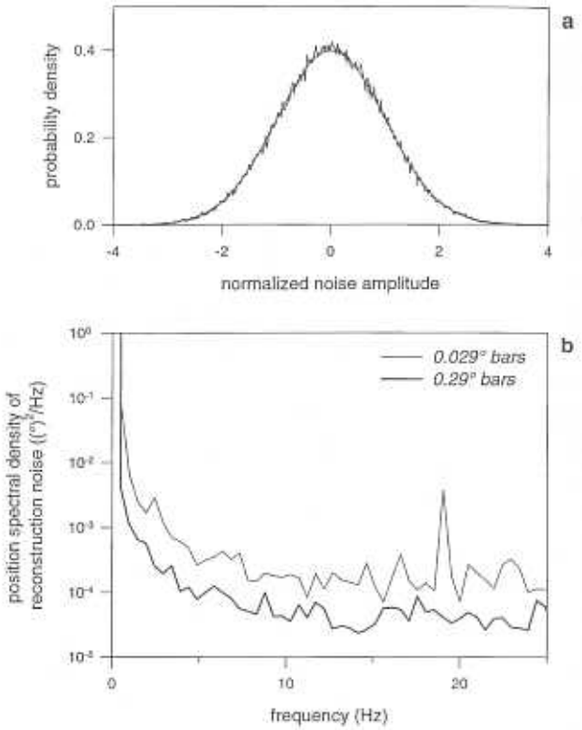



**Figure 4.26**
(a) Distribution of errors in an H1 estimation experiment. We normalize each sample $\tilde{\eta}(\omega)$ of the effective noise (computed for each of 400 sections of the stimulus and estimate) by the standard deviation $\sqrt{\langle|\eta(\omega)|^2\rangle/2}$. We then compute a histogram of these normalized noise amplitudes including points from each frequency $\omega$. A Gaussian with standard deviation of unity is also plotted for reference. (b) Power spectrum of effective displacement noise in H1 estimation experiment. The two solid curves shows results from experiments in which the bars composing the random pattern seen by the fly were 0.29° wide (thick line) and 0.029° (thin line). The stimulus power spectrum is given by the dashed line. Because H1 is a movement sensor and not a displacement sensor, low frequency displacements are not coded nearly as accurately as high frequency displacements. Above 10 Hz the effective displacement noise level for the 0.029° bars is such that an observer of the H1 spike train could estimate the amplitude of a sinusoidal modulation in the position of the pattern to an accuracy of about 0.06° with a 30 ms integration time; this is similar to the sensitivity measured in the step discrimination experiments.



spatiotemporal correlations of photoreceptor outputs—roughly speaking, we know that an object is moving because a receptor at site $n$ in the photoreceptor lattice "sees" the same signal as receptor $n-1$ saw some short time in the past. In an elegant experiment, Franceschini, Riehle, and le Nestour (1989) have recorded from H1 while stimulating just two individual photoreceptors that occupy neighboring positions in the retinal lattice, and they find that this most elementary "apparent motion" stimulus is indeed sufficient to trigger a response of the fly's motion sensitive neurons.

If we are looking at small motions of some particular pattern on the screen, then the voltage in a photoreceptor consists of two components. There is a constant voltage related to the static pattern, and there is a fluctuating voltage that has contributions both from the motion and from the photoreceptor noise. Since the pattern itself is random (by construction, in these experiments) we cannot extract the movement signal by combining photoreceptor voltages linearly, unless we have some extra knowledge about how to weight the different terms. More formally we can say that any translation invariant linear combination of voltages will vanish as we integrate over larger areas of the retina.

Since linear combinations of photoreceptor outputs cannot result in a nonzero signal, the simplest possible movement sensor will necessarily involve multiplying pairs of photoreceptor voltages. This is the basic idea of the correlation model for motion estimation, first proposed in the context of insect vision roughly forty years ago (Hassenstein and Reichardt 1956). There is an enormous body of evidence that fly optomotor behavior can be described, at least approximately, by a correlation model (Poggio and Reichardt 1976; Reichardt and Poggio, 1976; Buchner 1984), and the same can be said for the responses of H1 and the other movement sensitive neurons (Zaagman, Mastebroek, and Kuiper 1978; Borst and Egelhaaf 1989). One can show that, for discriminating among small step displacements, as in the experiments of the previous section, the computation of delayed nearest-neighbor correlations is in fact optimal—no other computation will lead to better discrimination performance in the same integration time (Bialek 1992; Rieke et al. 1996). We can find the limits to continuous estimation by making the guess that, so long as the displacements remain small, nearest neighbor correlation provides not only the optimal discrimination strategy but the optimal estimation strategy as well. For details, including a more mathematical justification of the focus on nearest neighbor correlation, see Bialek (1990, 1992) and Rieke et al. (1996); for a more general attack on the problem of optimal motion estimation, see Potters and Bialek (1994). Our concern here is not with the mathematical details but with the orders of magnitude: is a noise level of $10^{-4}\text{deg}^2/\text{Hz}$ close to the physical limit?

Motion estimation, as well as all other visual tasks, must become easier as we count more photons. This suggests that the physical limit to the effective noise level, $N_\theta^{\text{limit}}$, is inversely proportional to the photon counting rate $R$ in each receptor cell. Since we are interested in rigid motion, which is coherent across the entire visual field, we should also be able to lower the effective noise level by averaging the signals from $N$ photoreceptor cells. The effective displacement noise has units of deg$^2$/Hz, so we need an angular scale, and we know that this is set by diffraction through the lens, as in the discussion of human visual acuity; for flies this scale is much larger, $\phi_0 \sim 1.2°$, as seen in Fig. 4.12. From these arguments, and dimensional analysis, we expect

$$N_\theta^{\text{limit}} \propto \frac{1}{R} \cdot \frac{1}{N}\phi_0^2. \tag{4.45}$$

This is almost right. We also need to include the effect of the contrast in the images—if the world is uniformly grey, we cannot see it move. As noted in section 3.1.4, photon shot noise contributes an equivalent contrast noise of $1/R$, so that the signal to noise ratio in images with contrast $C$ is $RC^2\tau$, where $\tau$ is the integration time. Thus increasing the mean square contrast is just like increasing the photon counting rate, so that

$$N_\theta^{\text{limit}} \sim \frac{1}{RC^2} \cdot \frac{1}{N}\phi_0^2, \tag{4.46}$$

and this is (nearly) the result from a more rigorous analysis (Bialek 1990), which shows that there is a constant factor in front with a relatively large magnitude, so that the true physical limit is $\sim 60$ times larger than the simple expression in Eq. (4.46).

In the experiments on H1, the typical photon counting rates were $R \sim 10^4$ photons/s, the typical contrast of the images as seen through the photoreceptor aperture was $C \sim 0.16$, and roughly 2,500 receptor cells were illuminated. These parameters, substituted into Eq. (4.46) including the constant factor from the more rigorous theory, set the limiting noise level $N_\theta^{\text{limit}} \sim 10^{-4}$ deg$^2$/Hz. In Fig. 4.26 we see that the displacement noise from the linear reconstruction approaches this limit imposed by photon shot noise and diffraction, at least at high frequencies where this theory of the limiting noise level is valid (Bialek et al. 1991). We can increase the contrast seen by the photoreceptors by presenting random patterns with wider stripes, and if the fly continues to perform optimally this should, from Eq. (4.46), reduce the effective noise level still further, and this is also observed. The fly visual system thus performs an optimal and nearly noiseless extraction of the motion signal from the array of photoreceptor voltages, and this is accomplished in real time.



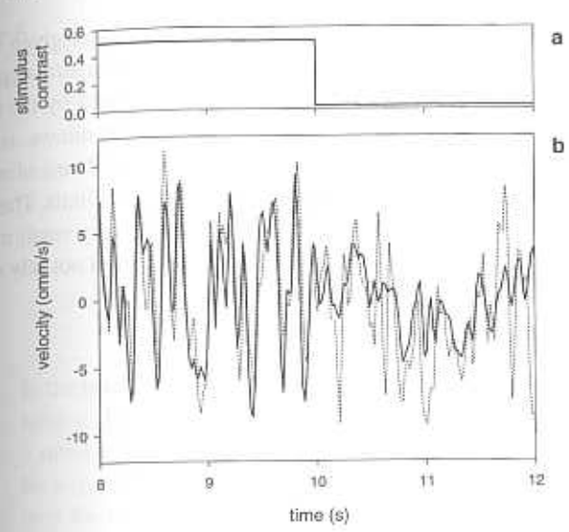



**Figure 4.27**
Section of stimulus waveform (dotted line) and reconstruction (solid line) over a 4 s time window, during which the root mean square contrast is switched from 0.5 to 0.02. During 10 s the fly watches a high-contrast pattern move randomly. At 10 s, the contrast suddenly is switched to very low (0.02). Clearly in low-contrast conditions it is more difficult to make accurate reconstructions than at high contrast. This is to be expected if the performance of H1 is limited by peripheral noise sources, such as photon arrival statistics.

The comparison of noise levels in reconstruction experiments at different contrast values may seem a bit indirect. After all, when the fly emerges from a wooded area into an open field, the contrast and other parameters of the visual environment change immediately. Certainly a sudden drop in contrast should dramatically reduce the precision of motion estimation—again assuming that the fly is limited by the physics of its inputs and not by some spurious internal noise generator. We can see this effect directly in the experiments of Fig. 4.27.

## 4.4 SUMMARY

As a first attempt at quantifying the reliability of neural computation, we can try to perform psychophysical discrimination experiments with single cells.

This approach, which began with Barlow and Levick, has given us several examples in which the performance of individual cells approaches the performance of the organism as a whole or the limits imposed by the physics of the input signals. The method of stimulus reconstruction allows us to extend these observations from discrimination to the more natural task of estimation, and again we see a precision that approaches the physical limit. The results of these very precise computations can be represented by very small numbers of spikes, demonstrating clearly that the nervous system need not rely on the law of large numbers to synthesize reliable percepts.

# Chapter 5
# Directions

In the preceding three chapters we have tried to develop a precise language for talking about the neural code, and then we have used this language to address a small set of questions. In this brief chapter we allow ourselves to look a little bit beyond these questions to three areas where we expect substantial progress over the next few years. In each case the issues are at least three decades old, and in some cases we can find the origins of the modern issues in discussions at the start of the twentieth century. Perhaps the ideas of the previous chapters allow us to come back to these problems with a new point of view.

In the preface to the paperback edition of his classic text on nuclear magnetic resonance, Abragam (1983) cautions prospective authors to "never put in a book (as opposed to a research paper) anything that you do not understand thoroughly." In this chapter we run the risk of violating this dictum, but we hope not to go too far. We hope that the reader enjoys these less complete descriptions of ideas in progress in the spirit they are intended.

## 5.1 ARRAYS OF NEURONS

Most of the text has been about the problem of the impoverished homunculus who tries to make inferences about the world by looking at the spike train from only one neuron. One possibility is that each cell can tell the organism something completely unique and independent, so that a picture of the whole sensory world can be built in a simple way by adding up the pieces provided by individual neurons. In the other extreme, the outputs of individual cells could be completely ambiguous, and meaningful interpretations might require comparisons among spikes from different cells.

In the context of simple models for neural firing, one can answer the question of how the signals from different cells should be combined to reach the most reliable estimates of the sensory stimulus. This mathematical approach

has a long tradition, going back to Siebert's work on the auditory nerve, as described in section 4.1.3 (Siebert 1965, 1970). A number of groups have looked at the experimental analog of Siebert's problem, asking how one could recover estimates of frequency or, more generally, the power spectrum by observing the whole population of auditory nerve fibers. There has been particular interest in the encoding of continuous sounds that approximate the acoustic elements of speech, such as periodic waveforms that match the power spectra of spoken vowels. For some of these developments see Young and Sachs (1979), Miller and Sachs (1983, 1984), and Winslow, Barta, and Sachs (1987).

There has been a steady stream of theoretical papers devoted to coding by populations of neurons, but few have addressed the central problem in Siebert's discussion, namely the connection of the code to the accuracy of the percepts the organism can form. More recently Seung and Sompolinsky (1993) have revived this theoretical problem, focusing on the encoding of angular variables such as orientation or direction, either for the visual system or for the motor system. In all of this work, one imagines an array of cells that are more or less identical except that each is tuned to a different value of some stimulus parameter, such as frequency in the auditory system or orientation in the visual system. In the simplest case we are interested in estimating this one stimulus parameter, and we would like to know how to combine the outputs of the many cells to form the best estimate. In addition, as emphasized by Seung and Sompolinsky (1993), we would like to know how the rules for extracting information of relevance to, for example, a particular psychophysical discrimination task, could be learned through feedback.

From the experimental side, much of the stimulus for renewed interest in these problems comes from Georgopoulos and coworkers, who studied the representation of movement directions in the primate motor cortex (Georgopoulos, Taira, and Lukashin 1993). At least in the context of certain behavioral paradigms, each cell seems to have a preferred direction of motion, and the firing rate of the cell is related to the cosine of the angle between the actual direction and the preferred direction. As a simple hypothesis for how information from many such cells can be combined, Georgopoulos, Schwartz, and Kettner (1986) suggested the construction of a population vector: a unit vector pointing toward the preferred direction of each cell is weighted by the firing rate of that cell, and all the vectors are summed. This vector has some interesting properties. In particular, it seems that the population vector rotates in anticipation of movement, inspiring us to believe that we are reading out the monkey's intentions (Georgopoulos et al. 1989).



From a quantitative point of view, there are many open questions about the population vector picture. Salinas and Abbott (1994) have emphasized that the population vector is subject to systematic biases if one records from a population of cells that do not uniformly cover the sphere of possible movement directions. But the population vector is just one possible linear combination of the firing rates in the different cells. Salinas and Abbott show that the optimal linear combination can generate estimates that are dramatically more accurate than the usual population vector, at least on the assumption that the spikes fired by different neurons are statistically independent.

The notion of statistical independence is a convenient approximation for theoretical developments, but, as in the discussion of reliability in MT neurons (section 4.1.4), we should be mindful that violations of this assumption can have qualitative effects on the performance of the system. For the motor system there has been less quantitative analysis of reliability than on the sensory side, but all of the theoretical work (Seung and Sompolinsky 1993; Salinas and Abbott 1994) encourages us to think that the problems are analogous. We might, then, expect that the pattern of correlations among cortical cells coding for visual movement direction (Zohary, Shadlen, and Newsome 1994) will also be found in the cortical cells coding for arm movement direction. But we should worry that, as in MT, correlation blocks the $\sqrt{N}$ improvement of estimation accuracy as the number of cells increases. In the case of MT, this observation is actually quite satisfying, since the reliability of single neurons is close to the reliability of the organism as a whole, but in the motor system it is less clear that single cortical neurons represent direction with an accuracy comparable to the accuracy of actual arm movements (Donchin and Bialek 1995).

The coding of direction in the motor cortex has analogies in the much smaller network of the cricket cercal system, where a handful of cells encode the direction of air motion. These cells take inputs from the primary afferent neurons, whose coding properties were discussed in section 3.3.1. Miller, Jacobs, and Theunissen (1991) have described the array of interneurons in a way that makes explicit the analogy to cortical maps or population codes, and Theunissen and Miller (1991) have given an information theoretic characterization of these cells, assuming that the spike count is the relevant output and that wind direction is the important stimulus variable. In richer, dynamic stimulus ensembles, the outputs of individual interneurons can be decoded using the linear reconstruction methods (Theunissen 1993), and the spike trains of these cells convey large amounts of information about the time dependence of the

wind velocity. Thus the array of spike counts has been used to infer wind directions, and the spike times of individual cells have been used to infer the time dependence of wind velocity. The availability of simultaneous recordings of the spike trains from essentially all of the cells in this small map (Gozani and Miller 1994) should make it possible to try experiments that aim at the reconstruction of fully natural time-dependent patterns of air flow.

The cricket cercal system provides an example in which the ideas of population coding—which are based, traditionally, on a firing rate or spike count description of coding in single cells—meet the ideas of temporal coding. More generally, we have argued that, for single cells, the dynamics of natural signals are such that significant information must be contained in the occurrence times of single spikes. In the case of the fly's motion sensitive neuron H1 we have seen explicitly how the occurrence time of the first spike after the onset of a stimulus provides much of the information available for discriminating among different possible stimulus settings (section 4.3.2). If the occurrence time of one spike in one cell can represent the value of the stimulus parameter coded by that one cell, then in a multiparameter stimulus encoded by an array of cells, stimulus identity will be represented by the relative arrival times of individual spikes from each cell. Such a representation in terms of relative timing, as opposed to relative rates, has been advocated by Hopfield (1995), who identifies several computational advantages for this scheme.

Another system that seems amenable to analysis of multineuron coding is the rat hippocampus. From the work of O'Keefe and collaborators, there is strong evidence that the responses of cells in this region provide the rat with information about its location in space. In particular, there are "place cells" that fire when the rat is in particular small regions of an enclosed environment (O'Keefe and Nadel 1978). Improvements in electrode technology now make it possible to record from large numbers of place cells as the rat explores its surroundings (O'Keefe and Recce 1993; Wilson and McNaughton 1993).

Wilson and McNaughton (1993) have used the spike trains from 50–100 place cells, recorded simultaneously, to reconstruct trajectories of the rat's motion through an enclosed region. This is very much in the spirit of the "organism's point of view" introduced in chapters 1 and 2—given the outputs of the place cells, what can the rat learn about where she is? The particular reconstruction strategy used by Wilson and McNaughton involves breaking time into discrete windows and then accumulating the spike count of each cell in this window. Then the array of cells gives a vector of spike counts, and one can try to map these vectors into spatial locations. One strategy is to assume that each cell "points" to a definite location in space, and take an average of



these locations weighted by spike counts in the same way that the population vector in the motor system takes a weighted average of movement directions. Another approach is to try and match the vector to one of a set of stored vectors that represent the average spike counts in each small region of the space. Both procedures seem to lead to reasonable reconstructions, although mathematically neither scheme is optimal. The weighted averaging approach can be improved to find the optimal linear estimator, as explained by Salinas and Abbott (1994).

In addition to the work on place cells, the hippocampus is one of the key systems for the study of cellular and synaptic mechanisms in learning and memory. Recently, Blum and Abbott (1996) have brought these themes together. They consider a network of cells with some fixed rules for reading out the signals, such as the rat's location, that are encoded by this array of cells. Then the synapses connecting the cells are modified according to some plausible rules, and if we hold fixed the rules for reading the code, then these synaptic modifications will change what the network tells us. For a simple set of learning rules, Blum and Abbott have found that when synapses in a place cell network are modified by the experience of moving along certain paths, then decoding the network's response generates a prediction of the next location along the most often used paths, rather than a measure of the current location. This suggests that the combination of place cells and synaptic plasticity could yield a simple mechanism of navigation, and they consider the application of this idea to some of the classic experiments on rat behavior.

New recording techniques have also made their way to the sensory periphery. Meister, Pine and Baylor (1994) developed a multi-electrode array that allows simultaneous recording from ~50 retinal ganglion cells. By monitoring activity in many nearby cells in the salamander retina, Meister and colleagues (Warland and Meister 1993, 1995; Meister, Lagnado, and Baylor 1995) found that the message conveyed by a single spike in one retinal ganglion cell depends on the activity of neighboring cells. Single ganglion cells generate spikes that are either coincident or noncoincident with spikes in neighboring cells, with coincidences defined as spikes that occur within the narrow peak of the correlation function, and these coincident spikes seem to "stand for" intensity variations in different regions of the visual field, as indicated in Fig. 5.1. Information theoretic analyses suggest that these coincident events allow the pair of cells to carry more information than the sum of the two cells in isolation (Warland and Meister 1993, 1995). DeVries and Baylor (1996) have used the electrode array to study how different types of ganglion cells sample the visual world. Several different classes of ganglion cells have

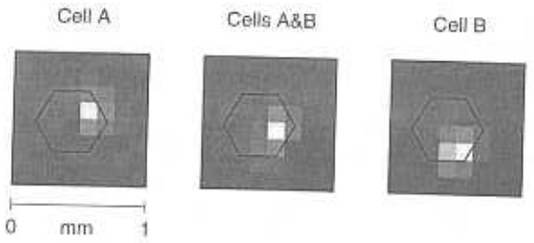



**Figure 5.1**
Decoding filters for two ganglion cells, A and B, in the salamander retina. Spike trains from two cells were filtered to estimate the time course of a randomly varying checkerboard stimulus. To search for messages carried specifically by correlations between the two cells, the original two spike trains were recoded from a pair of cells A and B into three trains of events, corresponding to cell A firing within 30 ms of cell B (A&B), cell A firing alone, and cell B firing alone. The decoder for each event train can be approximated by a product of a function of space and a function of time, and the spatial components are shown here. The hexagon represents the area spanned by the array of electrodes used in these experiments. The joint firing pattern A&B contributes to the estimation process in a spatial region in between the contributions from A and B, suggesting that higher visual processing centers can resolve finer spatial details if they recognize that two cells fired together. From experiments by Warland and Meister (1995).

been characterized in mammalian retina. In the rabbit, each of these classes of ganglion cell "tiles" the retina (DeVries and Baylor 1996), so that the receptive field centers of ganglion cells of the same class form a nearly perfect lattice, and the lattice constant is given by the width of the receptive field.

Neuroscientists once spoke wishfully of the day when one would be able to record from large numbers of neurons. At the time of this writing, simultaneous recordings from tens of cells are routine in many laboratories. In some preparations of order one hundred simultaneous spike trains are accessible if one works very hard, and it is not difficult to imagine that of order one thousand will become accessible within the next several years. Serious questions arise about how to understand and process the resulting volumes of data, which are at the same time too large to sift through by hand and too small to allow an exhaustive analysis of all possible interactions among the cells. It seems that the idea of decoding—how can one read out the response of the neural population to say something about what is going on in the world?—provides a useful framework for analysis, at least in the systems studied thus far. The possibility of asking this question in the context of naturalistic sensory environments is especially exciting.



## 5.2 NATURAL SIGNALS

We have emphasized that the problem of neural coding must be phrased with respect to some assumed sensory environment: Bayes' rule tells us that the interpretation of spike trains depends on our prior hypotheses about what is likely. Ideally, we would like to study the nervous system in the most natural of environments, but this is difficult. One problem is that true natural stimuli have an enormous amount of structure, and in studying a single cell or a small region of the brain we need to discover which structures are actually important for these particular cells: Do retinal ganglion cells "know" that the visual world is built out of objects, or does this become important only in the higher stages of cortical processing? To address these issues it is not enough to present the visual system with "natural" stimuli, we have to understand these stimuli in some detail.

Understanding the structure of natural stimuli is really a physics problem. In the case of olfaction, we actually know the equations that give rise to the dynamics of natural stimuli, since odors are carried to the nose (or antenna) by a combination of diffusion and advection. For very small creatures, like bacteria, diffusion is dominant, and the physics of this diffusion–dominated regime determines many of the signal processing strategies for chemical sensing in bacteria (Berg and Purcell 1977). For an insect seeking a pheromone source or finding food in an open field, transport of odorants is dominated by advection—the odor molecules are simply carried along as the air moves. Furthermore, the motion of the air itself is turbulent. What do we know about the spatial and temporal dynamics of odorants (or any tracer molecules) carried by turbulent flows?

The dynamics of odorants in turbulent flows has certain universal features—qualitative and even quantitative features that are independent of the detailed mechanism generating the flow. Because the equations are the same, we expect similar behavior in air and in water, provided that we look on appropriate scales. Turbulent flows act to dissipate local variations in the density of odor molecules, while macroscopic gradients are concentrated into short, steep steps in the otherwise smooth concentration profile. Thus while the pheromone concentration tends to be highest near the source, there is no smooth gradient which can be followed "uphill" to the source. Instead the concentration gradient is a very intermittent signal, having long spatial and temporal epochs near zero, punctuated by large brief pulses. One way of revealing this intermittency is to look at the probability distribution of the concentration signal, or of its time derivative; this distribution has a long, nearly exponential tail (Cas-

taing et al. 1989; Gollub et al. 1991). Understanding these distributions from first principles remains an important theoretical problem. For recent efforts see Shraiman and Siggia (1994) and references therein.

The intermittent structure of concentration profiles is crucial to understanding olfactory navigation. Not so long ago, much of the literature on insect olfaction in particular worked on the tacit assumption that turbulent flows were equivalent to enhanced diffusion, so that concentration profiles spread out smoothly from the source. This is wrong; instead the concentration profiles consist of very thin plumes that meander outward. Navigation models that assume the insect could follow a smooth gradient toward the source (which works, in principle, in the bacterial environment) are therefore also wrong, and they are wrong because of the physics, not the biology (Murlis, Elkinton, and Cardé 1992).

Recent experiments strongly suggest that insects "know" about the structure of turbulent flows and expect to see (smell, actually) these structures. Exposure to steady gradients does not seem to drive insects up the gradient, while exposures to repeated pulses (Vickers and Baker 1994) or simulated plumes (Mafra-Neto and Cardé 1994) lead to clear and steady flight toward the odor source. Similar considerations seem to apply to chemical navigation by lobsters (Atema 1995). It seems likely that by proper combination of experimental physics and ethology, it will be possible to do behavioral experiments in controlled models of the natural environment, studying the evolution of navigation strategies as a function of the crucial parameters in the turbulent flow. But already these experiments teach us that there must be interesting features of the neural circuits that are responsible for encoding and processing these complex dynamical signals.

One way to generate reasonably "natural" images is to visualize a turbulent flow. For example, if we look at the concentration profiles of an odorant carried by the flow, we see "objects" that correspond to the plumes discussed above. These plumes move around, bending and swaying, but retain their identity. This is one hint that the statistical structures of olfactory and visual signals might be related, which is at first sight (or smell) a strange idea.

Exploration of the statistics of natural images has been driven in part by Barlow's (1961) suggestion that the visual system "knows" about these statistics and uses them to provide more efficient representations of the visual world. We review some more recent attempts to quantify this idea in the next section. Laughlin (1981) presented a very simple version of this problem, in which a single cell must encode contrast variations with a graded voltage response. The distribution of contrasts in the environment has some shape, and



we might expect that this distribution has a long tail (see below). But with a limited range of voltages one would like to use the full range equally, maximizing the entropy of the voltage distribution and hence (with reasonable assumptions) maximizing the information the voltage can convey about the contrast signal. Laughlin measured the distribution of contrasts as seen through an aperture the size of a fly photoreceptor as it moves through the woods, and computed the ideal contrast-to-voltage conversion. This can be compared to the input/output relation of the second order neurons (the large monopolar cells) in the fly visual system, and the match is very good. Although it is just a beginning, Laughlin's result strongly suggests that this first synapse in fly vision is adapted to the statistical structure of natural scenes.

One evident feature of the visual world is that structure appears on all scales. Even if objects have a characteristic linear dimension, they can appear at any distance from us, and hence there is no characteristic angular scale. This notion of scale invariance was explored by Field (1987), who showed that the power spectra of various natural images has approximately the form one would expect from scaling considerations. The argument runs as follows. Imagine that we measure the contrast $\phi(\mathbf{x})$ at each point in a two dimensional gray-scale image. The contrast is defined as the difference between the intensity at point $\mathbf{x}$ and the mean intensity in the image, normalized by the mean intensity. Thus the contrast is a dimensionless number. Images are (statistically) invariant under translations, since we imagine that any object can appear at any location, as long as we stay away from the horizon or other large scale boundaries. Thus, if we average over an ensemble of similar images—for example, all the pictures we encounter on a walk through the woods—then we should find that the correlation function of the contrast (see section 3.1.4 for a discussion of the analogous functions in the time domain) has the form

$$\langle \phi(\mathbf{x})\phi(\mathbf{x}')\rangle = C(\mathbf{x}-\mathbf{x}') \tag{5.1}$$

$$= \int \frac{d^2k}{(2\pi)^2} S(\mathbf{k}) \exp[-i\mathbf{k}\cdot(\mathbf{x}-\mathbf{x}')]. \tag{5.2}$$

In this expression, $C(\mathbf{x})$ is the correlation function and $S(\mathbf{k})$ is the power spectrum. The spatial variable $\mathbf{x}$ is a vector in two dimensions with the units of (for example) degrees, and the Fourier variable $\mathbf{k}$ is the spatial frequency with units 1/degree. The contrast is dimensionless, and therefore the correlation function is also dimensionless. To satisfy Eq. (5.2), the power spectrum has to have units (degrees)$^2$. But then, if there is no characteristic angular scale, we must have (it seems)

$$S(\mathbf{k}) \sim \frac{A}{|\mathbf{k}|^2}, \tag{5.3}$$

with $A$ a dimensionless constant.

Field (1987) found that several images have a nearly $1/k^2$ power spectrum, and he argued that this particular distribution of power across spatial frequencies is well matched to the spatial frequency selectivities of cells in primary visual cortex. Specifically, he suggested that the pattern of spatial frequency and orientation tuning in cortex builds a set of filters such that each filter accepts an equal fraction of the total contrast variance in the image. This is a simple and thought provoking idea, and Field's work sparked a great deal of interest in the relation of natural image properties to the properties of the visual system.

Not all images have a simple scaling power spectrum, as emphasized by Tolhurst, Tadmor, and Chao (1992). But we should recall that, as discussed for time (rather than space) dependent signals in section 3.1.4, the power spectrum is the analog of the variance, and hence should be defined an average over an *ensemble of images*. Ruderman (1993; Ruderman and Bialek 1994) studied the statistics of an ensemble of images taken in the woods. This work was motivated in part by a search for a more general form of scaling. In physics we know that systems can exhibit scale invariance across a very broad range of length scales but the correlation functions need not obey the simple rules derived from dimensional analysis. The best studied examples are second order phase transitions, such as the gas–liquid critical point, where the power spectrum of density fluctuations in the fluid is almost of the form in Eq. (5.3), but instead of $1/k^2$ one observes $1/k^{2-\eta}$. The *anomalous dimension* $\eta$ can be measured in light–scattering experiments, and there are a set of other scaling relations for thermodynamic quantities, such as the specific heat, that also have anomalous behaviors. The correct calculation of these scaling behaviors is one of the triumphs of the modern renormalization group theory of critical phenomena (Wilson 1975, 1983; Ma 1976).

Power spectra measured for an ensemble of images taken in the woods show scale invariance over three orders of magnitude in spatial frequency, with an anomalous dimension of $\eta = 0.19 \pm 0.02$ (Ruderman and Bialek 1994). But measuring the power spectrum is only one test of scale invariance. If we really believe in scaling, then the particular choice of digitizing an image into pixels should be arbitrary—it should be possible to form "block pixels" by averaging over $2\times2$ blocks, and the statistics of the blocked images should be the same as those of the original ensemble. We have to be careful, however,



because when we average over blocks we reduce the total contrast of the images, so we have to renormalize the contrast. This sequence of blocking and renormalization is the essence of renormalization group theory (Wilson 1975, 1983), and in fact the block construction was identified as a crucial step long before the full theory was in place (Kadanoff 1966).

For the ensemble of images in the woods, most local statistical properties—the distribution of contrast, the distribution of contrast gradients, etc.—are invariant to the construction of block pixels, and an example of this scaling behavior is shown in Fig. 5.2. This is direct evidence that whatever statistical structure is present in these images is present on all angular scales. But the examination of the contrast and contrast gradient distributions also reveals

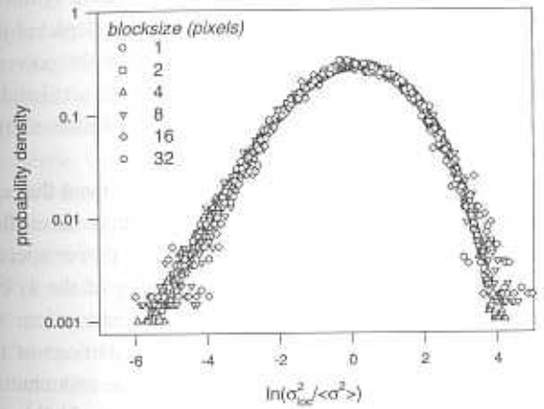



**Figure 5.2**
Local variances and scaling in natural images. The local variance $\sigma^2_{loc}$ is defined as the variance among the intensities in a $2 \times 2$ array of neighboring pixels. Histograms of the local variance are accumulated from an ensemble of images gathered in the woods, and then the local variance is normalized by its ensemble average. The histograms are converted into probability distributions for the natural logarithm of this normalized quantity, and the distributions are shown here on a logarithmic scale. Before analyzing the images, however, we are free to redigitize them into pixels that are larger than the pixels provided by the camera, in effect creating new pixels out of blocks of the original pixels. As explained in the text, this blocking changes (for example) the average local variance, but this effect is removed by the normalization procedure. Probability distributions of the (normalized) local variance measured on these blocked images overlay those measured in the original images, at least up to block sizes of $32 \times 32$ original pixels, demonstrating the scale invariance of the image ensemble. From experiments by D. Ruderman (1993).

that these distributions have long tails, very much like the distributions of odorant concentrations in a turbulent flow (Ruderman and Bialek 1994). In particular, if we filter images through the sorts of center–surround receptive fields that are thought to characterize the early stages of visual processing, then the distribution of outputs from this filtering is precisely exponential and invariant to the overall scale of the receptive field (Ruderman 1993).

Although much of the recent interest in image statistics was triggered by Field's (1987) observations on power spectra, the presence of long tails in the contrast distribution is one of several hints that the spectrum alone does not capture the statistical structure of these signals, and this is again reminiscent of the situation in turbulent flows. One can construct images that are chosen from Gaussian probability distributions, and hence miss the long tails, but nonetheless have power spectra like those of natural images. These synthetic images have no "objects"—they are much too soft and smooth to depict object boundaries. Alternatively, one can filter natural images so that the power spectrum becomes flat, like white noise, yet most images and the objects inside them are perfectly recognizable. There is clearly some important statistical structure we do not yet know how to quantify.

Nearly twenty years ago, Voss and Clarke (1977) showed that speech and music exhibit slow fluctuations in amplitude and frequency. These fluctuations occur on all time scales, and an analysis in terms of power spectra reveals an approximate $1/f$ spectrum, the time domain analog of the $1/k^2$ for two-dimensional images. Remarkably, music synthesized at random with these statistics has a rather pleasing sound. Probability distributions of the instantaneous sound pressure in musical pieces have the same exponential form as the distribution of the local contrast in images, and recently Nelkin (1995) has found that some of the hierarchical structures of the nonGaussian fluctuations in images (Ruderman 1993) are also found in natural sounds.

At present, the exploration of statistical structure in natural signals is at a very early but also a very enjoyable stage. The idea that there is something similar in the statistics of odor concentrations, image contrast, and acoustic waveforms is certainly appealing. After all, the cortex that understands images is not so different from the cortex that writes music. On the other hand, it may be that these universal features of natural signals, even if borne out by more detailed investigations, are irrelevant to our perception of the world. Whatever the outcome, even our current crude understanding of natural signals is enough to raise questions about how the nervous system encodes and processes signals with these very peculiar statistical properties.



## 5.3 OPTIMAL CODING AND COMPUTATION

At the beginning of this book, we warned the reader that much of our discussion would be driven by the desire to quantify the behavior of neurons. We hope that the discussions of chapters 3 and 4 have refined this notion of quantification to include the idea that one must place the performance of neurons on an absolute scale. Thus, being absolutely sure that a particular pattern of motion across the visual field increases the firing rate of a neuron by $23.5 \pm 0.7$ spikes per second isn't really interesting. Is this a large or a small change? Is this change enough to indicate reliably some parameters of the motion trajectory? Are the modulations of firing rate fast enough to convey information about interesting dynamic signals in the natural environment?

In trying to establish an absolute scale for neural performance, we were guided on the one hand by the behavior of the whole animal, and on the other hand by the notion of a physical limit. The intuitive example of a physical limit is diffraction, where we know that no optical system can be completely devoid of blur—there is a minimal blur set by the laws of physics. In the same sense, there is a maximum amount of information that neurons can transmit, and there is a maximum reliability with which the nervous system can estimate nontrivial features of the sensory environment. At the outset we have no reason to think that these physical limits are relevant to real brains—the design of the nervous system could be driven by completely independent criteria. Indeed the biology of today is the product of a long evolutionary history, and many authors have emphasized that artifacts of this history may dominate any notion of efficient or optimal design for today's organism.

It may come as a surprise, then, that in several different systems the performance of neurons does approach the relevant physical limits. Many authors have explored the idea that such near optimal performance is not a coincidence, but is rather a general principle from which many aspects of neural coding and computation can be understood. Given the evidence for performance close to the physical limits, it certainly makes sense to ask which features of neural coding and computation are essential to this remarkable performance. What is the structure of a neural code that allows such high rates of information transmission? What computations are necessary for maximum reliability in estimation problems? These are challenging theoretical problems, by no means trivial applications of known results. A large part of the difficulty is connected, again, with the description of the natural sensory environment. Most progress to date has, therefore, been made by studying a model world that is a simpler and less structured place than the real world, hoping that the

optimal strategies for dealing with this simple world will at least give us hints about optimal strategies for the real world.

Considerable attention has been given to the problem of optimal coding in the early stages of vision. These ideas go back, at least, to Barlow's discussion of retinal ganglion cells, where he suggested that the center-surround organization of receptive fields serves to make the responses of neighboring neurons less correlated than the intensity values in neighboring pixels of the visual stimulus (Barlow 1961). This is the idea of decorrelation, or redundancy reduction. If we imagine that each retinal ganglion cell has a limited capacity to carry information about the visual scene, then, to the extent that different cells are telling the brain about the same feature of the scene, we are not making maximum use of this limited capacity. More subtly, if one cell tells us about some feature of the world and a second cell tells us about a different feature, we still may not have an efficient code if one feature can be predicted from the other. In the case of written English, for example, Q and U are separate signals, but clearly it is often possible to predict the occurrence of a U from the observation of Q, and it would be inefficient to have separate 'Q' and 'U' neurons. There exists instead some optimal combination of Q–U sensitivities that minimizes the redundancy of messages carried by the different neurons and therefore makes maximum use of the information capacity of these cells.

Formally, the redundancy $R$ of two neurons $A$ and $B$ can be defined by measuring the information they provide when observed simultaneously, $I_{AB}$, and comparing this with the information they provide when observed separately, $I_A$ and $I_B$. The redundancy is

$$R = I_A + I_B - I_{AB}. \tag{5.4}$$

We can generalize this definition to include whole arrays of neurons, not just two. One candidate optimization principle for the neural code is, then, to minimize the redundancy $R$. But minimizing redundancy is not, by itself, a sensible principle, since it ignores the role of noise. Thus an array of completely independent messages is indistinguishable from an array of random bits, so if we reduce redundancy to zero we lose the ability to distinguish meaningful signals from random variations. Put another way, redundancy is good because it affords the signal some protection against the addition of noise. Presumably, the correct principle involves minimizing redundancy while holding something fixed, such as the total information transmission ($I_{AB}$ in our example), given some model of the noise in the system. This is exactly the principle explored by Atick and coworkers, who show how many of the observed receptive field properties of retinal ganglion cells can be derived from the solution to



this constrained optimization problem; for a review of these ideas, see Atick (1992).

As an alternative to minimizing redundancy, we can imagine that the early stages of neural coding are designed to maximize information transmission. Again, by itself this is not a well posed problem, and one needs to introduce some constraints. For proper choices of constraints, the minimization of redundancy and the maximization of information are actually the same problem. We have seen an example of information maximization in section 3.1.4, and to get some intuition for optimal coding problems we will look at a related example here.

Imagine that we have a time dependent signal $s(t)$ embedded in a background of Gaussian white noise $\eta_1(t)$. This might be, for example, a visual signal, and $\eta_1(t)$ may be the noise due to random photon arrivals that inevitably accompanies visual stimulation. Let us imagine that our "neuron" is a device that converts this signal into a voltage and then adds some additional Gaussian white voltage noise $\eta_2(t)$. For simplicity, we assume that the conversion process is linear, so it is completely described by a transfer function $F$, as in our discussion of fly photoreceptors and large monopolar cells in section 3.1.4. Thus we have a cell that gives an output voltage

$$V(t) = \int_0^\infty d\tau F(\tau)\left[s(t-\tau) + \eta_1(t-\tau)\right] + \eta_2(t), \tag{5.5}$$

and the question is how much information the output of the neuron, $V(t)$, provides about the input signal $s(t)$. If the world is simple, we can assume that the signal $s(t)$ is also chosen from a Gaussian distribution, so it is completely characterized by its power spectrum $S(\omega)$. Then, as discussed in section A.19, we can calculate analytically the rate of information transmission, and we can solve the problem of optimizing this rate subject to some reasonable constraints. For example, it is clear that the larger the magnitude of $F$ the less important is the noise $\eta_2(t)$, but this comes at the cost of very large excursions in the output voltage of our cell. We discuss the filter that maximizes information transmission while holding fixed the variance of these voltage fluctuations, corresponding to a limited dynamic range of the cell.

From what we learned about natural signals in the last section, we might expect that the power spectrum has an approximately scale invariant form, that is $S(\omega) \sim 1/\omega^\alpha$, with $\alpha \sim 1$. Under these conditions the signal to noise ratio is always high at very low frequencies, and in this limit the optimal filter is a highpass filter. Specifically, if we look at the Fourier transform of $F(t)$, which we call $\tilde{F}(\omega)$, then $|\tilde{F}(\omega)| \sim 1/\sqrt{S(\omega)}$. This means that once the signal is

passed through the optimal filter it comes out with a flat power spectrum—the optimal filter takes in correlated signals and puts out something that looks like white noise, so long as the signal to noise ratio is large. This is decorrelation, or redundancy reduction in the time domain, as emphasized by Srinivasan, Laughlin, and Dubs (1982) and then by van Hateren (1992) in their analyses of signal transfer across the first synapse in fly vision.

If we take seriously the naturalistic spectrum $\sim 1/\omega^n$, then there is the clear prediction that the optimal encoding filter has zero gain at zero frequency. The same argument can be made for an array of cells coding a spatially varying signal—if the spatial power spectrum has the natural form $S(\mathbf{k}) \sim 1/k^{2-\eta}$, then the optimal spatial filters have zero gain for $k = 0$—the cell that provides maximal information transmission does not respond to spatially uniform stimuli. Usually this is explained by saying that these constant stimuli "carry no information," but in the optimization of information transmission something else is happening: low frequency (not strictly constant) signals have lots of power, so they can be attenuated without concern about losing them in the noise, and this frees up the limited dynamic range of the cell to represent the more fragile signals at intermediate frequencies.

At sufficiently low frequencies, the signal to noise ratio is high, and optimal filter has a highpass character, but, as the frequency is increased the signal power decreases and we eventually reach a point where the signal gets lost in the noise. Since neurons have a limited dynamic range, there is no point in wasting this dynamic range to represent frequency components that are almost certainly noise alone. The optimal filter rolls off at high frequencies, suppressing this noise. As emphasized by Atick (1992), one can think of the optimal encoding filter as having two pieces—a lowpass filter to separate signal from noise, and a highpass filter to reduce redundancy. The combination makes a bandpass characteristic, and if we transform back into the time domain, the filters look like a smoothed version of a time derivative operation. In the case of spatial stimuli, the bandpass filter turns into the classical center–surround organization of ganglion cell receptive fields. The overall scale of the receptive field depends on the location of the bandpass peak, and this in turn varies as a function of the overall signal to noise ratio. Atick and Redlich (1990) show that this gives a good account of the changes in receptive field organization as the retina adapts to different light intensities. These changes are reflected not only in the responses of individual ganglion cells, but also in the sensitivity of human observers, which is compared to the theory in Fig. 5.3. In a similar spirit, van Hateren (1992) has shown that the temporal filtering in fly retina



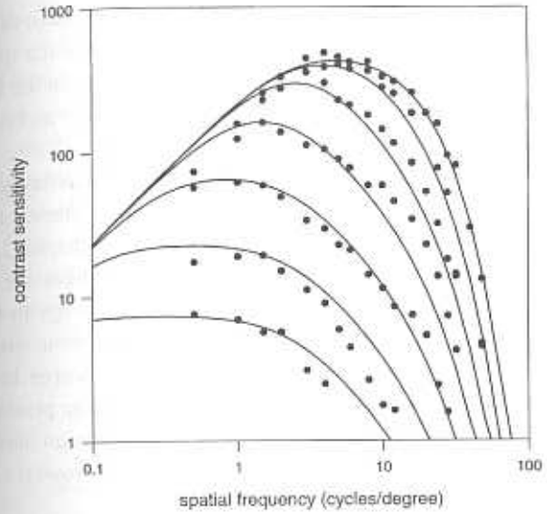



**Figure 5.3**
Contrast sensitivity at different adaptation levels, comparing human performance with the theoretically derived filters that minimize redundancy in retinal coding. Solid lines show the optimal filters at background light intensities increasing by factors of ten. The curve with the lowest sensitivity corresponds to the lowest light intensity, where we see lowpass behavior, and higher intensities correspond to higher sensitivities. Note also the bandpass behavior at high intensities. As explained in the text, this crossover from lowpass to bandpass filtering is driven by the increase of signal to noise ratio with increasing light intensity. Points are from psychophysical experiments by van Ness and Bouman (1967). Redrawn from Atick (1992).

adapts to mean light intensity as expected from signal to noise considerations for the optimal encoder.

We hope this example makes clear the structure of any optimal coding problem. One starts with a model of what the neuron can do to encode the signal—in this case just a linear filter or receptive field. Then one has to calculate various information theoretic quantities, and this requires assumptions about the statistics of the input signals and noise; in general even the qualitative features of the optimal filters can depend on the signal statistics, as in the problem discussed by Bialek, Ruderman, and Zee (1991). One would like to show that the system is optimized for signals that occur in the real world, so the model of the input statistics should capture some aspects of the natural signals. Finally, one must add some constraints to make the problem well posed. The

current state of our theoretical and experimental abilities is such that, at each step, we make approximations. Nonetheless, many features of the spatial, temporal, and chromatic (Atick, Li, and Redlich 1992) filtering in the first stages of visual processing can be understood, at least qualitatively and often semi-quantitatively, as solutions to these optimization problems.

We have emphasized the application of information theoretic optimization principles to the early stages of sensory processing, where there is now direct experimental evidence for efficient coding, as described in chapter 3. From the inception of information theory, however, there has also been the hope that these ideas could be used to describe higher level processing. In this spirit, several authors have discussed the application of optimization principles to the transmission of information between layers of cells deeper in the brain (Linsker 1990). There are also a number of algorithms for practical signal processing problems that emerge from the ideas of information maximization and redundancy reduction—see, for example, Bell and Sejnowksi (1995) and the review by Becker (1996).

We want to emphasize that theories of optimal coding do not just predict the form of receptive fields or temporal filters. The fundamental quantities in these theories are the magnitudes of the transmitted information and the redundancy among different cells. It should be clear that the development of techniques for the analysis of information transmission in arrays of neurons will make it possible to provide a much more stringent test of this theoretical point of view.

The theoretical work discussed thus far has not addressed itself to the behavior of spiking neurons. Rather than considering a cell that can transmit a continuous, filtered version of the input signals and noise, one could consider instead a cell that produces a spike each time this filtered waveform crosses a threshold. The first problem is to calculate the rate at which the resulting spike train can provide information about the input signal, and this is far from trivial. DeWeese (1995) has developed a general perturbation theory approach to this problem. Given the information transmission rate, one can ask again about the form of the optimal filter, but now there is also the question of the optimal setting for the threshold. If the threshold is set very high, spikes are infrequent and, by the arguments of section 2.3.1, we expect that it will be possible to decode the spike train using linear filters. Alternatively, if the optimal threshold is small, then the information is maximized just by maximizing the rate of spiking, which certainly leads toward the more classical rate coding ideas. In this broad class of models, the maximization of information transmission determines the optimal setting of the threshold, and this in turn controls the structure of the code.



DeWeese (1995) finds that there are indeed parameter regimes in which the optimal setting of the threshold produces very few spikes per characteristic time of the signal. One way of understanding this result is to think about a neuron in which spiking is a Poisson process modulated by the stimulus. We recall, as described in section 2.1.4, that a Poisson process is one in which spikes occur with some probability per unit time, $r[t; s(\tau)]$, which depends on the stimulus waveform $s(\tau)$ but not on the times of previous spikes. One can ask again how much information (in bits per second) the spike train provides about the time dependent stimulus. In general, this is very difficult to calculate, even for the Poisson model. But one can prove that the information transmitted by a Poisson neuron is *less* than a simple upper bound,

$$R_{\rm info} \leq \left\langle r[t; s(\tau)] \log_2 \left[ \frac{r[t; s(\tau)]}{\bar{r}} \right] \right\rangle \text{bits/s}, \tag{5.6}$$

where the average $\langle \cdots \rangle$ denotes an average over stimulus waveforms, and $\bar{r}$ is the average firing rate, that is, $\bar{r} = \langle r[t; s(\tau)] \rangle$. Notice that this bound is not sensitive to correlations between different times, only to the distribution of rates at one time. Thus, no matter how quickly the rates are varying, the transmitted information is never greater than that in Eq. (5.6). Different choices of time dependence, however, determine how close the neuron can get to this maximum. Indeed, the inequality is saturated, indicating maximal information transmission, precisely in the limit where the correlation time of the rate variations, and hence of the signal, is small, $\bar{r}\tau_c \to 0$. But, from our discussion in section 2.3.1, this is the limit that would guarantee linear decoding. These theoretical results raise the interesting possibility that the observation of sparse coding in the time domain is intimately connected to the observation of high information rates and coding efficiencies.

Animals are not interested in capturing and encoding information about the environment for its own sake, but rather in processing these input data to compute several very specific quantities. We emphasized in chapter 4 that the results of these specific computations can be extremely accurate, approaching the limits imposed by the signals and noise at the sensory input. We would like to have a theory of the computations required to make estimates and decisions at this limiting level of reliability.

In general, we know the answer to the problem of making optimal estimates: the brain should compute the mean value of the quantity it is trying to estimate, given the data provided by the receptor cells. The problem is to work out the form of this conditional mean in some cases of interest, and ask if the brain

really does this computation. We emphasize that, at least in principle, this idea of optimal estimation provides a parameter free prediction of what the nervous system should compute. The problem is that we don't really know how to calculate the optimal estimator for most interesting tasks, and again part of the problem is that the form of the optimal estimator depends on the distribution from which signals are drawn.

We reviewed in section 4.1.3 the theory of optimal pitch estimation, and here we emphasize that the theory makes nontrivial (and correct) predictions about the pitch of inharmonic signals (de Boer 1976). Human vision also provides several examples where the theory of optimal processing yields successful predictions of interesting features in our perceptions. These examples range from the detection of density variations and symmetry in random dot patterns (Barlow 1980) to aspects of three dimensional object recognition (Blake, Bulthoff, and Sheinberg 1993; Liu, Knill, and Kersten 1995), and even include the seemingly paradoxical case of ambiguous percepts (Bialek and DeWeese 1995). But these tests of the theory relate to the behavior of the entire human observer. Can we construct theories of optimal estimation that make predictions about the responses of individual neurons?

Returning to the problem of photon counting in the dark adapted visual system, we can ask whether we can understand processing strategies used in the retina given that this processing adds little or no noise to the signals present in the photoreceptors. The problem is to estimate the photon arrival rate, or, more generally, a functional of the photon arrival rate, from the currents in the photoreceptors. In the limit of low photon flux, all such estimation tasks share a common stage—a filtering step matched to the signal and noise spectra of the photoreceptors. It seems that such a universal stage of visual processing should occur early, perhaps in the transfer of photoreceptor signals to the second order cells such as bipolar cells. The hypothesis that the rod–bipolar transfer function implements such a matched filter allows successful, parameter free prediction of the bipolar cell's response to a dim flash (Bialek and Owen 1990; Rieke, Owen, and Bialek 1991).

The problem of counting photons is a very simple one, not involving the extraction of complex features in the image. The problem of estimating motion, which the fly seems to solve optimally as well, is a bit richer. Potters and Bialek (1994) discussed the theory of optimal motion estimation in a context motivated by the fly experiments and found that, as mentioned in section 4.3.3, there are limits in which the optimal estimate of motion across the visual field is based simply on the computation of delayed correlations among neighboring photoreceptors. But the problem has a structure beyond this simple limit.



More generally, because of Bayes' rule, we know that our best estimate of what is happening in the world combines the data that we receive from our sensory receptors with our prior knowledge about what to expect. It is not just that the optimal computation reaches different answers in different sensory environments, even the structure of the computation itself is different in different environments. Similarly, the maximally informative encoding of incoming signals also depends on the statistical structure of the ensemble from which these signals are drawn, as discussed for spiking neurons by DeWeese (1995).

If the sensory world were a simple place, completely characterized by a few low order statistical properties, then the matching of computational strategies to the stimulus ensemble, as required for optimal performance, could be achieved on evolutionary time scales. But we have seen, for example in the analysis of natural images (section 5.2), that the world does not have such a simple statistical structure. While the local structure of an image might be described with a simple statistical model, the parameters of this model fluctuate as we move through different regions of the scene, and these fluctuations have the same sorts of scale invariant structures found in the original image (Ruderman 1993). In such an inhomogeneous world, the optimal strategy for processing or encoding a small region of an image depends on the structure of the image on larger scales, and one can make a similar argument in the time domain—the optimal strategy for estimating motion, for example, on the short time scales of relevance to fly behavior will depend on features of the visual signal that can be measured only on longer time scales. Thus, given the complexity of the real world, optimal coding and computation are necessarily *adaptive* processes.

The word "adaptation" has multiple meanings. For neurobiology, the first meaning refers to the adaptation that Adrian discovered (Fig. 1.4), in which the response of a sensory neuron gradually fades away as stimuli are kept constant. A second meaning denotes the adaptation that is so difficult to characterize with Wiener kernels (section 2.1.3), in which the dynamics of neural responses to small signals depend strongly on the level of a large, constant background signal. There is a gap, however, between these notions of adaptation, describing the phenomenology of individual neurons, and the adaptation of organisms in the parlance of evolutionary biology. The theory of optimal coding and computation suggests a bridge between these different levels—adaptation in single neurons and circuits is a mechanism allowing maximally efficient use of resources for the crucial tasks of capturing and processing sensory information, and this can be viewed as one component of the organism's adaptation to its environment. In the same way that evolutionary adaptation

involves interactions between the organism and *all* the parameters of the environment, optimal signal processing requires, in principle, adaptation to the entire probability distribution of sensory inputs.

Can the early stages of visual processing, for example, adapt to the probability distribution of images or movies projected onto the retina? As a first step one can generate movies with Gaussian statistics, and ask if neural responses adapt to the variance, the correlation time, or the correlation length of contrast fluctuations in these movies. If the movies include true motion, simulating either the movement of objects or of the organism itself, then we can also ask about adaptation to the mean, variance, and correlation time of the motion velocity. Experiments of this type have been done in the fly's visual system, probing the responses of the motion sensitive neuron H1 (Zaagman, Mastebroek, and de Ruyter van Steveninck 1983; Maddess and Laughlin 1985; de Ruyter van Steveninck, Zaagman, and Mastebroek 1986; de Ruyter van Steveninck et al. 1996), and in the vertebrate retina, probing the responses of the ganglion cells (Smirnakis et al. 1995, 1996). Both systems provide clear evidence of adaptation to statistics, with the neural circuitry adjusting both its sensitivity and the time constants of its response to the statistical structure of the movie seen over the past few seconds.

The idea that the computations carried out by individual neurons reflect some form of optimal design is really an extension to the central nervous system of ideas that were applied long ago to the compound eye, as emphasized by Barlow (1981). Already in the nineteenth century, Mallock (1894) realized that, because of the small facet lenses in the compound eye, diffraction is a much more serious limit to insect vision than to our own vision. Barlow (1952) argued that the sizes of the lenses should be chosen to maximize the angular resolution of the system, and this predicts a simple scaling of the lens size as the square root of the size of the insect's head; this relation fits very well with data from a variety of insect species. Barlow's argument seems to have been rediscovered by Feynman, who made it part of his undergraduate lectures on physics (Feynman, Leighton, and Sands 1963). As in the analysis of neural computation, the proper formulation of the optimization problem for compound eyes requires that we take account of the sensory environment, and Snyder, Stavenga, and Laughlin (1977) showed how one could give a more general information theoretic approach to eye design, taking into account not only diffraction but also photon shot noise and image statistics. In a somewhat similar spirit, Barlow (1982) has discussed the origins of trichromatic vision as an optimal strategy for dealing with naturally occurring reflectance spectra, and Chittka and Menzel (1992) show that the particular choice of trichromatic



receptors made by bees and related insects serves to maximize the information gathered about floral identity.

It is somehow not surprising to learn that the design of the eye's optics is determined by very basic physical considerations. The fact that individual photoreceptors count single photons tells us that the physics of signals and noise in the molecular amplification processes of phototransduction must be taken very seriously in any attempt to understand the biological function of these cells. There is a vague feeling that as we move deeper into the nervous system these physical considerations become less important and some more uniquely biological considerations must dominate. But we have seen that primary sensory neurons can transmit information at rates very close to the physical limits set by spike train entropy, and that central neurons can give responses that are nearly as reliable as possible given the signals and noise at the receptor cells. It would seem that, at least for some tasks, nature has built computing machinery of surprising precision and adaptability.

# Epilogue
## Homage to the single spike

When we set out to write this book we had in mind several disparate ideas that we hoped to communicate to a more general audience. What emerged as we wrote was a surprising convergence toward one simple idea: Individual spikes are important. In the billions of neurons that are active as you read this text, each firing perhaps tens of spikes per second, it is difficult to believe that one spike more or less could matter. Yet we have seen that, under many conditions, behavioral decisions are made with of order one spike per cell (chapter 2), that individual spikes can convey several bits of information about incoming sensory stimuli (chapter 3), and that precise discriminations could, at least in principle, be based on the occurrence of individual spikes or spike pairs at definite times (chapter 4). These different results encourage us to take seriously the possibility that each spike that streams into our brain really does make a difference.

Long ago, Valbo and colleagues set out to correlate the human perception of touch with the activity of skin receptor afferents. The beauty of these experiments was that the activity of individual afferents could be recorded by placing a fine needle in the arm of an alert human subject. Although there are many issues concerning the interpretation of these data, there is strong evidence that the threshold of touch sensation for the observer is very close to the thresholds of individual afferents. But, more importantly, there is a trial by trial correlation between the presence or absence of a single action potential and the response of the observer (Valbo 1995). These experiments come very close to a direct demonstration that we can "feel" individual spikes.

Our story began, more or less, with Adrian's discovery that spikes are the units out of which our perceptions must be built. We end with the idea that each one of these units makes a definite and measurable contribution to those perceptions. The individual spike, so often averaged in with its neighbors, deserves more respect.

# Appendix
# Mathematical asides

Here we collect some of the mathematical details that are necessary to derive the results in the main text, as well as brief discussions of a few conceptual points. We try to give a fair bit of explanation about what is going on in the various manipulations, particularly when new tricks are introduced, minimizing the number of skipped steps at the cost of extra pages. We go to some lengths (quite literally) so that we never say "it can be shown that . . . ". On the other hand, many of the same mathematical tricks appear again and again, and we hope that by the last few asides the reader can begin to fill in the details if the ingredients are made clear.

## A.1 RATES AS EXPECTATION VALUES

Let us see how the empirically defined rate versus time, $r(t)$, relates to the probability distribution $P[\{t_i\}|s(\tau)]$. In analyzing real data we divide time into bins, each of width $\Delta\tau$, and we assume that these bins are sufficiently small (smaller than the refractory period) that we never observe more than one spike in each bin. Thus, if we look at a bin centered on time $t$, we can define a function $n(t)$ that is $n = 1$ if there is a spike in the bin and $n = 0$ if there is not.

Suppose we define a function $f(x)$ that is equal to zero if the magnitude of $x$ is larger than one half, and equal to one if the magnitude of $x$ is less than one half. That is,

$$f(x) = 1 \quad \text{if} \quad -1/2 \le x \le 1/2, \tag{A.1}$$

$$f(x) = 0 \quad \text{if} \quad x < -1/2 \text{ or } 1/2 < x. \tag{A.2}$$

Now if we look at the arrival time of one particular spike, $t_i$, we can evaluate $f[(t - t_i)/\Delta\tau]$, and this will count whether that particular spike is in a bin of

size $\Delta\tau$ centered on time $t$. To find whether *any* spike is in this bin we have to sum over all possible spikes, so the function $n(t)$ that counts spikes in bins can be written as

$$n(t) = \sum_i f\left[\frac{t - t_i}{\Delta\tau}\right]. \tag{A.3}$$

Again, the function $n(t)$ is what we get by counting spikes in the obvious way; the expression in terms of $f$ and the $\{t_i\}$ is a more formal mathematical description.

If we repeat the stimulus many times then we can find the average value of $n(t)$, which we have called $p(t)$ in the text. This is the *probability* that a spike will occur in a bin of width $\Delta\tau$ surrounding the time $t$, given that we have presented a particular stimulus $s(t)$. If we divide this probability by the size of the bins, then we obtain a probability per unit time, the firing rate. Strictly speaking, we want to define the rate versus time as the result of this procedure carried out with arbitrarily small bins. To describe this procedure with an equation, the rate is defined by averaging $n(t)$, normalizing to the bin size, and then letting the bin size go to zero:

$$r[t; s(\tau)] \equiv \lim_{\Delta\tau\to 0} \frac{1}{\Delta\tau}\langle n(t)\rangle. \tag{A.4}$$

But now we can substitute our formal expression for $n(t)$ in terms of the spike arrival times, Eq. (A.3), to rewrite the rate as an average over these times:

$$\begin{aligned} r[t; s(\tau)] &\equiv \lim_{\Delta\tau\to 0} \frac{1}{\Delta\tau}\langle n(t)\rangle, \\ &= \lim_{\Delta\tau\to 0} \frac{1}{\Delta\tau}\left\langle \sum_i f\left[\frac{t - t_i}{\Delta\tau}\right]\right\rangle, \qquad \text{(A.5)} \\ &= \left\langle \sum_i \lim_{\Delta\tau\to 0} \frac{1}{\Delta\tau} f\left[\frac{t - t_i}{\Delta\tau}\right]\right\rangle. \qquad \text{(A.6)} \end{aligned}$$

So it seems that the firing rate is naturally related to a slightly funny object, a function

$$\lim_{\Delta\tau\to 0} \frac{1}{\Delta\tau} f\left[\frac{t - t_i}{\Delta\tau}\right]. \tag{A.7}$$

This function has some interesting properties. First of all, since we are taking the bin size to zero, it is clear that the function must equal zero unless the spike



time $t_i$ is exactly the time $t$. On the other hand, at the point where $t_i = t$ the function is infinite because $f(0) = 1$ and we are dividing by $\Delta\tau \to 0$. Finally, if we are careful in applying the definition of $f$ given above, then we can show that this function has an integral over time of precisely one. That is,

$$\int_{-\infty}^{\infty} dt \lim_{\Delta\tau\to 0} \frac{1}{\Delta\tau} f\left[\frac{t - t_i}{\Delta\tau}\right] = 1; \tag{A.8}$$

we leave this calculation to the reader. The function with these three properties has a name: It is the "delta function" introduced by Dirac. It is obviously a bit strange, since it's value is either zero or infinity, so from a rigorous point of view it is not an ordinary function. But, as emphasized by Lighthill (1958), one can construct a rigorous theory along the lines used here, defining the delta function and other "generalized functions" as the limit of a sequence of functions where some parameter is taken to zero (in this case the bin width). Lighthill's discussion is exceptionally clear and concise, and has the added virtue of treating concepts from Fourier analysis that will be useful in connection with later chapters of the text.

The Dirac delta function $\delta(t)$ is defined as having the following properties:

$$\delta(t) = 0 \quad t \neq 0 \tag{A.9}$$

$$\int_{-\infty}^{\infty} dt\, \delta(t) = 1. \tag{A.10}$$

Then the funny object that arose in our discussion of the firing rate can be written in much more compact form:

$$\lim_{\Delta\tau\to 0} \frac{1}{\Delta\tau} f\left[\frac{t - t_i}{\Delta\tau}\right] = \delta(t - t_i). \tag{A.11}$$

This means that we can rewrite the firing rate itself:

$$\begin{aligned} r[t; s(\tau)] &= \left\langle \sum_i \lim_{\Delta\tau\to 0} \frac{1}{\Delta\tau} f\left[\frac{t - t_i}{\Delta\tau}\right]\right\rangle \\ &= \left\langle \sum_i \delta(t - t_i)\right\rangle. \qquad \text{(A.12)} \end{aligned}$$

This is a very simple expression, telling us that the firing rate is an average over a set of delta functions that pick out the times when the spikes arrive. The delta functions themselves are singular objects, being either zero or infinity, but the average of these functions is a perfectly reasonable function of time.

In Eq. (A.12) we have written the rate as an average quantity, formalizing (as emphasized in the text) the fact that rate is defined as an average over many presentations of the same stimulus. Let us flesh this out a little more. Each time we present the stimulus $s(\tau)$ we see slightly different values for the spike arrival times $\{t_i\}$. Thus these times are random variables, and when we write an average in Eq. (A.12) it is this randomness that we mean to average over. But this randomness is described by a probability distribution, the distribution of spike arrival times given the stimulus waveform, $P[\{t_i\}|s(\tau)]$. If we imagine that our experiment covers a time window from $t = 0$ to $t = T$, then when we average over the arrival times of the spike we need to integrate over all the individual spike arrival times $t_1, t_2, \cdots, t_N$ and sum over all spike counts in the window $0 < t < T$, being careful to weight each of these possibilities by the distribution $P[\{t_i\}|s(\tau)]$. Thus

$$r[t;s(\tau)] = \left\langle \sum_i \delta(t - t_i) \right\rangle$$
$$= \sum_N \int_0^T dt_1 \int_0^T dt_2 \cdots \int_0^T dt_N P[\{t_i\}|s(\tau)] \sum_{i=1}^N \delta(t - t_i). \tag{A.13}$$

The spike train is a set of pulses that occur exactly at times $t_i$. This stream of pulses can be described as the sum of delta functions, which we will call $\rho(t)$,

$$\rho(t) = \sum_i \delta(t - t_i). \tag{A.14}$$

The whole point of the discussion above has been to show that the firing rate is just the average spike train,

$$r[t;s(\tau)] = \langle \rho(t) \rangle, \tag{A.15}$$

where the average is taken with respect to the distribution $P[\{t_i\}|s(\tau)]$. We emphasize that this formulation in terms of $\rho(t)$, which is extremely compact and useful in subsequent mathematical discussions, is a bit tricky. We started our description of the spike train with a function $n(t)$ that is obtained from real data in a straightforward fashion by counting spikes in bins of finite size $\Delta\tau$, and $\rho(t)$ is related to the limiting behavior of this counting function as the bins shrink to zero size ($\Delta\tau \to 0$). In analyzing real experiments we will always need to keep some finite $\Delta\tau$, and it can be a delicate problem to decide how small we can make the bins given a data set of fixed size. As an example, if we



compute the average in Eq. (A.15) as an average over $K$ repeated presentations of the stimulus $s(t)$, it is clear that we will not recover a smooth function $r[t;s(\tau)]$ unless we keep $\Delta\tau$ at some small but nonzero value; we can accept smaller values of $\Delta\tau$ at larger $K$, but we can never, in practice, let $\Delta\tau$ shrink all the way to zero.

## A.2 TWO-POINT FUNCTIONS

In the same way that we define the rate as the average of the spike train, as in Eq. (A.15), we can consider averages of products of spike trains, such as

$$C(t,t') = \langle \rho(t)\rho(t') \rangle. \tag{A.16}$$

From the definition of $\rho(t)$ we can see that

$$C(t,t') = P(\text{spike at time } t \text{ \& spike at time } t') \tag{A.17}$$
$$= P(\text{spike at time } t|\text{spike at time } t') \times P(\text{spike at time } t'). \tag{A.18}$$

But the second term in this equation, the probability that a spike is fired at time $t'$, is just the firing rate $r(t')$. The conditional probability

$$P(\text{spike at time } t|\text{spike at time } t')$$

can be thought of as a conditional rate, that is the rate of spiking at time $t$ given that there was a spike at time $t'$. When we define the average in Eq. (A.16), we can choose to average over any dependence on the absolute time, for example by choosing stimuli from a stationary probability distribution (see section 3.1.4). Then the conditional rate, which is often called simply the correlation function of the spike train, depends only on the time difference $t - t'$, and the probability of spiking is just the average rate $\bar{r}$.

When we talk about correlations in the spike train, as in the correlation function of Fig. 2.5, it seems natural to hope that these correlations are characterized by a dimensionless number that measures the strength of the correlations. Thus, if we have two random variables $x$ and $y$, we know that we can define their correlation coefficient and that it is a pure number independent of the units in which we measure $x$ and $y$. Can we do the same thing for spike trains?

When the times $t$ and $t'$ that enter the correlation function are very far apart, we expect that the correlations between spikes become small, essentially because the system must, after some time, forget that a spike was fired. This gives us a plot as in Fig. 2.5, where the conditional rate decays as $|t - t'|$

becomes larger, reaching a plateau at the mean rate itself. At short times the conditional rate is small, because cells cannot produce spikes in rapid succession; this is called *refractoriness*. The area of this lobe in the correlation function has the dimensions of 1/time, or rate. This is the amount by which the mean firing rate is reduced as a result of refractoriness. Alternatively, if we look at the normalized correlation function, or conditional rate, the corresponding area is dimensionless. We can think of this area as the fraction of spikes that are "deleted" from the spike train as a result of refractoriness. Similarly, one can look at other features in the correlation function and compute their areas on the plot; each area gives a rate, or a dimensionless fraction if we analyze the conditional rates. If these rates are small compared to the mean rate of spiking, or if the fractions are small compared to 1, then correlations are weak. If the rates computed from the integrals of the correlation functions are comparable to the mean spike rate, then correlations are strong.

There is another useful interpretation of the correlation function. We recall that

$$\rho(t) = \sum_i \delta(t - t_i),$$

and that the delta function has unit area, so that

$$\int_0^T dt\, \delta(t - t_i) = 1 \quad \text{if } 0 \le t_i \le T \tag{A.19}$$

$$= 0 \quad \text{otherwise.} \tag{A.20}$$

Thus if we integrate $\rho(t)$ itself we count the spikes,

$$\int_0^T dt\, \rho(t) = N(T), \tag{A.21}$$

where $N(T)$ is the number of spikes in the time window from 0 to $T$. So if we integrate the correlation function we find

$$\int_0^T dt \int_0^T dt' C(t,t') = \int_0^T dt \int_0^T dt' \langle \rho(t)\rho(t') \rangle \tag{A.22}$$

$$= \left\langle \int_0^T dt \int_0^T dt' \rho(t)\rho(t') \right\rangle \tag{A.23}$$

$$= \left\langle \left[ \int_0^T dt\, \rho(t) \right] \left[ \int_0^T dt'\, \rho(t') \right] \right\rangle \tag{A.24}$$

$$= \langle [N(T)]^2 \rangle, \tag{A.25}$$



the mean square spike count. But the mean spike count is the time integral of the firing rate, which is in turn the average of $\rho(t)$. Thus we can write the variance of the spike count in terms of averages of $\rho(t)$:

$$\langle [\delta N(T)]^2 \rangle = \langle [N(T)]^2 \rangle - \langle N(T) \rangle^2 \tag{A.26}$$

$$= \int_0^T dt \int_0^T dt' C(t,t') - \left[ \left\langle \int_0^T dt \rho(t) \right\rangle \right]^2 \tag{A.27}$$

$$= \int_0^T dt \int_0^T dt' \langle \rho(t)\rho(t') \rangle - \int_0^T dt \int_0^T dt' \langle \rho(t) \rangle \langle \rho(t') \rangle \tag{A.28}$$

$$= \int_0^T dt \int_0^T dt' \left\langle [\rho(t) - \langle \rho(t) \rangle][\rho(t') - \langle \rho(t') \rangle] \right\rangle \tag{A.29}$$

$$= \int_0^T dt \int_0^T dt' \langle \delta\rho(t)\delta\rho(t') \rangle. \tag{A.30}$$

Thus the variance in spike counts is related to the integral of the correlation function $\langle \delta\rho(t)\delta\rho(t') \rangle$ of fluctuations in $\rho(t)$.

As we suggest in the discussion above, and as we explain more fully in section 3.1.4, correlation functions measure the "memory" that a system has for its previous states, and we expect that this memory will decay. This means that $\langle \delta\rho(t)\delta\rho(t') \rangle$ will be small as $t$ and $t'$ become widely separated, and when we do the integral in Eq. (A.30) for large values of $T$ most of the area in the region of integration will have the integrand near zero. Thus when we do the two dimensional integral over $t$ and $t'$, it is convenient to change coordinates and integrate over time differences $t - t' = \tau$ and the average time $(t + t')/2 = \bar{t}$:

$$\langle [\delta N(T)]^2 \rangle = \int_0^T dt \int_0^T dt' \langle \delta\rho(t)\delta\rho(t') \rangle$$

$$= \int_0^T d\bar{t} \int_{-2\bar{t}}^{2\bar{t}} d\tau \langle \delta\rho(\bar{t} + \tau/2)\delta\rho(\bar{t} - \tau/2) \rangle \tag{A.31}$$

$$\approx \int_0^T d\bar{t} \int_{-\infty}^{\infty} d\tau \langle \delta\rho(\bar{t} + \tau/2)\delta\rho(\bar{t} - \tau/2) \rangle, \tag{A.32}$$

where in the last step we use the fact that $\bar{t}$ is typically $\sim T/2$ and hence the integral over $\tau$ ranges over large values that completely cover the "memory"

of the correlation function. Finally, we can do the integral over $\bar{t}$ to pick up a factor of $T$, so that

$$\langle [\delta N(T)]^2 \rangle = DT \tag{A.33}$$

$$D = \int_{-\infty}^{\infty} d\tau \langle \delta\rho(\bar{t} + \tau/2)\delta\rho(\bar{t} - \tau/2)\rangle. \tag{A.34}$$

Notice that the variance in spike count grows in proportion to the time, in the same way that the mean square displacement grows as time for a diffusing particle, provided that the "diffusion constant" $D$ is finite. If the correlation function has long tails, then the integral that defines $D$ need not converge and the variance in spike count can grow as a different power of $T$ (Teich 1989).

Notice that, with well behaved correlations, the variance of the spike count in arbitrarily large windows $T$ is determined by an integral of the correlation function which has structure only on much shorter time scales. Again this is analogous to diffusion, where the mean square displacement of a particle moving through a fluid is determined by the statistics of collisions that last only a few picoseconds.

All of the discussion we have given here for correlations among spikes in one spike train can be repeated for correlations among spikes generated by two cells. We can define the functions $\rho_1(t)$ and $\rho_2(t)$ for the two cells, or more generally $\rho_i(t)$ and $\rho_j(t)$ for any two cells in a group. The natural definition of the crosscorrelation function is $\langle \rho_i(t)\rho_j(t')\rangle$, or we could subtract off the mean values to define $\langle \delta\rho_i(t)\delta\rho_j(t')\rangle$. The quantities that are called the cross-correlation function in the literature are all related to these objects, normalized in different ways.[1] In particular there is an analog of Eq. (A.33), which relates the covariance in spike counts to an integral of the crosscorrelation function,

$$\langle \delta N_i(T)\delta N_j(T)\rangle = D_{ij}T \tag{A.35}$$

$$D_{ij} = \int_{-\infty}^{\infty} d\tau \langle \delta\rho_i(\bar{t} + \tau/2)\delta\rho_j(\bar{t} - \tau/2)\rangle. \tag{A.36}$$

Again the important consequence of this equation is that correlations among spike counts in very large windows are determined by the crosscorrelation function of the spike trains, and these correlation functions typically have

1. The question of which normalization is best inspires passionate debate, and we are inclined to avoid the issue since it is a bit peripheral to our discussion. We do remark that many of the difficulties can be avoided by being careful to report the correlation functions with their correct units, whatever normalization procedure you choose.



structure on short time scales. This idea was used by Bair, Zohary, and Koch (1996) to analyze the correlation among MT neurons observed by Zohary, Shadlen, and Newsome (1994) and discussed in section 4.1.4. They find that the correlation functions have memory over time scales that are, on average, shorter than 10 ms, but that spike count correlations observed in 2 s windows are nonetheless consistent with predictions from Eq. (A.35).

## A.3 WIENER KERNELS

Methods developed by Volterra (1930) and Wiener (1958) provide a systematic characterization of a nonlinear system in terms of a set of filters or kernels. We recall, as explained in the text, that when we look at a function of one number $x$, we can expand the function $y = f(x)$ in a series of powers,

$$y = f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2}f''(x_0)(x - x_0)^2 + \cdots \tag{A.37}$$

$$= f_0 + f_1(x - x_0) + f_2(x - x_0)^2 + \cdots; \tag{A.38}$$

this is the Taylor series, Eq. (2.6). We want to generalize this to the case we take as "input" not a single number $x$ but a function of time $x(t)$, and similarly the "output" is not a single number $y$ but another function of time $y(t)$. We write, symbolically, the analogy to $y = f(x)$,

$$y(t) = F[x(t)], \tag{A.39}$$

and the transformation $F[x(t)]$ is called a *functional*. The Volterra series is the analog of the Taylor series:

$$\begin{aligned} y(t) = h_0 &+ \int d\tau_1 h_1(\tau)x(t - \tau_1) \\ &+ \int d\tau_1 \int d\tau_2 h_2(\tau_1, \tau_2)x(t - \tau_1)x(t - \tau_2) \\ &+ \int d\tau_1 \int d\tau_2 \int d\tau_3 h_3(\tau_1, \tau_2, \tau_3)x(t - \tau_1)x(t - \tau_2)x(t - \tau_3) \\ &+ \cdots, \end{aligned} \tag{A.40}$$

where the $h_n$ are called the Volterra kernels. The functions $h_n$ act as expansion coefficients to describe the transformation from input $x(t)$ to output $y(t)$ in the same way that the coefficients $f_n$ describe the mapping from $x$ to $y$. Notice that now the expansion coefficients themselves are functions rather than numbers. This description can be extended to multiple inputs, and to the case

where the input at each instant of time is itself a function of spatial variables, as in vision; see, for example, the discussion by Poggio and Reichardt (1976). There are theorems (Volterra 1930) which guarantee that, under certain conditions, the proper choice of the kernels $h_n$ will provide a complete description of any transformation $x(t) \to y(t)$.

In principle the integrals in Eq. (A.40) should range over all possible values of the time variables $\tau_1, \tau_2, \cdots$. In practice we are interested in using this approach to characterize real physical devices, in which the output can depend only on inputs that have already occurred—this is causality, the statement that your current state depends on events in the past, but not on events in the future. From causality, then, $y(t)$ can be written in terms of $x(t-\tau)$ using only positive values of $\tau$, and hence the integrals can always be restricted to positive $\tau$,

$$\begin{aligned} y(t) = h_0 &+ \int_0^\infty d\tau_1\, h_1(\tau) x(t-\tau_1) \\ &+ \int_0^\infty d\tau_1 \int_0^\infty d\tau_2\, h_2(\tau_1,\tau_2) x(t-\tau_1) x(t-\tau_2) \\ &+ \int_0^\infty d\tau_1 \int_0^\infty d\tau_2 \int_0^\infty d\tau_3\, h_3(\tau_1,\tau_2,\tau_3) \\ &\quad \times x(t-\tau_1) x(t-\tau_2) x(t-\tau_3) \\ &+ \cdots . \end{aligned} \tag{A.41}$$

Notice that, since the functions $h_n$ appear only under integral signs, we can always choose these functions to be symmetric when we interchange the values of the different $\tau_i$, that is,

$$h_2(\tau_1,\tau_2) = h_2(\tau_2,\tau_1) \tag{A.42}$$

$$h_3(\tau_1,\tau_2,\tau_3) = h_3(\tau_2,\tau_3,\tau_1) \tag{A.43}$$

$$= h_3(\tau_1,\tau_3,\tau_2) \tag{A.44}$$

$$= h_3(\tau_3,\tau_2,\tau_1), \tag{A.45}$$

and so on.

Wiener reformulated Volterra's expansion in a way that makes clear how one can measure the individual coefficients. He imagined that "input" functions $x(t)$ would be drawn from a probability distribution, so that each term in the series would become a random variable. Then, one could simplify matters by choosing an expansion in which different terms are statistically independent. In principle one can carry through this procedure (sometimes called "orthogonalization" of the terms) for any probability distribution $P[x(t)]$, but the simplest case is where $x(t)$ is Gaussian white noise. For Gaussian white noise, the average value of $x(t)$ is zero, and all odd order correlation functions vanish, that is,



$$\langle x(t) \rangle = 0, \tag{A.46}$$

$$\langle x(t_1) x(t_2) x(t_3) \rangle = 0, \tag{A.47}$$

$$\langle x(t_1) x(t_2) x(t_3) x(t_4) x(t_5) \rangle = 0, \tag{A.48}$$

$$\cdots .$$

The "white" description means that all frequency components are present with equal strength (the power spectrum is independent of frequency; see section 3.1.4), and in the time domain this means that correlations extend only over vanishingly short times. Thus, the usual correlation function—which we should now be careful to call the two-point correlation function, for obvious reasons—has the form

$$\langle x(t) x(t') \rangle = S_x \delta(t - t'), \tag{A.49}$$

where $S_x$ is the power spectrum of $x$ and the delta function is discussed in section A.1. All of the higher, even order correlation functions have the same feature of being nonzero only when the different time variables are equal to one another, and one can find these correlation functions by enumerating all possible ways of pairing up the factors of $x(t)$. For example, the four point correlation function has the form

$$\begin{aligned} \langle x(t_1) x(t_2) x(t_3) x(t_4) \rangle &= \langle x(t_1) x(t_2) \rangle \langle x(t_3) x(t_4) \rangle \\ &\quad + \langle x(t_1) x(t_3) \rangle \langle x(t_2) x(t_4) \rangle \\ &\quad + \langle x(t_1) x(t_4) \rangle \langle x(t_2) x(t_3) \rangle \end{aligned} \tag{A.50}$$

$$\begin{aligned} &= S_x^2 \delta(t_1 - t_2) \delta(t_3 - t_4) \\ &\quad + S_x^2 \delta(t_1 - t_3) \delta(t_2 - t_4) \\ &\quad + S_x^2 \delta(t_1 - t_4) \delta(t_2 - t_3). \end{aligned} \tag{A.51}$$

We can use these definitions of Gaussian white noise to analyze the Volterra expansion in Eq. (A.41). We notice, for example, that the term with two factors of $x(t)$—that is, the term $\propto h_2(\tau_1,\tau_2)$—has a nonzero average, which we could have chosen to be part of the constant term $h_0$. Similarly, the term with three factors of $x(t)$—the term $\propto h_3(\tau_1,\tau_2,\tau_3)$—is correlated with the term

that has only one factor of $x(t)$, and so on. The Wiener series removes all these correlations by writing

$$y(t) = G_0 + G_1[x(t)] + G_2[x(t)] + G_3[x(t)] + \cdots, \tag{A.52}$$

where the individual terms are defined by

$$G_0 = g_0 \tag{A.53}$$

$$G_1[x(t)] = \int_0^\infty d\tau_1\, g_1(\tau_1) x(t-\tau_1) \tag{A.54}$$

$$\begin{aligned} G_2[x(t)] = &\int_0^\infty d\tau_1 \int_0^\infty d\tau_2\, g_2(\tau_1,\tau_2) x(t-\tau_1)x(t-\tau_2) \\ &- S_x \int_0^\infty d\tau\, g_2(\tau,\tau) \end{aligned} \tag{A.55}$$

$$\begin{aligned} G_3[x(t)] = &\int_0^\infty d\tau_1 \int_0^\infty d\tau_2 \int_0^\infty d\tau_3\, g_3(\tau_1,\tau_2,\tau_3) \\ &\times x(t-\tau_1)x(t-\tau_2)x(t-\tau_3) \\ &- 3S_x \int d\tau_1 d\tau_2\, g_3(\tau_1,\tau_1,\tau_2) x(t-\tau_2), \end{aligned} \tag{A.56}$$

$$\cdots.$$

The coefficients of the Wiener expansion, $g_0, g_1(\tau_1), g_2(\tau_1,\tau_2), \cdots$, are called Wiener kernels. Like the Volterra coefficients $h_n$, we can choose the Wiener kernels to be causal, so that they vanish for negative values of the $\tau_i$, and to be symmetric under the interchange of the different $\tau_i$. The beauty of the Wiener expansion is that successive terms in the expansion are independent; the pieces subtracted from each term in Eq. (A.3) impose this independence for the case of Gaussian white noise inputs. The independence of each term in the series means that we can measure the terms individually—unlike the usual situation in fitting a polynomial to set of data points, for example, we can determine the correct coefficient of the first order term without knowing in advance whether we will need to use a second or a third order term to provide a better fit.

The kernels can be measured by correlating the output $y(t)$ with the successive powers of the white noise input $x(t)$. For example, correlating $y(t)$ with $x(t)$ we obtain

$$\begin{aligned} \langle y(t)x(t-\tau)\rangle = &\langle x(t-\tau)G_0\rangle + \langle x(t-\tau)G_1[x(t)]\rangle \\ &+ \langle x(t-\tau)G_2[x(t)]\rangle + \langle x(t-\tau)G_3[x(t)]\rangle + \cdots, \end{aligned} \tag{A.57}$$



and we can evaluate each of the terms individually:

$$\langle x(t-\tau)G_0\rangle = \langle x(t-\tau)g_0\rangle \tag{A.58}$$

$$= 0, \tag{A.59}$$

$$\langle x(t-\tau)G_1[x(t)]\rangle = \int_0^\infty d\tau_1\, g_1(\tau_1)\langle x(t-\tau_1)x(t-\tau)\rangle \tag{A.60}$$

$$= \int_0^\infty d\tau_1\, g_1(\tau_1) S_x \delta(\tau-\tau_1) \tag{A.61}$$

$$= S_x g_1(\tau), \tag{A.62}$$

$$\begin{aligned} \langle x(t-\tau)G_2[x(t)]\rangle = &\int_0^\infty d\tau_1 \int_0^\infty d\tau_2\, g_2(\tau_1,\tau_2)\langle x(t-\tau_1)x(t-\tau_2)x(t-\tau)\rangle \\ &- S_x \int_0^\infty d\tau_1\, g_2(\tau_1,\tau_1)\langle x(t-\tau)\rangle \end{aligned} \tag{A.63}$$

$$= 0, \tag{A.64}$$

$$\begin{aligned} \langle x(t-\tau)G_3[x(t)]\rangle = &\int_0^\infty d\tau_1 \int_0^\infty d\tau_2 \int_0^\infty d\tau_3\, g_3(\tau_1,\tau_2,\tau_3) \\ &\times \langle x(t-\tau_1)x(t-\tau_2)x(t-\tau_3)x(t-\tau)\rangle \\ &- 3S_x \int_0^\infty d\tau_1 \int_0^\infty d\tau_2\, g_3(\tau_1,\tau_1,\tau_2)\langle x(t-\tau_2)x(t-\tau)\rangle \end{aligned} \tag{A.65}$$

$$\begin{aligned} = &\int_0^\infty d\tau_1 \int_0^\infty d\tau_2 \int_0^\infty d\tau_3\, g_3(\tau_1,\tau_2,\tau_3) S_x^2 \\ &\times [S_x^2\delta(\tau_1-\tau_2)\delta(\tau_3-\tau) + S_x^2\delta(\tau_1-\tau_3)\delta(\tau_2-\tau) \\ &+ S_x^2\delta(\tau_1-\tau)\delta(\tau_2-\tau_3)] \\ &- 3S_x \int_0^\infty d\tau_1 \int_0^\infty d\tau_2\, g_3(\tau_1,\tau_1,\tau_2) S_x \delta(\tau-\tau_2) \end{aligned} \tag{A.66}$$

$$\begin{aligned} = S_x^2 \int_0^\infty d\tau_1 \big[&g_3(\tau_1,\tau_1,\tau) + g_3(\tau_1,\tau,\tau_1) \\ &+ g(\tau,\tau_1,\tau_1) - 3g(\tau_1,\tau_1,\tau)\big] \end{aligned} \tag{A.67}$$

$$= 0, \tag{A.68}$$

$$\cdots.$$

Thus all the terms except the one $\propto G_1[x(t)]$ vanish, and we find

$$\langle y(t)x(t-\tau)\rangle = S_x g_1(\tau). \tag{A.69}$$

Evidently correlating the output with the input isolates the first kernel, and correlation with higher powers of the input isolates higher kernels, so we can solve for the kernels in terms of these crosscorrelations:

$$g_0 = \langle y(t) \rangle, \tag{A.70}$$

$$g_1(\tau) = \frac{1}{S_x} \langle y(t) x(t-\tau) \rangle, \tag{A.71}$$

$$g_2(\tau_1, \tau_2) = \frac{1}{S_x^2} \langle y(t) x(t-\tau_1) x(t-\tau_2) \rangle, \tag{A.72}$$

$$\cdots .$$

As in the case of the Volterra series, there are theorems to guarantee that a Wiener series with sufficiently many terms provides a complete description of a broad class of nonlinear systems.

For the case of spiking neurons we take the input to be the stimulus $s(t)$ and the output to be the function $\rho(t)$ defined in section A.1; we recall that the expectation value of $\rho(t)$ is the time dependent firing rate $r[t; s(\tau)]$. If we follow the Wiener prescription summarized by Eq. (A.3), the first Wiener kernel of a spiking neuron is given by

$$g_1(\tau) = \frac{1}{S} \langle \rho(t) s(t-\tau) \rangle. \tag{A.73}$$

But $\rho(t)$ is a sum of delta functions at the times where spikes occur, as defined in Eq. (A.14), so that

$$\begin{aligned} g_1(\tau) &= \frac{1}{S} \langle \rho(t) s(t-\tau) \rangle \\ &= \frac{1}{S} \left\langle \sum_i \delta(t - t_i) s(t-\tau) \right\rangle \qquad \text{(A.74)} \\ &= \frac{1}{S} \left\langle \sum_i \delta(t - t_i) s(t_i-\tau) \right\rangle, \qquad \text{(A.75)} \end{aligned}$$

where in the last step we make use of the fact that the delta function $\delta(t - t_i)$ is zero unless $t = t_i$, so we can replace $t$ by $t_i$ in any function that multiplies the delta function. But now we recall that the Wiener kernel, which describes the response of the system, should be independent of the time $t$ at which we measure this response. If we want, then, we can integrate over $t$, which ranges from 0 to $T$ in our experiment, and then divide by $T$. This is useful because of Eq's. (A.19, A.20), which tell us that when we integrate the delta function we obtain one if the time $t_i$ is in our integration window, and zero otherwise. But



then the time integral of the sum of delta functions just counts the number of spikes in the window, as in Eq. (A.21), and in this case we take the window to be the whole experiment:

$$\begin{aligned} g_1(\tau) &= \frac{1}{S} \left\langle \sum_i \delta(t - t_i) s(t_i - \tau) \right\rangle \\ &= \frac{1}{T} \int_0^T dt \frac{1}{S} \left\langle \sum_i \delta(t - t_i) s(t_i - \tau) \right\rangle \qquad \text{(A.76)} \\ &= \frac{1}{S} \left\langle \sum_i \frac{1}{T} \int_0^T dt \delta(t - t_i) s(t_i - \tau) \right\rangle \qquad \text{(A.77)} \\ &= \frac{1}{S} \left\langle \frac{N(T)}{T} s(t_i - \tau) \right\rangle. \qquad \text{(A.78)} \end{aligned}$$

Now if the duration of our experiment is long enough, the number of spikes we observe per unit time, $N(T)/T$, will just be the average firing rate $\bar{r}$, with vanishingly small ($\propto 1/\sqrt{T}$) fluctuations. Then, finally,

$$g_1(\tau) = \frac{\bar{r}}{S} \langle s(t_i - \tau) \rangle. \tag{A.79}$$

Thus, up to normalizing factors, we see that the first Wiener kernel of a spiking neuron, $g_1(\tau)$ is exactly the average stimulus measured a time $\tau$ before the occurrence of a spike, as promised in the text (de Boer and Kuyper 1968).

## A.4 POISSON MODEL I

The Poisson model for spike firing is defined by the assumption that the occurrence of each spike is independent of all the others, given that the stimulus waveform $s(\tau)$ is fixed. This stimulus determines the firing rate as a function of time, which we write as $r[t; s(\tau)]$ to remind ourselves (as before) that it depends both on time and on the particular stimulus being presented.

As discussed in the text, the fact that each spike is independent of the others means that the probability of observing spikes at times $t_1, t_2, \cdots, t_N$ must be proportional to a product of the rates evaluated at these times, that is

$$\begin{aligned} P[\{t_i\}|s(\tau)] &\propto r[t_1; s(\tau)] r[t_2; s(\tau)] \cdots r[t_N; s(\tau)] \\ &= \prod_{i=1}^{N} r[t_i; s(\tau)]. \end{aligned} \tag{A.80}$$

But to get the exact form of the distribution we must include a factor that measures the probability of *no* spikes occurring at any other times. We recall that the probability of a spike occurring in a bin of size $\Delta\tau$ surrounding time $t$ is, by the original definition of the rate, $p(t) = r[t; s(\tau)]\Delta\tau$. Then the probability of no spike must be $1 - p(t)$. So we need to form a product of factors $1 - p(t)$ for all times not equal to the special $t_i$ where we observed spikes. Let's call this factor $F$,

$$F = \prod_{\{t_n\} \neq \{t_i\}} [1 - p(t_n)]; \tag{A.81}$$

as a shorthand we write

$$F = \prod_{n \neq i} [1 - p(t_n)]. \tag{A.82}$$

Then the probability of observing spikes in bins surrounding the $t_i$ is

$$P[\{t_i\}|s(\tau)](\Delta\tau)^N = \frac{1}{N!} F \prod_{i=1}^{N} (r[t_i; s(\tau)]\Delta\tau), \tag{A.83}$$

where the $N!$ corrects for all the different ways of assigning labels $1, 2, \cdots, N$ to the spikes we observe.

To proceed we pull out all the factors related to the $t_i$ and isolate the terms independent of these times:

$$\begin{aligned} P[\{t_i\}|s(\tau)](\Delta\tau)^N &= \frac{1}{N!} F \prod_{i=1}^{N} (r[t_i; s(\tau)]\Delta\tau) \\ &= \frac{1}{N!} \prod_{n \neq i} [1 - p(t_n)] \times \prod_{i=1}^{N} (r[t_i; s(\tau)]\Delta\tau) \qquad \text{(A.84)} \\ &= \frac{1}{N!} \prod_{n \neq i} (1 - r[t_n; s(\tau)]\Delta\tau) \\ &\quad \times \prod_{i=1}^{N} (r[t_i; s(\tau)]\Delta\tau) \qquad \text{(A.85)} \\ &= \frac{1}{N!} \prod_{n} (1 - r[t_n; s(\tau)]\Delta\tau) \\ &\quad \times \prod_{i=1}^{N} \left( \frac{r[t_i; s(\tau)]\Delta\tau}{1 - r[t_i; s(\tau)]\Delta\tau} \right). \qquad \text{(A.86)} \end{aligned}$$



where we must be careful to remember that $\prod_n$ denotes a product over *all* possible times $t_n$. To simplify this product we remember that products can be turned into sums by taking logarithms. Thus if we have a product of terms $x_1 x_2 \cdots x_N$, we can write each $x$ as the exponential of its logarithm:

$$\begin{aligned} x &= \exp(\ln x) \qquad \text{(A.87)} \\ \Rightarrow x_1 x_2 \cdots x_N &= \exp(\ln x_1) \exp(\ln x_2) \cdots \exp(\ln x_N) \qquad \text{(A.88)} \\ &= \exp(\ln x_1 + \ln x_2 + \cdots + \ln x_N), \qquad \text{(A.89)} \end{aligned}$$

where in the last step we use the fact that $\exp(A)\exp(B) = \exp(A + B)$. We are interested in using this trick on the product over all times in Eq. (A.86), that is

$$\begin{aligned} \prod_n (1 - r[t_n; s(\tau)]\Delta\tau) &= \prod_n \exp\left[\ln (1 - r[t_n; s(\tau)]\Delta\tau)\right] \qquad \text{(A.90)} \\ &= \exp\left[\sum_n \ln (1 - r[t_n; s(\tau)]\Delta\tau)\right], \qquad \text{(A.91)} \end{aligned}$$

so that when we substitute back into Eq. (A.86) we find

$$\begin{aligned} P[\{t_i\}|s(\tau)](\Delta\tau)^N &= \frac{1}{N!} \prod_n (1 - r[t_n; s(\tau)]\Delta\tau) \\ &\quad \times \prod_{i=1}^{N} \left( \frac{r[t_i; s(\tau)]\Delta\tau}{1 - r[t_i; s(\tau)]\Delta\tau} \right) \\ &= \frac{1}{N!} \exp\left[\sum_n \ln (1 - r[t_n; s(\tau)]\Delta\tau)\right] \\ &\quad \times \prod_{i=1}^{N} \left( \frac{r[t_i; s(\tau)]\Delta\tau}{1 - r[t_i; s(\tau)]\Delta\tau} \right). \qquad \text{(A.92)} \end{aligned}$$

Now $\Delta\tau$ is very small, which means that we need to take the logarithm of numbers that are almost equal to one. We recall that the log of one is zero, and the Taylor series of the logarithm in the neighborhood of one is

$$\ln(1 + x) = x - \frac{1}{2}x^2 + \frac{1}{3}x^3 - \cdots. \tag{A.93}$$

In this case we apply this expansion to

$$\ln (1 - r[t_n; s(\tau)]\Delta\tau) = -r[t_n; s(\tau)]\Delta\tau - \frac{1}{2}(r[t_n; s(\tau)]\Delta\tau)^2 + \cdots, \tag{A.94}$$

so our expression for the probability can be written as

$$P[\{t_i\}|s(\tau)](\Delta\tau)^N = \frac{1}{N!}\exp\left[\sum_n \ln\left(1 - r[t_n;s(\tau)]\Delta\tau\right)\right] \times \prod_{i=1}^{N}\left(\frac{r[t_i;s(\tau)]\Delta\tau}{1 - r[t_i;s(\tau)]\Delta\tau}\right)$$
$$= \frac{1}{N!}\exp\left[\sum_n (-r[t_n;s(\tau)]\Delta\tau) - \frac{1}{2}\sum_n (-r[t_n;s(\tau)]\Delta\tau)^2 + \cdots\right] \times \prod_{i=1}^{N}\left(\frac{r[t_i;s(\tau)]\Delta\tau}{1 - r[t_i;s(\tau)]\Delta\tau}\right). \quad \text{(A.95)}$$

What does it mean that we sum over bins, with each bin weighted by its width $\Delta\tau$? We recall that this sort of sum converges, as the bins become small, to an integral. That is,

$$\lim_{\Delta\tau\to 0}\sum_n f(t_n)\Delta\tau = \int dt\, f(t) \quad \text{(A.96)}$$

for any function $f(t)$. In the present case this means that

$$\lim_{\Delta\tau\to 0}\exp\left[\sum_n (-r[t_n;s(\tau)]\Delta\tau) - \frac{1}{2}\sum_n (-r[t_n;s(\tau)]\Delta\tau)^2 + \cdots\right]$$
$$= \exp\left[-\int dt\, r[t;s(\tau)] - \frac{1}{2}\Delta\tau\int dt\, (r[t;s(\tau)])^2 + \cdots\right]. \quad \text{(A.97)}$$

Now we notice that the second integral in the exponential has an extra factor of $\Delta\tau$, which comes from the $(\Delta\tau)^2$ in the previous expression, but if we really let $\Delta\tau$ go to zero this must be negligible as long as the rate doesn't become infinite. Similarly, we have in Eq. (A.95) factors like

$$\frac{r[t_i;s(\tau)]\Delta\tau}{1 - r[t_i;s(\tau)]\Delta\tau},$$

and again as $\Delta\tau \to 0$ we can expand this in powers of $\Delta\tau$ and drop all but the first term. This is equivalent to replacing the denominator of the fraction by 1.



So, when the dust clears, the expression for the probability of the spike arrival times becomes

$$P[\{t_i\}|s(\tau)] = \lim_{\Delta\tau\to 0}\frac{1}{(\Delta\tau)^N}\frac{1}{N!}\exp\left[\sum_n (-r[t_n;s(\tau)]\Delta\tau) - \frac{1}{2}\sum_n (-r[t_n;s(\tau)]\Delta\tau)^2 + \cdots\right] \times \prod_{i=1}^{N}\left(\frac{r[t_i;s(\tau)]\Delta\tau}{1 - r[t_i;s(\tau)]\Delta\tau}\right) \quad \text{(A.98)}$$
$$= \frac{1}{N!}\exp\left[-\int dt\, r[t;s(\tau)]\right]\prod_{i=1}^{N} r[t_i;s(\tau)]. \quad \text{(A.99)}$$

The integral over time should refer to the whole duration of our observations, which we will say ranges from $t = 0$ to $t = T$. Thus

$$P[\{t_i\}|s(\tau)] = \frac{1}{N!}\exp\left[-\int_0^T dt\, r[t;s(\tau)]\right]\prod_{i=1}^{N} r[t_i;s(\tau)], \quad \text{(A.100)}$$

as promised in the text, Eq. (2.18).

Now we indicate the steps involved in checking the normalization of the probability distribution in Eq. (A.100). We want to calculate the total probability, which involves taking the term with $N$ spikes and integrating over all $N$ arrival times, then summing on $N$:

$$\sum_{N=0}^{\infty}\int_0^T dt_1\int_0^T dt_2\cdots\int_0^T dt_N P[\{t_i\}|s(t)]$$
$$= \sum_{N=0}^{\infty}\int_0^T dt_1\int_0^T dt_2\cdots\int_0^T dt_N\frac{1}{N!}\exp\left[-\int_0^T dt r[t;s(\tau)]\right] \times \prod_{i=1}^{N} r[t_i;s(\tau)]. \quad \text{(A.101)}$$

Notice that the exponential does not depend on the $\{t_i\}$ or on $N$, so we can take it outside the sum and integral.

$$\sum_{N=0}^{\infty}\int_0^T dt_1\int_0^T dt_2\cdots\int_0^T dt_N P[\{t_i\}|s(t)]$$

$$=\sum_{N=0}^{\infty}\int_0^T dt_1\int_0^T dt_2\cdots\int_0^T dt_N\frac{1}{N!}\exp\left[-\int_0^T dt r[t;s(\tau)]\right]$$

$$\times\prod_{i=1}^{N} r[t_i;s(\tau)]$$

$$=\exp\left[-\int_0^T dt r[t;s(\tau)]\right]$$

$$\times\sum_{N=0}^{\infty}\frac{1}{N!}\int_0^T dt_1\int_0^T dt_2\cdots\int_0^T dt_N r[t_1;s(\tau)]$$

$$\times r[t_2;s(\tau)]\cdots r[t_N;s(\tau)]. \qquad (A.102)$$

Although we have to integrate over all the $N$ different $t_i$ together (an $N$ dimensional integral), we see that the integrand is just a product of terms that depend on each individual $t_i$. This means that really we have a product of $N$ one dimensional integrals:

$$\sum_{N=0}^{\infty}\int_0^T dt_1\int_0^T dt_2\cdots\int_0^T dt_N P[\{t_i\}|s(t)]$$

$$=\exp\left[-\int_0^T dt r[t;s(\tau)]\right]$$

$$\times\sum_{N=0}^{\infty}\frac{1}{N!}\int_0^T dt_1\int_0^T dt_2\cdots\int_0^T dt_N r[t_1;s(\tau)]$$

$$\times r[t_2;s(\tau)]\cdots r[t_N;s(\tau)]$$

$$=\exp\left[-\int_0^T dt r[t;s(\tau)]\right]\sum_{N=0}^{\infty}\frac{1}{N!}\int_0^T dt_1 r[t_1;s(\tau)]$$

$$\times\int_0^T dt_2 r[t_2;s(\tau)]\cdots\int_0^T dt_N r[t_N;s(\tau)]$$

$$=\exp\left[-\int_0^T dt r[t;s(\tau)]\right]\sum_{N=0}^{\infty}\frac{1}{N!}\left(\int_0^T dt r[t;s(\tau)]\right)^N. \qquad (A.103)$$



Now we have to remember that the series expansion of the exponential function is (with $0! = 1$ by definition)

$$\exp(x)=\sum_{N=0}^{\infty}\frac{1}{N!}x^N, \qquad (A.104)$$

so we can actually do the sum in Eq. (A.103):

$$\exp\left[-\int_0^T dt r[t;s(\tau)]\right]\sum_{N=0}^{\infty}\frac{1}{N!}\left(\int_0^T dt r[t;s(\tau)]\right)^N$$

$$=\exp\left[-\int_0^T dt r[t;s(\tau)]\right]\exp\left[+\int_0^T dt r[t;s(\tau)]\right]$$

$$=1, \qquad (A.105)$$

which completes our check on the normalization of the distribution in Eq. (2.18).

## A.5 POISSON MODEL II

It is a useful exercise to derive the expressions for the spike count distribution, Eq. (2.19), as well as the mean and variance of the spike count, Eq.'s (2.20) and (2.21). The derivation serves to remind us that all these quantities follow from the general definition of the Poisson model in Eq. (2.18).

To find the distribution of spike counts we take the full probability distribution $P[\{t_i\}|s(\tau)]$, pick out the term involving $N$ spikes, and then integrate over all the possible arrival times of these spikes. That is,

$$P(N)=\int_0^T dt_1\int_0^T dt_2\cdots\int_0^T dt_N P[\{t_i\}|s(\tau)]. \qquad (A.106)$$

Substituting from Eq. (2.18), we have

$$P(N)=\int_0^T dt_1\int_0^T dt_2\cdots\int_0^T dt_N P[\{t_i\}|s(\tau)]$$

$$=\int_0^T dt_1\int_0^T dt_2\cdots\int_0^T dt_N\frac{1}{N!}\exp\left[-\int_0^T dt r[t;s(\tau)]\right]$$

$$\times\prod_{i=1}^{N} r[t_i;s(\tau)]. \qquad (A.107)$$

As in the discussion of Eq. (A.103) we notice that the exponential factor can be taken outside the integral, and that really we have a product of $N$ one dimensional integrals rather than a full $N$ dimensional integral:

$$P(N) = \int_0^T dt_1 \int_0^T dt_2 \cdots \int_0^T dt_N \frac{1}{N!} \exp\left[-\int_0^T dt\, r[t;s(\tau)]\right] \times \prod_{i=1}^{N} r[t_i;s(\tau)]$$

$$= \frac{1}{N!} \exp\left[-\int_0^T dt\, r[t;s(\tau)]\right] \times \int_0^T dt_1 \int_0^T dt_2 \cdots \int_0^T dt_N \prod_{i=1}^{N} r[t_i;s(\tau)] \quad \text{(A.108)}$$

$$= \frac{1}{N!} \exp\left[-\int_0^T dt\, r[t;s(\tau)]\right] \left(\int_0^T dt\, r[t;s(\tau)]\right)^N \quad \text{(A.109)}$$

$$= \frac{1}{N!} \exp(-Q) Q^N, \quad \text{(A.110)}$$

where we have defined

$$Q = \int_0^T dt\, r[t;s(\tau)]. \quad \text{(A.111)}$$

In particular, the probability that no spikes occur in the time from $t = 0$ to $t = T$ is $P(0) = \exp(-Q)$, or

$$P(0) = \exp\left[-\int_0^T dt\, r[t;s(\tau)]\right]. \quad \text{(A.112)}$$

With the probability distribution of spike counts from Eq. (A.110), we can compute the mean and the variance of the count. To obtain the mean we compute

$$\langle N \rangle \equiv \sum_{N=0}^{\infty} P(N) N \quad \text{(A.113)}$$

$$= \sum_{N=0}^{\infty} \frac{1}{N!} \exp(-Q) Q^N N \quad \text{(A.114)}$$



$$= \exp(-Q) \sum_{N=0}^{\infty} \frac{1}{N!} Q^N N. \quad \text{(A.115)}$$

Now we have already made use of the series expansion for the exponential, Eq. (A.104), and to sum this last series we notice that

$$Q^N N = Q \frac{\partial}{\partial Q} Q^N, \quad \text{(A.116)}$$

so that

$$\langle N \rangle = \exp(-Q) \sum_{N=0}^{\infty} \frac{1}{N!} Q^N N$$

$$= \exp(-Q) \sum_{N=0}^{\infty} \frac{1}{N!} Q \frac{\partial}{\partial Q} Q^N \quad \text{(A.117)}$$

$$= \exp(-Q) Q \frac{\partial}{\partial Q} \sum_{N=0}^{\infty} \frac{1}{N!} Q^N \quad \text{(A.118)}$$

$$= \exp(-Q) Q \frac{\partial}{\partial Q} \exp(+Q), \quad \text{(A.119)}$$

where in the last step we recognize the series for the exponential. Now the derivative of the exponential is just the exponential itself,

$$\frac{\partial}{\partial Q} \exp(+Q) = \exp(+Q), \quad \text{(A.120)}$$

so that

$$\langle N \rangle = \exp(-Q) Q \frac{\partial}{\partial Q} \exp(+Q)$$

$$= \exp(-Q) Q \exp(+Q) = Q. \quad \text{(A.121)}$$

We see that the mean spike count is what we have called $Q$, the integral of the firing rate, as promised in the text.

We can do a very similar calculation to find the variance of the count distribution. We start by computing the average of $N^2$,

$$\langle N^2 \rangle = \sum_{N=0}^{\infty} N^2 P(N). \quad \text{(A.122)}$$

Substituting for $P(N)$ from Eq. (A.110) and rearranging, we have

$$\langle N^2 \rangle = \sum_{N=0}^{\infty} N^2 P(N)$$

$$= \sum_{N=0}^{\infty} N^2 \exp(-Q) \frac{1}{N!} Q^N \tag{A.123}$$

$$= \exp(-Q) \sum_{N=0}^{\infty} \frac{1}{N!} N^2 Q^N. \tag{A.124}$$

The trick is once again to write the extra factors of $N$ (here $N^2$) in terms of derivatives with respect to $Q$. Now we know that

$$\frac{\partial^2}{\partial Q^2} Q^N = N(N-1) Q^{N-2}, \tag{A.125}$$

so we can write

$$Q^2 \frac{\partial^2}{\partial Q^2} Q^N = (N^2 - N) Q^N, \tag{A.126}$$

which is almost what we want. But we can use the formula in Eq. (A.116) to finish the job, obtaining

$$N^2 Q^N = Q^2 \frac{\partial^2}{\partial Q^2} Q^N + Q \frac{\partial}{\partial Q} Q^N. \tag{A.127}$$

Now we can substitute into Eq. (A.124) and follow the steps corresponding to Eq's. (A.117) through (A.121):

$$\langle N^2 \rangle = \exp(-Q) \sum_{N=0}^{\infty} \frac{1}{N!} N^2 Q^N$$

$$= \exp(-Q) \sum_{N=0}^{\infty} \frac{1}{N!} \left[ Q^2 \frac{\partial^2}{\partial Q^2} Q^N + Q \frac{\partial}{\partial Q} Q^N \right] \tag{A.128}$$

$$= \exp(-Q) Q^2 \frac{\partial^2}{\partial Q^2} \sum_{N=0}^{\infty} \frac{1}{N!} Q^N + \exp(-Q) Q \frac{\partial}{\partial Q} \sum_{N=0}^{\infty} \frac{1}{N!} Q^N \tag{A.129}$$

$$= \exp(-Q) Q^2 \frac{\partial^2}{\partial Q^2} \exp(+Q) + \exp(-Q) Q \frac{\partial}{\partial Q} \exp(+Q) \tag{A.130}$$



$$= \exp(-Q) Q^2 \exp(+Q) + \exp(-Q) Q \exp(+Q) \tag{A.131}$$

$$= Q^2 + Q. \tag{A.132}$$

Now since we have already identified $Q$ as equal to the mean spike count, this means that the mean square spike count can be written as

$$\langle N^2 \rangle = \langle N \rangle^2 + \langle N \rangle. \tag{A.133}$$

But the variance of the count is defined by

$$\langle (\delta N)^2 \rangle \equiv \langle N^2 \rangle - \langle N \rangle^2 \tag{A.134}$$

$$= [\langle N \rangle^2 + \langle N \rangle] - \langle N \rangle^2 = \langle N \rangle. \tag{A.135}$$

Thus the variance of the count for a Poisson process is equal to the mean count.

## A.6 ESTIMATION FROM INDEPENDENT RESPONSES

Imagine that we have observed a sequence of responses from a neuron, $R_1$, $R_2$, $\cdots$, $R_k$, and from these responses we would like to reconstruct the stimulus $s(t)$. To simplify the calculation we assume that each of the responses is statistically independent of all the others if the stimulus is fixed.

The assumption that each response $R_i$ is independent of the others can be stated mathematically as

$$P[R_1, R_2, \cdots, R_k | s] = \prod_{i=1}^{k} P[R_i | s] \tag{A.136}$$

That is, given the stimulus $s$, the probability of observing *all* of the responses $R_1, R_2, \cdots, R_k$ is the product of the probabilities of observing the responses individually. Now, by Bayes' rule, we can write the probability distribution for the stimulus given the responses in terms of the distribution of responses given the stimulus:

$$P[s | R_1, R_2, \cdots, R_k] = \frac{P[R_1, R_2, \cdots, R_k | s] P_0[s]}{P[R_1, \cdots, R_k]}. \tag{A.137}$$

But then we can use Eq. (A.136) to substitute for $P[R_1, R_2, \cdots, R_k | s]$:

$$P[s | R_1, R_2 \cdots, R_k] = \frac{P[R_1, R_2, \cdots, R_k | s] P_0[s]}{P[R_1, R_2, \cdots, R_k]}$$

$$= \left( \prod_{i=1}^{k} P[R_i | s] \right) \frac{P_0[s]}{P[R_1, R_2, \cdots, R_k]}. \tag{A.138}$$

In the experiments discussed in section 2.2.3 we have learned about the distributions $P[s|R]$, not $P[R|s]$, so we use Bayes' rule once more, now in the form

$$P[R_i|s] = \frac{P[s|R_i]P[R_i]}{P_0[s]}. \tag{A.139}$$

Substituting into Eq. (A.138) we find that

$$\begin{aligned} P[s|R_1, R_2, \cdots, R_k] &= \left(\prod_{i=1}^{k} P[R_i|s]\right) \frac{P_0[s]}{P[R_1, R_2, \cdots, R_k]} \\ &= \left(\prod_{i=1}^{k} \frac{P[s|R_i]P[R_i]}{P_0[s]}\right) \frac{P_0[s]}{P[R_1, R_2, \cdots, R_k]} \end{aligned} \tag{A.140}$$

$$= \left(\frac{1}{P[R_1, R_2, \cdots, R_k]} \prod_{i=1}^{k} P[R_i]\right) \times P_0[s] \times \left(\prod_{i=1}^{k} \frac{P[s|R_i]}{P_0[s]}\right). \tag{A.141}$$

We notice that the first factor is a constant, given that this particular combination of the $\{R_i\}$ was observed. The second factor is the a priori probability density. The third factor is a product of ratios of the probabilities of the relevant conditional velocity waveforms to those of the a priori waveforms. These conditional probabilities correspond to the response-conditional ensembles defined in section 2.2.3.

In the experiments on H1 (de Ruyter van Steveninck and Bialek 1988), the response-conditional ensembles were approximated as Gaussian. To be concrete, we discretize time so that the stimulus $s(t)$ becomes a vector $\mathbf{s}$ of values $(s_1, s_2, \cdots, s_n, \cdots)$ at discrete times $t_1, t_2, \cdots, t_n, \cdots$. Given that we have seen the response $R_i$ the mean stimulus waveform is a vector that we call $\mathbf{w}_{R_i}$, and the fluctuations around this mean are described by a covariance matrix that we write as $C_{R_i}$. Somewhat schematically, then, we can write the relevant conditional probability as

$$P[s|R_i] \propto \exp\left[-\frac{1}{2}(\mathbf{s} - \mathbf{w}_{R_i})^T \cdot C_{R_i}^{-1} \cdot (\mathbf{s} - \mathbf{w}_{R_i})\right], \tag{A.142}$$

where $\mathbf{v}^T$ denotes the transpose of the vector $\mathbf{v}$. In similar notation, the total ensemble of stimuli can be written as



$$P_0[s] \propto \exp\left[-\frac{1}{2}\mathbf{s}^T \cdot C_0^{-1} \cdot \mathbf{s}\right], \tag{A.143}$$

and now we have the ingredients to substitute into Eq. (A.141). Dropping all the constant factors—that is, the factors independent of $\mathbf{s}$—we find:

$$P[s|R_1, R_2, \cdots, R_k] \tag{A.144}$$

$$= \left(\frac{1}{P[R_1, R_2, \cdots, R_k]} \prod_{i=1}^{k} P[R_i]\right) \times P_0[s] \times \left(\prod_{i=1}^{k} \frac{P[s|R_i]}{P_0[s]}\right)$$

$$\propto \exp\left[-\frac{1}{2}\mathbf{s}^T \cdot C_0^{-1} \cdot \mathbf{s}\right] \prod_{i=1}^{k} \frac{\exp\left[-\frac{1}{2}(\mathbf{s} - \mathbf{w}_{R_i})^T \cdot C_{R_i}^{-1} \cdot (\mathbf{s} - \mathbf{w}_{R_i})\right]}{\exp\left[-\frac{1}{2}\mathbf{s}^T \cdot C_0^{-1} \cdot \mathbf{s}\right]} \tag{A.145}$$

$$\propto \exp\left[-\frac{1}{2}\mathbf{s}^T \cdot C_0^{-1} \cdot \mathbf{s} - \frac{1}{2}\sum_{i=1}^{k}(\mathbf{s} - \mathbf{w}_{R_i})^T \cdot C_{R_i}^{-1} \cdot (\mathbf{s} - \mathbf{w}_{R_i}) + \frac{1}{2}\sum_{i=1}^{k}\mathbf{s}^T \cdot C_0^{-1} \cdot \mathbf{s}\right] \tag{A.146}$$

$$\propto \exp\left\{-\frac{1}{2}\mathbf{s}^T \cdot \left[C_0^{-1} + \sum_{i=1}^{k}(C_{R_i}^{-1} - C_0^{-1})\right] \cdot \mathbf{s} + \sum_{i=1}^{k} C_{R_i}^{-1} \cdot \mathbf{w}_{R_i}\right\}. \tag{A.147}$$

Now we can see that, given the observed responses, the most likely value of the stimulus, which is also the mean value (because the distribution is Gaussian), is our best estimate:

$$\mathbf{s}_{est} = \left[\sum_{i=1}^{k}(C_{R_i}^{-1} - C_0^{-1}) + C_0^{-1}\right]^{-1} \cdot \sum_{i=1}^{k} C_{R_i}^{-1} \cdot \mathbf{w}_{R_i}. \tag{A.148}$$

Note that the $R_i$ occur at different instants, and the corresponding velocity vectors and covariances must be shifted in accordance with these occurrence times.

## A.7 CONDITIONAL MEAN AS OPTIMAL ESTIMATOR

One of the most useful facts in the theory of estimation is that if we choose $\chi^2$ as our measure of errors, then the estimator that makes the smallest errors

is always the conditional mean. Thus, if we observe $y$, then to estimate $x$ we should compute the average value of $x$ in the conditional probability distribution $P(x|y)$. We will call this conditional mean $\langle x\rangle_y$, and it is defined by

$$\langle x\rangle_y = \int dx\, P(x|y)x. \tag{A.149}$$

Here we review the proof that $\langle x\rangle_y$ is indeed the optimal estimator in that it minimizes $\chi^2$. In the process we shall introduce the notion of functional differentiation, which plays a key role in several of the calculations described in these asides.

Let us imagine that we observe $y$ and compute some function $F(y)$ that we call our estimate of $x$. Our mean square error is, by definition, an average over the joint distribution $P(x,y)$,

$$\begin{aligned}\chi^2 &= \left\langle |F(y)-x|^2 \right\rangle \\ &= \int dx \int dy\, P(x,y)[F(y)-x]^2.\end{aligned} \tag{A.150}$$

To find the minimum value of $\chi^2$ we must try different functions $F(y)$. When we do an ordinary minimization problem, we know that the minimum of a function is a place where the derivative is equal to zero, and we also need to check that the second derivative is positive. Here $\chi^2$ is a function of a function, which is called a *functional*. The idea is the same, however: We can find the minimum of $\chi^2$ by asking what happens when we make a small change $F(y) \to F(y)+\delta F(y)$, where $\delta F(y)$ is an arbitrary small function. At the minimum, the change is zero to first order in $\delta F$ and is positive at second order.

So we want to evaluate the change $\delta\chi^2$ when we make the change $F(y) \to F(y)+\delta F(y)$:

$$\chi^2[F(y)+\delta F(y)] = \int dx \int dy\, P(x,y)[F(y)+\delta F(y)-x]^2 \tag{A.151}$$

$$\begin{aligned}&= \int dx \int dy\, P(x,y)[F(y)-x]^2 \\ &\quad + 2\int dx \int dy\, P(x,y)\delta F(y)[F(y)-x] \\ &\quad + \int dx \int dy\, P(x,y)[\delta F(y)]^2.\end{aligned} \tag{A.152}$$



Now we notice that, of the three terms in this expression, the first one is just $\chi^2[F(y)]$, so we have:

$$\begin{aligned}\chi^2[F(y)+\delta F(y)] &= \int dx \int dy\, P(x,y)[F(y)-x]^2 \\ &\quad + 2\int dx \int dy\, P(x,y)\delta F(y)[F(y)-x] \\ &\quad + \int dx \int dy\, P(x,y)[\delta F(y)]^2 \\ &= \chi^2[F(y)] + 2\int dx \int dy\, P(x,y)\delta F(y)[F(y)-x] \\ &\quad + \int dx \int dy\, P(x,y)[\delta F(y)]^2.\end{aligned} \tag{A.153}$$

The change in $\chi^2$ consists of one "first order" term $\propto \delta F(y)$, one "second order" term $\propto [\delta F(y)]^2$, and no other terms; this is because $\chi^2$ is quadratic in $F(y)$. Furthermore, the second order term is obviously positive no matter what function we choose for $\delta F(y)$, so all we really have to do is find the place where the first order term vanishes.

By analogy with ordinary derivatives and the Taylor series expansion, we define the *functional derivatives* of $\chi^2$ with respect to $F(y)$:

$$\begin{aligned}\chi^2[F(y)+\delta F(y)] &= \chi^2[F(y)] + \int dy\, \delta F(y) \frac{\delta\chi^2}{\delta F(y)} \\ &\quad + \frac{1}{2}\int dy\, [\delta F(y)]^2 \frac{\delta^2\chi^2}{\delta F(y)\delta F(y)} + \cdots,\end{aligned} \tag{A.154}$$

where in this simple case the higher powers of $\delta F$ indicated by $\cdots$ are not present. Comparing Eq. (A.153) with the definition of functional derivatives in Eq. (A.154), we identify the first order terms:

$$\begin{aligned}\int dy\, \delta F(y) \frac{\delta\chi^2}{\delta F(y)} &= 2\int dx \int dy\, P(x,y)\delta F(y)[F(y)-x] \\ &= 2\int dy\, \delta F(y) \int dx\, P(x,y)[F(y)-x].\end{aligned} \tag{A.155}$$

To proceed we interchange the order of the $x$ and $y$ integrations and expand the joint distribution $P(x,y)$ in terms of the conditional distribution for $x$ given $y$, that is, $P(x,y) = P(x|y)P(y)$, to obtain

$$\frac{\delta\chi^2}{\delta F(y)} = 2\int dy\,\delta F(y)\int dx\,P(x,y)[F(y)-x]$$
$$= 2\int dy\,\delta F(y)\int dx\,P(x|y)P(y)[F(y)-x]$$
$$= 2\int dy\,\delta F(y)P(y)\int dx\,P(x|y)[F(y)-x]$$
$$= 2\int dy\,\delta F(y)P(y)$$
$$\times\left[\int dx\,P(x|y)F(y) - \int dx\,P(x|y)x\right]. \quad \text{(A.156)}$$

Of the two terms in brackets, the first one is easy to evaluate because we can take $F(y)$ outside the integral over $x$, and then use the normalization of $P(x|y)$. That is,

$$\int dx\,P(x|y)F(y) = F(y)\int dx\,P(x|y) \quad \text{(A.157)}$$
$$= F(y), \quad \text{(A.158)}$$

since the normalization condition is

$$\int dx\,P(x|y) = 1. \quad \text{(A.159)}$$

The second bracketed term in Eq. (A.156) is $\int dx\,P(x|y)x$, and comparing with Eq. (A.149) we see that this is the conditional mean $\langle x\rangle_y$. So we can go back to Eq. (A.156) and substitute for both of the terms in brackets:

$$\int dy\,\delta F(y)\frac{\delta\chi^2}{\delta F(y)} = 2\int dy\,\delta F(y)P(y)$$
$$\times\left[\int dx\,P(x|y)F(y) - \int dx\,P(x|y)x\right]$$
$$= 2\int dy\,\delta F(y)P(y)\left[F(y)-\langle x\rangle_y\right]. \quad \text{(A.160)}$$

This equation must be true for an *arbitrary* choice of the function $\delta F(y)$, which allows us to identify the terms inside the integral:

$$\int dy\,\delta F(y)\frac{\delta\chi^2}{\delta F(y)} = 2\int dy\,\delta F(y)P(y)\left[F(y)-\langle x\rangle_y\right]$$
$$\Rightarrow \frac{\delta\chi^2}{\delta F(y)} = 2P(y)\left[F(y)-\langle x\rangle_y\right] \quad \text{(A.161)}$$



The condition that first order changes in $\chi^2$ vanish for any choice of the function $\delta F(y)$ is that this functional derivative be equal to zero:

$$0 = \frac{\delta\chi^2}{\delta F(y)} \quad \text{(A.162)}$$
$$= 2P(y)\left[F(y)-\langle x\rangle_y\right] \quad \text{(A.163)}$$
$$= \left[F(y)-\langle x\rangle_y\right] \quad \text{(A.164)}$$
$$\Rightarrow F(y) = \langle x\rangle_y. \quad \text{(A.165)}$$

Thus we see that, to minimize $\chi^2$, the best estimator is equal to the conditional mean, as promised.

## A.8 PRACTICAL CALCULATIONS OF RECONSTRUCTION FILTERS

In Chapter 2 we discuss estimation of a continuous sensory stimulus by filtering the spike train; the estimate takes the form

$$s_{\text{est}}(t) = \sum_i K_1(t-t_i) + \sum_{i,j} K_2(t-t_i, t-t_j) + \cdots, \quad \text{(A.166)}$$

where the spikes occur at times $\{t_i\}$ and the $K_n$ are estimation filters. With this formulation, estimating the stimulus becomes a matter of choosing the $K_n$. We choose the filters to minimize the error function

$$E = \left\langle \int dt\,|s(t)-s_{\text{est}}(t)|^2 G[s(t)]\right\rangle, \quad \text{(A.167)}$$

where $G[s]$ imposes a variable weight on large deviations in the stimulus $s$. We begin by setting $G[s]=1$, in which case the error function is the mean-square error, $E=\chi^2$. For simplicity we also assume that the average value of the stimulus is zero.

### A.8.1 The "acausal-shifted" calculation

Let us start the analysis by restricting ourselves to linear filtering, that is $K_1(\tau)$, and ignoring the constraint of causality. Then our problem is to find the kernel $K_1(\tau)$ that minimizes

$$\chi^2[K_1(\tau)] = \left\langle \int dt\,\left| s(t) - \sum_i K_1(t-t_i)\right|^2\right\rangle. \quad \text{(A.168)}$$

Although we can proceed with $\chi^2$ in this form, it is easier to change to the frequency domain, using the ideas discussed in section 3.1.4 and in the text by Lighthill (1958). In particular, Parseval's theorem tells us that the integral over

time can be converted into an integral over frequency. In general,

$$\int dt\, F^2(t) = \int \frac{d\omega}{2\pi} |\tilde{F}(\omega)|^2, \tag{A.169}$$

where $\tilde{F}(\omega)$ is the Fourier transform of $F(t)$, defined by

$$\tilde{F}(\omega) = \int dt\, F(t) \exp[+i\omega t]. \tag{A.170}$$

In our case we need to compute the Fourier transform of the difference between the true stimulus $s(t)$ and our estimate $\sum_i K_1(t - t_i)$. By definition, the Fourier transform of $s(t)$ is $\tilde{s}(\omega)$, and the Fourier transform of our estimate can be written in a simple form:

$$\int dt \sum_i K_1(t - t_i) \exp[+i\omega t]$$

$$= \sum_i \int dt\, K_1(t - t_i) \exp[+i\omega(t - t_i)] \exp[+i\omega t_i] \tag{A.171}$$

$$= \sum_i \exp[+i\omega t_i] \int d\tau\, K_1(\tau) \exp[+i\omega\tau] \tag{A.172}$$

$$= \left[\sum_i \exp[+i\omega t_i]\right] \tilde{K}_1(\omega), \tag{A.173}$$

where $\tilde{K}_1(\omega)$ is the Fourier transform of the kernel $K_1(\tau)$. Now we can rewrite our expression for $\chi^2$ as an integral over frequencies,

$$\chi^2[K_1(\tau)] = \left\langle \int dt \left| s(t) - \sum_i K_1(t - t_i) \right|^2 \right\rangle$$

$$= \int \frac{d\omega}{2\pi} \left\langle \left| \tilde{s}(\omega) - \tilde{K}_1(\omega) \left[\sum_i \exp[+i\omega t_i]\right] \right|^2 \right\rangle \tag{A.174}$$

$$= \int \frac{d\omega}{2\pi} \langle |\tilde{s}(\omega)|^2 \rangle - \int \frac{d\omega}{2\pi} \tilde{K}_1(\omega) \left\langle \tilde{s}^*(\omega) \left( \sum_i \exp[+i\omega t_i] \right) \right\rangle$$

$$- \int \frac{d\omega}{2\pi} \tilde{K}_1^*(\omega) \left\langle \tilde{s}(\omega) \left( \sum_i \exp[-i\omega t_i] \right) \right\rangle$$



$$+ \int \frac{d\omega}{2\pi} |\tilde{K}_1(\omega)|^2 \left\langle \left| \sum_i \exp[+i\omega t_i] \right|^2 \right\rangle. \tag{A.175}$$

We see that each frequency component of the kernel makes an independent contribution to $\chi^2$, so we can follow the procedure of section A.7 for each component. The result is that $\chi^2$ will be minimized if we choose, at every frequency $\omega$, a kernel that satisfies the condition

$$\langle \tilde{s}(\omega) \sum_i \exp(-i\omega t_i) \rangle = \langle \tilde{K}_1(\omega) \sum_{i,j} \exp[i\omega(t_i - t_j)] \rangle. \tag{A.176}$$

The filter $K_1$ depends only on the stimulus ensemble, so it can be taken out of the average. Then we can solve Eq. (A.176) for the filter,

$$\tilde{K}_1(\omega) = \frac{\langle \tilde{s}(\omega) \sum_i \exp(-i\omega t_i) \rangle}{\langle |\sum_i \exp(i\omega t_i)|^2 \rangle}. \tag{A.177}$$

The filter is the Fourier transform of the average stimulus surrounding a spike divided by the power spectrum of the spike train (see sections 3.1.4 and A.2). Thus the filter is completely determined by the experimental stimulus $s(t)$ and measured spike times $\{t_i\}$ in response to this stimulus.

Equation (A.177) represents the best *acausal* filter. To carry out real-time estimation of the stimulus we turn this filter into a causal filter. We can insure that the filter $K_1(\tau)$ is causal by setting $K_1(\tau < -\tau_{\rm delay}) = 0$ and shifting the filter by a delay $\tau_{\rm delay}$. Thus we define a filter $K_1^{\rm shift}(\tau) = \theta(\tau) K_1(\tau - \tau_{\rm delay})$, where $\theta(\tau < 0) = 0$ and $\theta(\tau > 0) = 1$. In general $K_1^{\rm shift}$ will no longer minimize $\chi^2$; if, however, the main features of the filter are confined to times $\tau > -\tau_{\rm delay}$ the truncation process will not significantly change the filter characteristics. We can check the influence of imposing causality in this way using the second approach to calculating the filter.

A limitation of this approach to the calculation is the inability to go beyond the first term in Eq. (A.166). As we discuss in section 2.3.1, from studies of models for spike generation we believe that the contribution of higher order terms will be relatively small, but this statement clearly requires verification. Such verification is provided by the second approach to calculating the filter.

### A.8.2 Power series expansions of the $K_n$

To calculate the causal filters directly we expand the filter in a power series of explicitly causal functions, $f_\mu(\tau)$ where $f_\mu(\tau < 0) = 0$. In this case, our estimate of the stimulus is

$$s_{\rm est}(t) = \sum_\mu a_\mu \int d\tau\, f_\mu(\tau) \sum_i \delta(t - t_i - \tau)$$
$$+ \sum_{\mu,\nu} b_{\mu\nu} \int d\tau\, d\tau' f_\mu(\tau) f_\nu(\tau') \sum_i \delta(t - t_i - \tau) \sum_j \delta(t - t_j - \tau')$$
$$+ \cdots, \tag{A.178}$$

where the expansion coefficients $a_\mu$ determine the shape of the linear filter, $b_{\mu\nu}$ determine the shape of second order filter, etc. For computational purposes the power series expansion must be truncated after a finite number of terms, which limits the calculation to filters that can be created from the resulting finite basis set. Comparing filters calculated using this method with the acausal-shifted filters serves to check if this truncation limits the choice of filters, as well as telling us something about the influence of the causality constraint on the form of the filters.

We choose a particular delay time for the reconstruction, and vary the power series coefficients $a_\mu, b_{\mu\nu}, \cdots$ to minimize

$$E(\tau_{\rm delay}) = \left\langle \int dt\, |s(t - \tau_{\rm delay}) - s_{\rm est}(t)|^2 G[s(t - \tau_{\rm delay})] \right\rangle. \tag{A.179}$$

Again we consider $G[s] = 1$ so we are minimizing the mean square error $\chi^2$. Restricting ourselves for the moment to the linear coefficients $a_\mu$, the condition $\partial\chi^2/\partial a_\beta = 0$ leads to

$$a_\beta = \sum_\mu F_\mu (N^{-1})_{\mu\beta}, \tag{A.180}$$

where $N^{-1}$ denotes the inverse of the matrix $N$,

$$F_\mu = \left\langle \int dt\, s(t - \tau_{\rm delay}) \sum_i f_\mu(t - t_i) \right\rangle$$
$$= \int d\tau\, f_\mu(\tau) \left\langle \sum_i s(\tau + t_i) \right\rangle, \tag{A.181}$$

and

$$N_{\mu\nu} = \left\langle \int dt \sum_{i,j} f_\mu(t - t_i) f_\nu(t - t_j) \right\rangle$$
$$= \int d\tau d\tau' f_\mu(\tau) f_\nu(\tau') \int dt \left\langle \sum_{i,j} \delta(t - t_i - \tau) \delta(t - t_j - \tau') \right\rangle. \tag{A.182}$$



Note that the coefficients depend on the experimental stimulus $s(\tau)$ and spike times $\{t_i\}$; in particular, $N$ depends on the two-point correlation function (or autocorrelation function) of the spike train and the average stimulus surrounding a spike; this is much the same structure we saw for the acausal-shifted filter, Eq. (A.177).

The calculation is essentially the same for higher order terms. We can rewrite Eq. (A.8.2) in a more general form:

$$s_{\rm est}(t) = \sum_n x_n y_n(t), \tag{A.183}$$

where

$$\vec{x} = (a_1, a_2, \ldots, b_{11}, b_{12}, \ldots, b_{22}, b_{23}, \ldots) \tag{A.184}$$

is a vector containing the expansion coefficients, and

$$\vec{y}(t) = \begin{pmatrix} \sum_i f_1(t - t_i) \\ \sum_i f_2(t - t_i) \\ \vdots \\ \sum_{i,j} f_1(t - t_i) f_1(t - t_j) \\ \sum_{i,j} f_1(t - t_i) f_2(t - t_j) \\ \vdots \\ \sum_{i,j} f_2(t - t_i) f_2(t - t_j) \\ \sum_{i,j} f_2(t - t_i) f_3(t - t_j) \\ \vdots \end{pmatrix}, \tag{A.185}$$

contains products of the basis functions $f_\mu(\tau)$ convolved with the spike train. Our minimization condition is now $\partial\chi^2(\tau_{\rm delay})/\partial x_n = 0$, which leads to

$$x_n = \sum_m F'_m (N'^{-1})_{mn} \tag{A.186}$$

where

$$F'_m = \left\langle \int dt\, s(t - \tau_{\rm delay}) y_m(t) \right\rangle \tag{A.187}$$

and

$$N'_{mn} = \left\langle \int dt\, y_m(t) y_n(t) \right\rangle. \tag{A.188}$$

These higher order filters depend on correlation functions of the spike train and correlation functions of the stimulus and spike trains.

## A.9 ENTROPY OF GAUSSIAN DISTRIBUTIONS

Here we indicate the steps in calculating the entropy of the Gaussian distribution. We begin with the definition of entropy for distributions of continuous variables, Eq. (3.4),

$$S = -\int dx\, P(x) \log_2 P(x) \quad \text{bits.} \tag{A.189}$$

It is convenient to work with natural logarithms, so we rewrite this as

$$\begin{aligned} S &= -\int dx\, P(x) \log_2 P(x) \\ &= -\frac{1}{\ln 2} \int dx\, P(x) \ln P(x). \end{aligned} \tag{A.190}$$

For a Gaussian distribution we have

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[ -\frac{(x-M)^2}{2\sigma^2} \right], \tag{A.191}$$

$$\ln P(x) = -\ln\left(\sqrt{2\pi\sigma^2}\right) - \frac{(x-M)^2}{2\sigma^2}. \tag{A.192}$$

So the entropy can be obtained by doing the following integral:

$$\begin{aligned} S &= -\frac{1}{\ln 2} \int dx\, P(x) \ln P(x) \\ &= \frac{1}{\ln 2} \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \int dx \exp\left[ -\frac{(x-M)^2}{2\sigma^2} \right] \\ &\qquad \times \left[ \ln\left(\sqrt{2\pi\sigma^2}\right) + \frac{(x-M)^2}{2\sigma^2} \right]. \end{aligned} \tag{A.193}$$

We can do this integral explicitly, or we can notice that it has the form of an expectation value, or average over the Gaussian distribution. We adopt the notation

$$\langle f(x) \rangle = \frac{1}{\sqrt{2\pi\sigma^2}} \int dx \exp\left[ -\frac{(x-M)^2}{2\sigma^2} \right] f(x) \tag{A.194}$$

for the average of any function $f(x)$. Then the entropy is



$$\begin{aligned} S &= -\frac{1}{\ln 2} \int dx\, P(x) \ln P(x) \\ &= -\frac{1}{\ln 2} \langle \ln P(x) \rangle \qquad \text{(A.195)} \\ &= -\frac{1}{\ln 2} \left\langle -\ln\left(\sqrt{2\pi\sigma^2}\right) - \frac{(x-M)^2}{2\sigma^2} \right\rangle \qquad \text{(A.196)} \\ &= \frac{1}{\ln 2} \left[ \ln\left(\sqrt{2\pi\sigma^2}\right) + \frac{1}{2\sigma^2} \langle (x-M)^2 \rangle \right] \qquad \text{(A.197)} \end{aligned}$$

Notice that the average $\langle (x-M)^2 \rangle$ is, by definition, the variance $\sigma^2$, and then this cancels the $\sigma^2$ in the denominator:

$$\begin{aligned} S &= \frac{1}{\ln 2} \left[ \ln\left(\sqrt{2\pi\sigma^2}\right) + \frac{1}{2\sigma^2} \langle (x-M)^2 \rangle \right] \\ &= \frac{1}{\ln 2} \left[ \ln\left(\sqrt{2\pi\sigma^2}\right) + \frac{1}{2} \right] \qquad \text{(A.198)} \\ &= \frac{1}{\ln 2} \left[ \frac{1}{2} \ln(2\pi\sigma^2) + \frac{1}{2} \right] \qquad \text{(A.199)} \\ &= \frac{1}{2\ln 2} \ln(2\pi e\sigma^2) \quad \text{bits,} \qquad \text{(A.200)} \end{aligned}$$

where in last step we make use of the fact that $\ln e = 1$. Finally we go back to using logs to the base two:

$$\begin{aligned} S &= \frac{1}{2\ln 2} \ln(2\pi e\sigma^2) \\ &= \frac{1}{2} \log_2(2\pi e\sigma^2) \quad \text{bits,} \end{aligned} \tag{A.201}$$

which completes the calculation.

## A.10 APPROXIMATING THE ENTROPY OF SPIKE TRAINS

We begin with Eq. (3.21) for the entropy of spike trains:

$$S = -\frac{T}{\Delta\tau \ln 2} \left[ (\bar{r}\Delta\tau) \ln(\bar{r}\Delta\tau) + (1 - \bar{r}\Delta\tau) \ln(1 - \bar{r}\Delta\tau) \right]. \tag{A.202}$$

We are interested in the limiting behavior when the bin size $\Delta\tau$ is small. To find this limit we make use of the Taylor expansion for the natural logarithm,

$$\ln(1+x) = x - \frac{1}{2}x^2 + \cdots, \tag{A.203}$$

as illustrated in Fig. 2.8. In this case we identify $x = -\bar{r}\Delta\tau$, so that

$$\ln(1-\bar{r}\Delta\tau) = -\bar{r}\Delta\tau - \frac{1}{2}(\bar{r}\Delta\tau)^2 + \cdots. \tag{A.204}$$

Substituting into Eq. (A.202), we find

$$\begin{aligned} S &= -\frac{T}{\Delta\tau \ln 2}\left[(\bar{r}\Delta\tau)\ln(\bar{r}\Delta\tau) + (1-\bar{r}\Delta\tau)\ln(1-\bar{r}\Delta\tau)\right] \\ &= -\frac{T}{\Delta\tau \ln 2}\left[(\bar{r}\Delta\tau)\ln(\bar{r}\Delta\tau) + (1-\bar{r}\Delta\tau)(-\bar{r}\Delta\tau + \cdots)\right] \qquad \text{(A.205)} \\ &\approx -\frac{T}{\Delta\tau \ln 2}\left[(\bar{r}\Delta\tau)\ln(\bar{r}\Delta\tau) - \bar{r}\Delta\tau\right], \qquad \text{(A.206)} \end{aligned}$$

where we drop terms proportional to the square of the bin size $(\Delta\tau)^2$. Now we pull out the common factor of $\bar{r}\Delta\tau$ and write the extra "1" as the natural logarithm of $e$, as in Eq. (A.200):

$$\begin{aligned} S &\approx -\frac{T}{\Delta\tau \ln 2}\left[(\bar{r}\Delta\tau)\ln(\bar{r}\Delta\tau) - \bar{r}\Delta\tau\right] \\ &= -\frac{T}{\Delta\tau \ln 2}(\bar{r}\Delta\tau)\left[\ln(\bar{r}\Delta\tau) - 1\right] \qquad \text{(A.207)} \\ &= -\frac{\bar{r}T}{\ln 2}\left[\ln(\bar{r}\Delta\tau) - \ln e\right] \qquad \text{(A.208)} \\ &= -\frac{\bar{r}T}{\ln 2}\ln\left(\frac{\bar{r}\Delta\tau}{e}\right) \qquad \text{(A.209)} \\ &= (\bar{r}T)\frac{1}{\ln 2}\ln\left(\frac{e}{\bar{r}\Delta\tau}\right) \qquad \text{(A.210)} \\ &= (\bar{r}T)\log_2\left(\frac{e}{\bar{r}\Delta\tau}\right), \qquad \text{(A.211)} \end{aligned}$$

where in the last step we change back to a logarithm to the base two. Finally, dividing the entropy by the duration $T$ of the spike train we obtain the formula for the entropy rate with small bins,

$$S/T \approx \bar{r}\log_2\left(\frac{e}{\bar{r}\Delta\tau}\right), \tag{A.212}$$

which is Eq. (3.22) of the text.



## A.11 MAXIMUM ENTROPY AND SPIKE COUNTS

One of the most useful techniques in the application of information theory is *maximum entropy*. We use this idea literally, to mean the probability distribution that has maximum entropy given the values of certain average quantities. As a first application, we derive the maximum entropy distribution of spike counts given the mean count or average firing rate. This problem is more general—we could be counting spikes, vesicles at a synapse, photons arriving at a rod cell, or cars arriving at an intersection. We want to find the probability distribution for the count $n$ that has the largest entropy, given that we know the mean count $\langle n \rangle$.

We have some distribution $p(n)$, and the entropy of this distribution is

$$S[p(n)] = -\sum_{n=0}^{\infty} p(n)\log_2 p(n). \tag{A.213}$$

We would like to search all possible distributions and find the one for which $S[p(n)]$ is maximal. As in the discussion of optimal estimation in section A.7, the idea is to examine the changes in $S$ when we make small changes in the function $p(n)$, and look for a function $p(n)$ such that these changes are zero at first order and negative (since we are looking for a maximum) at second order.

The difficulty is that we cannot make arbitrary changes in the function $p(n)$. First of all, this function is a probability distribution, and hence it must obey the normalization condition

$$1 = \sum_{n=0}^{\infty} p(n). \tag{A.214}$$

Second, we assume that the average count is known, and so the distribution has to obey the equation

$$\langle n \rangle = \sum_{n=0}^{\infty} np(n). \tag{A.215}$$

Our problem, then, is to maximize the entropy while holding fixed the two sums in Eq's. (A.214) and (A.215).

There is a general approach to solving such constrained maximization problems, and this is the method of Lagrange multipliers (Mathews and Walker 1964). If we have some function $f(x_1, x_2, \cdots, x_N)$ and we want to find the maximum of this function while holding some other function $g(x_1, x_2, \cdots, x_N)$ fixed and equal to $g_0$, we look for the maximum of the new function

$$F(x_1, x_2, \cdots, x_N; \lambda) = f(x_1, x_2, \cdots, x_N) - \lambda g(x_1, x_2, \cdots, x_N), \tag{A.216}$$

where $\lambda$ is called the Lagrange multiplier. When we find the set of variables $x_1, x_2, \cdots, x_N$ that maximizes $F$, this extremal point depends on $\lambda$. So at the end of the calculation we adjust $\lambda$ to the point where $g$ has the correct value $g_0$.

Let us see how this technique works for maximizing the entropy. Since we have two constraints, normalization and the mean count, we need two Lagrange multipliers, that we will call $\lambda_1$ and $\lambda_2$. Then we want to maximize the function

$$\bar{S}[p(n)] = -\sum_{n=0}^{\infty} p(n)\log_2 p(n) - \lambda_1 \sum_{n=0}^{\infty} p(n) - \lambda_2 \sum_{n=0}^{\infty} np(n). \quad \text{(A.217)}$$

We want to examine what happens when we make a small but arbitrary change $p(n) \to p(n) + \delta p(n)$, and identify the functional derivatives of $\bar{S}$ with respect to $p(n)$ (see section A.7). We begin with the expression for $\bar{S}$,

$$\begin{aligned}\bar{S}[p(n)+\delta p(n)] = &-\sum_{n=0}^{\infty}[p(n)+\delta p(n)]\log_2[p(n)+\delta p(n)] \\ &- \lambda_1 \sum_{n=0}^{\infty}[p(n)+\delta p(n)] - \lambda_2 \sum_{n=0}^{\infty} n[p(n)+\delta p(n)].\end{aligned} \quad \text{(A.218)}$$

Of the three pieces in this expression, it is clear that all the hard work is involved in evaluating the first one, which involves the logarithm; the other two pieces are already "expanded" into terms that are either $\propto p(n)$ or $\propto \delta p(n)$.

Let us pull out the first piece on the right side of Eq. (A.218), work through all the steps, and then put it back together with all the other terms. We begin by converting to natural logarithms and breaking the log into a sum of two terms, one of which depends on $\delta p(n)$ and one of which does not:

$$\sum_{n=0}^{\infty}[p(n)+\delta p(n)]\log_2[p(n)+\delta p(n)]$$

$$= \frac{1}{\ln 2}\sum_{n=0}^{\infty}[p(n)+\delta p(n)]\ln[p(n)+\delta p(n)] \quad \text{(A.219)}$$

$$= \frac{1}{\ln 2}\sum_{n=0}^{\infty}[p(n)+\delta p(n)]\ln\left(p(n)\left[1+\frac{\delta p(n)}{p(n)}\right]\right) \quad \text{(A.220)}$$



$$= \frac{1}{\ln 2}\sum_{n=0}^{\infty}[p(n)+\delta p(n)]\ln p(n) + \frac{1}{\ln 2}\sum_{n=0}^{\infty}[p(n)+\delta p(n)]\ln\left[1+\frac{\delta p(n)}{p(n)}\right]. \quad \text{(A.221)}$$

To proceed further we need the Taylor series expansion of the natural logarithm,

$$\ln(1+x) = x - \frac{1}{2}x^2 + \frac{1}{3}x^3 - \cdots, \quad \text{(A.222)}$$

which is useful for small $x$. Looking at Eq. (A.221), we can use this expansion to write

$$\ln\left[1+\frac{\delta p(n)}{p(n)}\right] = \frac{\delta p(n)}{p(n)} - \frac{1}{2}\left(\frac{\delta p(n)}{p(n)}\right)^2 + \cdots. \quad \text{(A.223)}$$

Substituting back into Eq. (A.221), we find

$$\begin{aligned}&\sum_{n=0}^{\infty}[p(n)+\delta p(n)]\log_2[p(n)+\delta p(n)] \\ &= \frac{1}{\ln 2}\sum_{n=0}^{\infty}[p(n)+\delta p(n)]\ln p(n) + \frac{1}{\ln 2}\sum_{n=0}^{\infty}[p(n)+\delta p(n)]\ln\left[1+\frac{\delta p(n)}{p(n)}\right] \\ &= \frac{1}{\ln 2}\sum_{n=0}^{\infty}[p(n)+\delta p(n)]\ln p(n) + \frac{1}{\ln 2}\sum_{n=0}^{\infty}[p(n)+\delta p(n)]\left[\frac{\delta p(n)}{p(n)} - \frac{1}{2}\left(\frac{\delta p(n)}{p(n)}\right)^2 + \cdots\right] \quad \text{(A.224)} \\ &= \frac{1}{\ln 2}\sum_{n=0}^{\infty}[p(n)+\delta p(n)]\ln p(n) \\ &\quad + \frac{1}{\ln 2}\sum_{n=0}^{\infty}\left[p(n)\cdot\frac{\delta p(n)}{p(n)} - p(n)\frac{1}{2}\left(\frac{\delta p(n)}{p(n)}\right)^2 + \delta p(n)\frac{\delta p(n)}{p(n)} + \cdots\right]\end{aligned} \quad \text{(A.225)}$$

$$= \frac{1}{\ln 2}\sum_{n=0}^{\infty} p(n)\ln p(n) + \sum_{n=0}^{\infty}\delta p(n)\left[\frac{1}{\ln 2}\ln p(n) + \frac{1}{\ln 2}\right]$$

$$+ \frac{1}{2}\sum_{n=0}^{\infty}[\delta p(n)]^2\left[\frac{1}{\ln 2}\cdot\frac{1}{p(n)}\right] + \cdots. \tag{A.226}$$

Now we are ready to substitute from Eq. (A.226) back into our original expression, Eq. (A.218), for $\tilde{S}[p(n)+\delta p(n)]$ and collect terms that have the same powers of $\delta p(n)$:

$$\begin{aligned}\tilde{S}[p(n)+\delta p(n)] &= -\sum_{n=0}^{\infty}[p(n)+\delta p(n)]\log_2[p(n)+\delta p(n)] \\ &\quad - \lambda_1\sum_{n=0}^{\infty}[p(n)+\delta p(n)] - \lambda_2\sum_{n=0}^{\infty} n[p(n)+\delta p(n)]. \\ &= -\frac{1}{\ln 2}\sum_{n=0}^{\infty} p(n)\ln p(n) \\ &\quad - \sum_{n=0}^{\infty}\delta p(n)\left[\frac{1}{\ln 2}\ln p(n) + \frac{1}{\ln 2}\right] \\ &\quad - \frac{1}{2}\sum_{n=0}^{\infty}[\delta p(n)]^2\left[\frac{1}{\ln 2}\cdot\frac{1}{p(n)}\right] - \cdots \\ &\quad - \lambda_1\sum_{n=0}^{\infty}[p(n)+\delta p(n)] - \lambda_2\sum_{n=0}^{\infty} n[p(n)+\delta p(n)]\end{aligned} \tag{A.227}$$

$$\begin{aligned}&= -\frac{1}{\ln 2}\sum_{n=0}^{\infty} p(n)\ln p(n) - \lambda_1\sum_{n=0}^{\infty} p(n) - \lambda_2\sum_{n=0}^{\infty} p(n)n \\ &+ \sum_{n=0}^{\infty}\delta p(n)\left[-\frac{1}{\ln 2}\ln p(n) - \frac{1}{\ln 2} - \lambda_1 - \lambda_2 n\right] \\ &+ \frac{1}{2}\sum_{n=0}^{\infty}[\delta p(n)]^2\left[-\frac{1}{\ln 2}\cdot\frac{1}{p(n)}\right] + \cdots.\end{aligned} \tag{A.228}$$

This equation provides all the ingredients necessary to identify the functional derivatives of $\tilde{S}$.



We want to write

$$\tilde{S}[p(n)+\delta p(n)] = \tilde{S}[p(n)] + \sum_{n=0}^{\infty}\frac{\delta\tilde{S}}{\delta p(n)}\delta p(n) + \frac{1}{2}\sum_{n=0}^{\infty}\frac{\delta^2\tilde{S}}{\delta p(n)\delta p(n)}[\delta p(n)]^2 + \cdots. \tag{A.229}$$

Comparing this with Eq. (A.228), we notice that the terms in the first line of the equation, which do not involve $\delta p(n)$, really do add up to give $\tilde{S}[p(n)]$ as defined in Eq. (A.217). This has to be true, so it is a good way to check that we didn't make the mistake of dropping any of the terms. Then we match up the terms $\propto \delta p(n)$ to find the first functional derivative:

$$\sum_{n=0}^{\infty}\frac{\delta\tilde{S}}{\delta p(n)}\delta p(n) = \sum_{n=0}^{\infty}\delta p(n)\left[-\frac{1}{\ln 2}\ln p(n) - \frac{1}{\ln 2} - \lambda_1 - \lambda_2 n\right] \tag{A.230}$$

$$\Rightarrow \frac{\delta\tilde{S}}{\delta p(n)} = -\left[\frac{1}{\ln 2}\cdot\ln p(n) + \frac{1}{\ln 2} + \lambda_1 + n\lambda_2\right]. \tag{A.231}$$

Similarly, we match up the terms $\propto [\delta p(n)]^2$ to find the second functional derivative:

$$\frac{1}{2}\sum_{n=0}^{\infty}\frac{\delta^2\tilde{S}}{\delta p(n)\delta p(n)}[\delta p(n)]^2 = \frac{1}{2}\sum_{n=0}^{\infty}[\delta p(n)]^2\left[-\frac{1}{\ln 2}\cdot\frac{1}{p(n)}\right] \tag{A.232}$$

$$\Rightarrow \frac{\delta^2\tilde{S}}{\delta p(n)\delta p(n)} = -\frac{1}{\ln 2}\cdot\frac{1}{p(n)}. \tag{A.233}$$

Note that the second derivative is always negative, so if we can find the place where the first derivative is zero then this point is guaranteed to be a maximum.

The first functional derivative of $\tilde{S}$ is zero when the probability distribution satisfies the equation

$$0 = \frac{\delta\tilde{S}}{\delta p(n)} \tag{A.234}$$

$$= -\left[\frac{1}{\ln 2}\cdot\ln p(n) + \frac{1}{\ln 2} + \lambda_1 + n\lambda_2\right] \tag{A.235}$$

$$\ln p(n) = -1 - (\lambda_1 + n\lambda_2)(\ln 2). \tag{A.236}$$

or

$$p(n) = \frac{1}{Z}\exp(-\lambda n), \tag{A.237}$$

$$Z = \exp[1 + \lambda_1 \ln 2] \tag{A.238}$$

$$\lambda = \lambda_2 \ln 2. \tag{A.239}$$

At this point we have proved that the maximum entropy distribution has the exponential form of Eq. (A.237). Next we have to choose the values of $\lambda_1$ and $\lambda_2$ to insure that the two conditions in Eq's. (A.214) and (A.215) are satisfied. Equivalently, we have to choose the parameters $Z$ and $\lambda$.

The normalization condition is

$$1 = \sum_{n=0}^{\infty} p(n)$$

$$= \sum_{n=0}^{\infty} \frac{1}{Z}\exp(-\lambda n) \tag{A.240}$$

$$= \frac{1}{Z}\sum_{n=0}^{\infty} \exp(-\lambda n) \tag{A.241}$$

$$= \frac{1}{Z}\sum_{n=0}^{\infty} \left[\exp(-\lambda)\right]^n. \tag{A.242}$$

So we see that, to maintain normalization, we must have

$$Z = \sum_{n=0}^{\infty} \left[\exp(-\lambda)\right]^n. \tag{A.243}$$

This is a geometric series, which can be summed exactly. In general,

$$\sum_{n=0}^{\infty} x^n = \frac{1}{1-x}, \tag{A.244}$$

so in the expression for $Z$ we identify $x = \exp(-\lambda)$ and find

$$Z = \sum_{n=0}^{\infty} \left[\exp(-\lambda)\right]^n$$

$$= \frac{1}{1-\exp(-\lambda)}. \tag{A.245}$$



To fix the average count we must have

$$\langle n \rangle = \sum_{n=0}^{\infty} n p(n) \tag{A.246}$$

$$= \sum_{n=0}^{\infty} n \frac{1}{Z}\exp(-\lambda n) \tag{A.247}$$

$$= \frac{1}{Z}\sum_{n=0}^{\infty} n \exp(-\lambda n). \tag{A.248}$$

We know how to sum the series $\sum \exp(-\lambda n)$ that arose in satisfying the normalization condition, but now we have to sum a slightly different series. To do this we notice that

$$n\exp(-\lambda n) = (-1)\frac{\partial}{\partial \lambda}\exp(-\lambda n), \tag{A.249}$$

so that

$$\langle n \rangle = \frac{1}{Z}\sum_{n=0}^{\infty} n \exp(-\lambda n)$$

$$= \frac{1}{Z}\sum_{n=0}^{\infty} (-1)\frac{\partial \exp(-\lambda n)}{\partial \lambda} \tag{A.250}$$

$$= -\frac{1}{Z}\frac{\partial}{\partial \lambda}\sum_{n=0}^{\infty} \exp(-\lambda n) \tag{A.251}$$

$$= -\frac{1}{Z}\frac{\partial}{\partial \lambda}\sum_{n=0}^{\infty} \left[\exp(-\lambda)\right]^n \tag{A.252}$$

$$= -\frac{1}{Z}\frac{\partial}{\partial \lambda}\frac{1}{1-\exp(-\lambda)} \tag{A.253}$$

$$= -\frac{1}{Z}(-1)\frac{\exp(-\lambda)}{[1-\exp(-\lambda)]^2} \tag{A.254}$$

$$= \frac{\exp(-\lambda)}{1-\exp(-\lambda)}, \tag{A.255}$$

where in the last step we substitute the expression for $Z$ from Eq. (A.245). Thus, to fix the average count, we must have

$$\langle n\rangle = \frac{\exp(-\lambda)}{1-\exp(-\lambda)}$$
$$= \frac{1}{\exp(\lambda)-1}, \tag{A.256}$$

or

$$\exp(\lambda) = 1 + 1/\langle n\rangle \tag{A.257}$$
$$\lambda = \ln(1 + 1/\langle n\rangle), \tag{A.258}$$

as promised in the text.

Finally, we need to compute the entropy itself:

$$S = -\frac{1}{\ln 2}\sum_{n=0}^{\infty} p(n)\ln p(n)$$
$$= -\frac{1}{\ln 2}\sum_{n=0}^{\infty}\frac{1}{Z}\exp(-\lambda n)\ln\left[\frac{1}{Z}\exp(-\lambda n)\right] \tag{A.259}$$
$$= -\frac{1}{\ln 2}\frac{1}{Z}\sum_{n=0}^{\infty}\exp(-\lambda n)(-\ln Z - n\lambda) \tag{A.260}$$
$$= -\frac{1}{\ln 2}(-\ln Z - \lambda\langle n\rangle) \tag{A.261}$$
$$= \log_2(Z) + \frac{1}{\ln 2}\lambda\langle n\rangle. \tag{A.262}$$

Now we substitute for $Z$ from Eq. (A.245) and for $\lambda$ from Eq. (A.258):

$$S = \log_2(Z) + \frac{1}{\ln 2}\lambda\langle n\rangle$$
$$= \log_2\left[\frac{1}{1-\exp(-\lambda)}\right] + \frac{1}{\ln 2}\lambda\langle n\rangle \tag{A.263}$$
$$= \log_2(1+\langle n\rangle) + \langle n\rangle\frac{1}{\ln 2}\ln(1+1/\langle n\rangle) \tag{A.264}$$
$$= \log_2(1+\langle n\rangle) + \langle n\rangle\log_2(1+1/\langle n\rangle) \quad \text{bits.} \tag{A.265}$$

To summarize, Eq. (A.265) gives us the maximum possible entropy of a counting distribution, given that we fix the mean count $\langle n\rangle$.



## A.12 THE GAUSSIAN CHANNEL

Here we indicate the steps leading to Eq. (3.35), the mutual information for a Gaussian channel. We begin with the definition of the information,

$$I = \int dy \int ds P(y,s)\log_2\left[\frac{P(y,s)}{P(y)P(s)}\right]. \tag{A.266}$$

We simplify this expression by writing the joint distribution $P(y,s)$ in terms of the conditional distribution $P(y|s)$ and the prior distribution $P(s)$, using

$$P(y,s) = P(y|s)\times P(s). \tag{A.267}$$

Substituting into Eq. (A.266), we can cancel the prior distribution from the numerator and denominator in the logarithm:

$$I = \int dy \int ds P(y,s)\log_2\left[\frac{P(y,s)}{P(y)P(s)}\right]$$
$$= \int dy \int ds P(y|s)P(s)\log_2\left[\frac{P(y|s)P(s)}{P(y)P(s)}\right] \tag{A.268}$$
$$= \int dy \int ds P(y|s)P(s)\log_2\left[\frac{P(y|s)}{P(y)}\right]. \tag{A.269}$$

It will be convenient to transform our logarithms into natural logarithms, and then to expand out the log of a ratio as the difference of logs:

$$I = \int dy \int ds P(y|s)P(s)\log_2\left[\frac{P(y|s)}{P(y)}\right]$$
$$= \frac{1}{\ln 2}\int dy \int ds P(y|s)P(s)\ln\left[\frac{P(y|s)}{P(y)}\right] \tag{A.270}$$
$$= \frac{1}{\ln 2}\int dy \int ds P(y|s)P(s)\left[\ln P(y|s) - \ln P(y)\right]. \tag{A.271}$$

We are interested in the case of the Gaussian channel, where both the distribution of the signal $P(s)$ and the conditional distribution of the output given the signal $P(y|s)$ are Gaussian, as discussed in section 3.1.3:

$$P(s) = \frac{1}{\sqrt{2\pi\langle s^2\rangle}}\exp\left[-\frac{s^2}{2\langle s^2\rangle}\right]$$
$$P(y|s) = \frac{1}{\sqrt{2\pi\langle \eta^2\rangle}}\exp\left[-\frac{(y-gs)^2}{2\langle \eta^2\rangle}\right]. \tag{A.272}$$

Each of the variables $s$ and $y$ can take on any values from $-\infty$ to $+\infty$, so all the integrals in Eq. (A.271) have to be taken over this range. From Eq. (A.271) it is clear that we also need to know $P(y)$, the distribution of outputs. This turns out to be Gaussian, but it takes a few steps to find the final form of this Gaussian. We begin with the definition of $P(y)$ as an integral over possible input signals:

$$P(y) = \int_{-\infty}^{\infty} ds\, P(y|s)P(s) \tag{A.273}$$

$$= \int_{-\infty}^{\infty} ds \frac{1}{\sqrt{2\pi\langle\eta^2\rangle}} \exp\left[-\frac{(y-gs)^2}{2\langle\eta^2\rangle}\right] \frac{1}{\sqrt{2\pi\langle s^2\rangle}} \exp\left[-\frac{s^2}{2\langle s^2\rangle}\right] \tag{A.274}$$

$$= \frac{1}{2\pi\sqrt{\langle s^2\rangle\langle\eta^2\rangle}} \int_{-\infty}^{\infty} ds \exp\left[-\frac{(y-gs)^2}{2\langle\eta^2\rangle} - \frac{s^2}{2\langle s^2\rangle}\right] \tag{A.275}$$

$$= \frac{1}{2\pi\sqrt{\langle s^2\rangle\langle\eta^2\rangle}} \times \int_{-\infty}^{\infty} ds \exp\left[-\frac{1}{2}s^2\left(\frac{1}{\langle s^2\rangle} + \frac{g^2}{\langle\eta^2\rangle}\right) + s\left(\frac{gy}{\langle\eta^2\rangle}\right) - \frac{y^2}{2\langle\eta^2\rangle}\right] \tag{A.276}$$

$$= \frac{1}{2\pi\sqrt{\langle s^2\rangle\langle\eta^2\rangle}} \exp\left[-\frac{y^2}{2\langle\eta^2\rangle}\right] \times \int_{-\infty}^{\infty} ds \exp\left[-\frac{1}{2}s^2\left(\frac{1}{\langle s^2\rangle} + \frac{g^2}{\langle\eta^2\rangle}\right) + s\left(\frac{gy}{\langle\eta^2\rangle}\right)\right]. \tag{A.277}$$

To do the integral over $s$, we need the general form of Gaussian integrals,

$$\int_{-\infty}^{\infty} ds \exp\left[-\frac{1}{2}As^2 + Bs\right] = \sqrt{\frac{2\pi}{A}} \exp\left[\frac{B^2}{2A}\right]. \tag{A.278}$$

We see that Eq. (A.277) is of the same form as the integral in Eq. (A.278) if we make the following identifications:



$$A = \frac{1}{\langle s^2\rangle} + \frac{g^2}{\langle\eta^2\rangle}$$

$$B = \frac{gy}{\langle\eta^2\rangle}. \tag{A.279}$$

Thus we have

$$P(y) = \frac{1}{2\pi\sqrt{\langle s^2\rangle\langle\eta^2\rangle}} \exp\left[-\frac{y^2}{2\langle\eta^2\rangle}\right] \times \int_{-\infty}^{\infty} ds \exp\left[-\frac{1}{2}s^2\left(\frac{1}{\langle s^2\rangle} + \frac{g^2}{\langle\eta^2\rangle}\right) + s\left(\frac{gy}{\langle\eta^2\rangle}\right)\right]$$

$$= \frac{1}{2\pi\sqrt{\langle s^2\rangle\langle\eta^2\rangle}} \exp\left[-\frac{y^2}{2\langle\eta^2\rangle}\right] \times \sqrt{\frac{2\pi}{1/\langle s^2\rangle + g^2/\langle\eta^2\rangle}} \exp\left[\frac{(gy/\langle\eta^2\rangle)^2}{1/\langle s^2\rangle + g^2/\langle\eta^2\rangle}\right] \tag{A.280}$$

$$= \frac{1}{\sqrt{2\pi(g^2\langle s^2\rangle + \langle\eta^2\rangle)}} \times \exp\left[-\frac{y^2}{2}\left(-\frac{1}{\langle\eta^2\rangle} + \frac{g^2/\langle\eta^2\rangle^2}{1/\langle s^2\rangle + g^2/\langle\eta^2\rangle}\right)\right] \tag{A.281}$$

$$= \frac{1}{\sqrt{2\pi(g^2\langle s^2\rangle + \langle\eta^2\rangle)}} \times \exp\left[-\frac{y^2}{2}\left(-\frac{1}{\langle\eta^2\rangle} + \frac{1}{\langle\eta^2\rangle}\cdot\frac{g^2\langle s^2\rangle}{\langle\eta^2\rangle + g^2\langle s^2\rangle}\right)\right] \tag{A.282}$$

$$= \frac{1}{\sqrt{2\pi(g^2\langle s^2\rangle + \langle\eta^2\rangle)}} \exp\left[-\frac{y^2}{2(g^2\langle s^2\rangle + \langle\eta^2\rangle)}\right] \tag{A.283}$$

$$= \frac{1}{\sqrt{2\pi\langle y^2\rangle}} \exp\left[-\frac{y^2}{2\langle y^2\rangle}\right], \tag{A.284}$$

where we have identified the variance at the output,

$$\langle y^2\rangle = g^2\langle s^2\rangle + \langle\eta^2\rangle. \tag{A.285}$$

Now we have the ingredients to substitute back into Eq. (A.271):

$$I = \frac{1}{\ln 2}\int_{-\infty}^{\infty} dy \int_{-\infty}^{\infty} ds\, P(y|s)P(s)\left[\ln P(y|s) - \ln P(y)\right]$$
$$= \frac{1}{\ln 2}\int_{-\infty}^{\infty} dy \int_{-\infty}^{\infty} ds\, P(y|s)P(s)$$
$$\times \left[-\ln\sqrt{2\pi\langle\eta^2\rangle} - \frac{(y-gs)^2}{2\langle\eta^2\rangle} + \ln\sqrt{2\pi\langle y^2\rangle} + \frac{y^2}{2\langle y^2\rangle}\right]. \tag{A.286}$$

Of the four terms that we need to evaluate, two are just constants so we can do the integrals over $y$ and $s$ easily:

$$\int_{-\infty}^{\infty} dy \int_{-\infty}^{\infty} ds\, P(y|s)P(s)\left[-\ln\sqrt{2\pi\langle\eta^2\rangle} + \ln\sqrt{2\pi\langle y^2\rangle}\right]$$
$$= \left[-\ln\sqrt{2\pi\langle\eta^2\rangle} + \ln\sqrt{2\pi\langle y^2\rangle}\right]\int_{-\infty}^{\infty} dy \int_{-\infty}^{\infty} ds\, P(y|s)P(s)$$
$$= \left[-\ln\sqrt{2\pi\langle\eta^2\rangle} + \ln\sqrt{2\pi\langle y^2\rangle}\right]\int_{-\infty}^{\infty} ds\, P(s)\int_{-\infty}^{\infty} dy\, P(y|s)$$
$$= \left[-\ln\sqrt{2\pi\langle\eta^2\rangle} + \ln\sqrt{2\pi\langle y^2\rangle}\right]\int_{-\infty}^{\infty} ds\, P(s)$$
$$= -\ln\sqrt{2\pi\langle\eta^2\rangle} + \ln\sqrt{2\pi\langle y^2\rangle}, \tag{A.287}$$

where in each of the last two steps we use the normalization conditions,

$$1 = \int_{-\infty}^{\infty} dy\, P(y|s), \tag{A.288}$$

$$1 = \int_{-\infty}^{\infty} ds\, P(s). \tag{A.289}$$

The other two terms are also quite simple if we do the integrals in the right order. One of the terms is

$$\int_{-\infty}^{\infty} ds \int_{-\infty}^{\infty} dy\, P(y|s)P(s)\left[-\frac{(y-gs)^2}{2\langle\eta^2\rangle}\right]$$



$$= -\int_{-\infty}^{\infty} ds\, P(s)\int_{-\infty}^{\infty} dy\, P(y|s)\frac{(y-gs)^2}{2\langle\eta^2\rangle}$$
$$= -\int_{-\infty}^{\infty} ds\, P(s)\frac{1}{2\langle\eta^2\rangle}\int_{-\infty}^{\infty} dy\, P(y|s)(y-gs)^2. \tag{A.290}$$

But remember that $P(y|s)$ describes $y$ as a Gaussian variable with a mean of $gs$ and a variance $\langle\eta^2\rangle$, so that

$$\int_{-\infty}^{\infty} dy\, P(y|s)(y-gs)^2 = \langle\eta^2\rangle. \tag{A.291}$$

Thus we see that

$$\int_{-\infty}^{\infty} ds \int_{-\infty}^{\infty} dy\, P(y|s)P(s)\left[-\frac{(y-gs)^2}{2\langle\eta^2\rangle}\right]$$
$$= -\int_{-\infty}^{\infty} ds\, P(s)\frac{1}{2\langle\eta^2\rangle}\int_{-\infty}^{\infty} dy\, P(y|s)(y-gs)^2$$
$$= -\int_{-\infty}^{\infty} ds\, P(s)\frac{1}{2\langle\eta^2\rangle}\langle\eta^2\rangle$$
$$= -\int_{-\infty}^{\infty} ds\, P(s)\frac{1}{2}$$
$$= -\frac{1}{2}. \tag{A.292}$$

Essentially the same argument allows us to evaluate the fourth term:

$$\int_{-\infty}^{\infty} ds \int_{-\infty}^{\infty} dy\, P(y|s)P(s)\frac{y^2}{2\langle y^2\rangle}$$
$$= \int_{-\infty}^{\infty} dy\, \frac{y^2}{2\langle y^2\rangle}\int_{-\infty}^{\infty} ds\, P(y|s)P(s)$$
$$= \int_{-\infty}^{\infty} dy\, \frac{y^2}{2\langle y^2\rangle}P(y)$$
$$= \frac{\langle y^2\rangle}{2\langle y^2\rangle}$$
$$= \frac{1}{2}. \tag{A.293}$$

Finally we can put all of the pieces together to compute the information itself:

$$I = \frac{1}{\ln 2}\int_{-\infty}^{\infty} dy \int_{-\infty}^{\infty} ds\, P(y|s)P(s) \times \left[-\ln\sqrt{2\pi\langle\eta^2\rangle} - \frac{(y-gs)^2}{2\langle\eta^2\rangle} + \ln\sqrt{2\pi\langle y^2\rangle} + \frac{y^2}{2\langle y^2\rangle}\right]$$

$$= \frac{1}{\ln 2}\left[-\ln\sqrt{2\pi\langle\eta^2\rangle} - \frac{1}{2} + \ln\sqrt{2\pi\langle y^2\rangle} + \frac{1}{2}\right] \quad (A.294)$$

$$= \frac{1}{\ln 2}\ln\sqrt{\frac{\langle y^2\rangle}{\langle\eta^2\rangle}}, \quad (A.295)$$

$$= \frac{1}{\ln 2}\times\frac{1}{2}\ln\left[\frac{\langle y^2\rangle}{\langle\eta^2\rangle}\right] \quad (A.296)$$

$$= \frac{1}{2}\log_2\left[\frac{\langle y^2\rangle}{\langle\eta^2\rangle}\right] \quad (A.297)$$

$$= \frac{1}{2}\log_2\left[1+\frac{g^2\langle s^2\rangle}{\langle\eta^2\rangle}\right], \quad (A.298)$$

which completes the calculation.

## A.13 GAUSSIANS AND MAXIMUM ENTROPY

We want to find the probability distribution $P(x)$ that has the largest possible entropy assuming that we know the mean and variance of $x$. The idea of the calculation is the same as in section A.11, where we discussed maximum entropy for distributions of spike counts. Again we use the technique of Lagrange multipliers.

We want to maximize the quantity

$$S = -\int_{-\infty}^{\infty} dx\, P(x)\log_2 P(x), \quad (A.299)$$

but we have several constraints to obey. First, we know that the function $P(x)$, being a probability distribution, must be normalized, so that

$$\int_{-\infty}^{\infty} dx\, P(x) = 1. \quad (A.300)$$



Next, we know the mean value of $x$,

$$\int_{-\infty}^{\infty} dx\, P(x)x = \langle x\rangle. \quad (A.301)$$

Finally, we know the variance $\sigma^2$ of $x$. It is easier to write this constraint by saying that we know the average value of $x^2$, that is

$$\int_{-\infty}^{\infty} dx\, P(x)x^2 = \langle x^2\rangle = \langle x\rangle^2 + \sigma^2. \quad (A.302)$$

As in A.3.2, we define a function $\tilde{S}$ that adds to the entropy one Lagrange multiplier term for each of the constraints:

$$\tilde{S}[P(x)] = -\int_{-\infty}^{\infty} dx\, P(x)\log_2 P(x) - \lambda_1\int_{-\infty}^{\infty} dx\, P(x) - \lambda_2\int_{-\infty}^{\infty} dx\, P(x)x - \lambda_3\int_{-\infty}^{\infty} dx\, P(x)x^2. \quad (A.303)$$

Now we try to find the function $P(x)$ that maximizes $\tilde{S}$.

We want to evaluate $\tilde{S}$ with a function $P(x)+\delta P(x)$, and expand in powers of $\delta P(x)$, identifying the functional derivatives. We start with

$$\tilde{S}[P(x)+\delta P(x)] = -\int_{-\infty}^{\infty} dx[P(x)+\delta P(x)]\log_2[P(x)+\delta P(x)] - \lambda_1\int_{-\infty}^{\infty} dx[P(x)+\delta P(x)] - \lambda_2\int_{-\infty}^{\infty} dx[P(x)+\delta P(x)]x - \lambda_3\int_{-\infty}^{\infty} dx[P(x)+\delta P(x)]x^2. \quad (A.304)$$

As before, all of the difficulty comes from the first, logarithmic, term. So we isolate this term and work out its expansion, then substitute back into Eq. (A.304) at the end. We begin by converting to natural logarithms, then breaking the log apart into two terms:

$$\int_{-\infty}^{\infty} dx[P(x)+\delta P(x)]\log_2[P(x)+\delta P(x)] = \frac{1}{\ln 2}\int_{-\infty}^{\infty} dx[P(x)+\delta P(x)]\ln[P(x)+\delta P(x)] \quad (A.305)$$

$$= \frac{1}{\ln 2}\int_{-\infty}^{\infty} dx[P(x)+\delta P(x)]\ln\left(P(x)\left[1+\frac{\delta P(x)}{P(x)}\right]\right) \tag{A.306}$$

$$= \frac{1}{\ln 2}\int_{-\infty}^{\infty} dx[P(x)+\delta P(x)]\ln P(x) + \frac{1}{\ln 2}\int_{-\infty}^{\infty} dx[P(x)+\delta P(x)]\ln\left[1+\frac{\delta P(x)}{P(x)}\right]. \tag{A.307}$$

Now we use once more the Taylor expansion of the natural logarithm, Eq. (A.93), in this case approximating

$$\ln\left[1+\frac{\delta P(x)}{P(x)}\right] = \frac{\delta P(x)}{P(x)} - \frac{1}{2}\left(\frac{\delta P(x)}{P(x)}\right)^2 + \cdots. \tag{A.308}$$

Substituting into Eq. (A.307) and collecting terms that have the same powers of $\delta P(x)$, we find:

$$\begin{aligned}
&\int_{-\infty}^{\infty} dx[P(x)+\delta P(x)]\log_2[P(x)+\delta P(x)] \\
&\quad = \frac{1}{\ln 2}\int_{-\infty}^{\infty} dx[P(x)+\delta P(x)]\ln P(x) \\
&\qquad + \frac{1}{\ln 2}\int_{-\infty}^{\infty} dx[P(x)+\delta P(x)]\ln\left[1+\frac{\delta P(x)}{P(x)}\right] \\
&\quad = \frac{1}{\ln 2}\int_{-\infty}^{\infty} dx[P(x)+\delta P(x)]\ln P(x) \\
&\qquad + \frac{1}{\ln 2}\int_{-\infty}^{\infty} dx[P(x)+\delta P(x)]\left[\frac{\delta P(x)}{P(x)} - \frac{1}{2}\left(\frac{\delta P(x)}{P(x)}\right)^2 + \cdots\right]
\end{aligned} \tag{A.309}$$

$$\begin{aligned}
&= \frac{1}{\ln 2}\int_{-\infty}^{\infty} dx[P(x)+\delta P(x)]\ln P(x) \\
&\quad + \frac{1}{\ln 2}\int_{-\infty}^{\infty} dx\left[\delta P(x) + \frac{1}{2}\left(\frac{\delta P(x)}{P(x)}\right)^2 + \cdots\right]
\end{aligned} \tag{A.310}$$



$$\begin{aligned}
&= \frac{1}{\ln 2}\int_{-\infty}^{\infty} dx P(x)\ln P(x) \\
&\quad + \frac{1}{\ln 2}\int_{-\infty}^{\infty} dx\delta P(x)\left[\ln P(x)+1\right] \\
&\quad + \frac{1}{\ln 2}\int_{-\infty}^{\infty} dx[\delta P(x)]^2\left[\frac{1}{2}\cdot\frac{1}{P(x)}\right] + \cdots,
\end{aligned} \tag{A.311}$$

Now we go back to the definition of $\tilde{S}[P(x)+\delta P(x)]$ in Eq. (A.304) and substitute for the logarithmic term:

$$\begin{aligned}
\tilde{S}[P(x)+\delta P(x)] = &-\int_{-\infty}^{\infty} dx[P(x)+\delta P(x)]\log_2[P(x)+\delta P(x)] \\
&- \lambda_1\int_{-\infty}^{\infty} dx[P(x)+\delta P(x)] \\
&- \lambda_2\int_{-\infty}^{\infty} dx[P(x)+\delta P(x)]x \\
&- \lambda_3\int_{-\infty}^{\infty} dx[P(x)+\delta P(x)]x^2 \\
= &\frac{1}{\ln 2}\int_{-\infty}^{\infty} dx P(x)\ln P(x) \\
&\quad + \frac{1}{\ln 2}\int_{-\infty}^{\infty} dx\delta P(x)\left[\ln P(x)+1\right] \\
&\quad + \frac{1}{\ln 2}\int_{-\infty}^{\infty} dx[\delta P(x)]^2\left[\frac{1}{2}\cdot\frac{1}{P(x)}\right] + \cdots \\
&- \lambda_1\int_{-\infty}^{\infty} dx[P(x)+\delta P(x)] \\
&- \lambda_2\int_{-\infty}^{\infty} dx[P(x)+\delta P(x)]x \\
&- \lambda_3\int_{-\infty}^{\infty} dx[P(x)+\delta P(x)]x^2
\end{aligned} \tag{A.312}$$

The functional derivatives are defined by the equation

$$\tilde{S}[P(x)+\delta P(x)] = \tilde{S}[P(x)] + \int dx\frac{\delta\tilde{S}}{\delta P(x)}\delta P(x) + \frac{1}{2}\int dx\frac{\delta^2\tilde{S}}{\delta P(x)\delta P(x)}[\delta P(x)]^2 + \cdots. \tag{A.313}$$

If we identify the terms proportional to $\delta P(x)$ we can isolate the first functional derivative,

$$\int dx \frac{\delta \tilde{S}}{\delta P(x)} \delta P(x) = -\int dx \delta P(x) \left\{ \frac{1}{\ln 2} [\ln P(x) + 1] + \lambda_1 + \lambda_2 x + \lambda_3 x^2 \right\} \tag{A.314}$$

$$\Rightarrow \frac{\delta \tilde{S}}{\delta P(x)} = -\frac{1}{\ln 2} [\ln P(x) + 1] - \lambda_1 - \lambda_2 x - \lambda_3 x^2. \tag{A.315}$$

Similarly, we identify terms proportional to $[\delta P(x)]^2$ to find the second functional derivative,

$$\frac{1}{2} \int dx \frac{\delta^2 \tilde{S}}{\delta P(x) \delta P(x)} [\delta P(x)]^2 = \frac{1}{2} \int dx [\delta P(x)]^2 \left[ -\frac{1}{\ln 2} \cdot \frac{1}{P(x)} \right] \tag{A.316}$$

$$\Rightarrow \frac{\delta^2 \tilde{S}}{\delta P(x) \delta P(x)} = -\frac{1}{\ln 2} \cdot \frac{1}{P(x)}. \tag{A.317}$$

As in the discussion of previous sections, the maximum is obtained when the first functional derivative is zero and the second functional derivative is negative. Clearly the second derivative is automatically negative, so all we need to do is solve the problem of the first derivative being zero:

$$0 = \frac{\delta \tilde{S}}{\delta P(x)} \tag{A.318}$$

$$= -\frac{1}{\ln 2} [\ln P(x) + 1] - \lambda_1 - \lambda_2 x - \lambda_3 x^2. \tag{A.319}$$

We rewrite this equation and collect the dependence on $x$ into a slightly more suggestive (if cumbersome) form:

$$0 = -\frac{1}{\ln 2} [\ln P(x) + 1] - \lambda_1 - \lambda_2 x - \lambda_3 x^2$$

$$\frac{1}{\ln 2} \ln P(x) = -\frac{1}{\ln 2} - \lambda_1 - \lambda_2 x - \lambda_3 x^2$$

$$\ln P(x) = -(1 + \lambda_1 \ln 2) - x(\lambda_2 \ln 2) - x^2(\lambda_3 \ln 2)$$

$$= -(\lambda_3 \ln 2) \left[ x^2 + 2x \left( \frac{\lambda_2}{2\lambda_3} \right) \right] - (1 + \lambda_1 \ln 2)$$



$$= -\frac{1}{2}(2\lambda_3 \ln 2) \left[ x^2 + 2x \left( \frac{\lambda_2}{2\lambda_3} \right) + \left( \frac{\lambda_2}{2\lambda_3} \right)^2 \right]$$

$$- \left[ 1 + \lambda_1 \ln 2 - \frac{1}{2}(2\lambda_3 \ln 2) \left( \frac{\lambda_2}{2\lambda_3} \right)^2 \right]$$

$$= -\frac{1}{2(2\lambda_3 \ln 2)^{-1}} \left( x + \frac{\lambda_2}{2\lambda_3} \right)^2$$

$$- \left[ 1 + \lambda_1 \ln 2 - \frac{1}{2}(2\lambda_3 \ln 2) \left( \frac{\lambda_2}{2\lambda_3} \right)^2 \right] \tag{A.320}$$

$$P(x) = \exp \left[ -1 - \lambda_1 \ln 2 + \frac{1}{2}(2\lambda_3 \ln 2) \left( \frac{\lambda_2}{2\lambda_3} \right)^2 \right]$$

$$\times \exp \left[ -\frac{1}{2(2\lambda_3 \ln 2)^{-1}} \left( x + \frac{\lambda_2}{2\lambda_3} \right)^2 \right], \tag{A.321}$$

where in the last step we have taken the exponential of both sides of the equation.

We recall that a Gaussian distribution is of the form

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{1}{2\sigma^2} (x - \langle x \rangle)^2 \right]. \tag{A.322}$$

Comparing this with the maximum entropy distribution in Eq. (A.321), we see that they have the same form if we make the following identifications of the parameters:

$$\langle x \rangle = -\frac{\lambda_2}{2\lambda_3} \tag{A.323}$$

$$\sigma^2 = \frac{1}{2\lambda_3 \ln 2} \tag{A.324}$$

$$\frac{1}{\sqrt{2\pi\sigma^2}} = \exp \left[ -1 - \lambda_1 \ln 2 + \frac{1}{2}(2\lambda_3 \ln 2) \left( \frac{\lambda_2}{2\lambda_3} \right)^2 \right]. \tag{A.325}$$

When we do calculations with Lagrange multipliers we have to set the values of these multipliers at the end of the calculation to make sure that the constraints are satisfied. In this case we have to make sure that the average $x$ has the correct value, and this is the condition expressed in Eq. (A.323). Then we

have to fix the variance, and this is the condition expressed in Eq. (A.324). Finally, we have to make sure that the distribution is normalized, and this is the condition expressed in Eq. (A.325). All three of these equations can be solved, giving the following values for the Lagrange multipliers:

$$\lambda_1 = -\frac{1}{\ln 2} + \frac{1}{2}\log_2(2\pi\sigma^2) + \frac{1}{2\ln 2}\frac{\langle x\rangle^2}{\sigma^2} \tag{A.326}$$

$$\lambda_2 = -\frac{\langle x\rangle}{\sigma^2 \ln 2} \tag{A.327}$$

$$\lambda_3 = \frac{1}{2\sigma^2 \ln 2}. \tag{A.328}$$

To summarize, we have found that the maximum entropy distribution consistent with a given mean and variance is Eq. (A.321), where the three Lagrange multipliers must be set to the values in Eq's. (A.326–A.328). But once we make these substitutions, the distribution we have found is exactly the familiar Gaussian distribution of Eq. (A.322). So the complete answer to our problem is: Given a certain mean and a certain variance, the maximum entropy distribution is Gaussian.

## A.14 WIENER–KHINCHINE THEOREM

In this section we review the connection between the correlation function and the power spectrum. The two quantities turn out to be a Fourier transform pair, and this fact is the theorem referred to in the title of this section.

We begin with the definition of the correlation function as the average value of the product of functions evaluated at times separated by $\tau$:

$$C(\tau) \equiv \langle f(t) f(t-\tau)\rangle. \tag{A.329}$$

Note that in this definition the average is taken over the probability distribution of random functions $f(t)$, so we should imagine taking many samples of the function in successive experiments, then average over these samples. In practice one often averages over a single long experiment, replacing the ensemble average by a time average. As discussed in the text, the idea that this replacement should work is called ergodicity, and one often sees the correlation function defined explicitly in terms of a time average. We prefer the ensemble approach because it forces us to keep in mind the picture of the function $f(t)$ being chosen from a probability distribution in exactly the same way



that the occurrence of heads or tails is chosen from a probability distribution when we flip a coin.

We have discussed in section 3.1.4 that we can describe a random function in terms of its Fourier coefficients, and that at least for Gaussian random functions this description is especially simple. So we write

$$f(t) = \sum_{n=-\infty}^{\infty} f_n \exp[-i\omega_n t]. \tag{A.330}$$

Substituting into the definition of the correlation function we have

$$C(\tau) = \langle f(t) f(t-\tau)\rangle$$

$$= \left\langle \sum_{n=-\infty}^{\infty} f_n \exp[-i\omega_n t] \sum_{m=-\infty}^{\infty} f_m \exp[-i\omega_m (t-\tau)] \right\rangle \tag{A.331}$$

$$= \sum_{n=-\infty}^{\infty}\sum_{m=-\infty}^{\infty} \langle f_n f_m\rangle \exp[-i\omega_n t]\exp[-i\omega_m(t-\tau)], \tag{A.332}$$

where in the last step we notice that the only random elements that we need to average are the $f_n$.

The reason that the Fourier representation is so convenient is that the covariance matrix of the $f_n$s is simple, as in Eq. (3.52),

$$\langle f_n [f_n]^*\rangle = \langle f_n f_{-n}\rangle = \sigma^2(\omega_n) \tag{A.333}$$

$$\langle f_n f_m\rangle = 0 \qquad m \neq -n. \tag{A.334}$$

These relations imply that in the expression for the correlation function, the only terms that survive the averaging to give nonzero contributions are those with $m = -n$, and we recall that $\omega_{-n} = -\omega_n$:

$$C(\tau) = \sum_{n=-\infty}^{\infty}\sum_{m=-\infty}^{\infty} \langle f_n f_m\rangle \exp[-i\omega_n t]\exp[-i\omega_m(t-\tau)]$$

$$= \sum_{n=-\infty}^{\infty} \langle f_n f_{-n}\rangle \exp[-i\omega_n t]\exp[-i\omega_{-n}(t-\tau)] \tag{A.335}$$

$$= \sum_{n=-\infty}^{\infty} \langle f_n f_{-n}\rangle \exp[-i\omega_n t]\exp[+i\omega_n(t-\tau)] \tag{A.336}$$

$$= \sum_{n=-\infty}^{\infty} \sigma^2(\omega_n)\exp(-i\omega_n\tau). \tag{A.337}$$

Finally, we recall that sums over discrete frequencies become integrals over a continuous frequency variable once we let our time window $T$ become large, as discussed in connection with the manipulations from Eq. (3.62) to (3.65). Thus

$$\begin{aligned} C(\tau) &= \sum_{n=-\infty}^{\infty} \sigma^2(\omega_n) \exp(-i\omega_n \tau) \\ &\to \int_{-\infty}^{\infty} \frac{d\omega}{2\pi} [T\sigma^2(\omega)] \exp(-i\omega\tau) \\ &= \int_{-\infty}^{\infty} \frac{d\omega}{2\pi} S(\omega) \exp(-i\omega\tau), \end{aligned} \tag{A.338}$$

where in the final step we recognize the power spectrum as defined in Eq. (3.66). Thus we see that the correlation function is the Fourier transform of the power spectrum, as promised.

## A.15 MAXIMIZING INFORMATION TRANSMISSION

If we choose signals from a Gaussian distribution and these signals are corrupted by the addition of Gaussian noise, then the rate of information transmission is given by Eq. (3.72):

$$R_{\text{info}} = \frac{1}{2} \int_{-\infty}^{\infty} \frac{d\omega}{2\pi} \log_2 \left[ 1 + \frac{S(\omega)}{N(\omega)} \right], \tag{A.339}$$

where $S(\omega)$ is the power spectrum of the signal and $N(\omega)$ is the power spectrum of the noise. We are interested in knowing the maximum value of $R_{\text{info}}$ given that we have a fixed noise spectrum $N(\omega)$ and a fixed total variance for the signal. More precisely, we would like to know how to take our "budget" for signal variance and distribute it over frequency so as to maximize the information transmission.

The total signal variance is related to the power spectrum by Eq. (3.65),

$$\langle s^2 \rangle = \int_{-\infty}^{\infty} \frac{d\omega}{2\pi} S(\omega). \tag{A.340}$$

So the problem we want to solve is maximizing $R_{\text{info}}$ while holding $\langle s^2 \rangle$ fixed. As in previous sections, we introduce a Lagrange multiplier and maximize a new function



$$\tilde{R} = R_{\text{info}} - \lambda \langle s^2 \rangle \tag{A.341}$$

$$= \frac{1}{2} \int_{-\infty}^{\infty} \frac{d\omega}{2\pi} \log_2 \left[ 1 + \frac{S(\omega)}{N(\omega)} \right] - \lambda \int_{-\infty}^{\infty} \frac{d\omega}{2\pi} S(\omega). \tag{A.342}$$

We want to evaluate $\tilde{R}$ with a signal spectrum $S(\omega) + \delta S(\omega)$ and expand to find the functional derivatives, so we start with

$$\begin{aligned} \tilde{R}[S(\omega) + \delta S(\omega)] = &\frac{1}{2} \int_{-\infty}^{\infty} \frac{d\omega}{2\pi} \log_2 \left[ 1 + \frac{S(\omega) + \delta S(\omega)}{N(\omega)} \right] \\ &- \lambda \int_{-\infty}^{\infty} \frac{d\omega}{2\pi} [S(\omega) + \delta S(\omega)] \end{aligned} \tag{A.343}$$

Once again all of the difficulty comes from the logarithmic term, so we tackle this first. We convert to natural logarithms then rearrange the log to isolate the dependence on $\delta S(\omega)$:

$$\int_{-\infty}^{\infty} \frac{d\omega}{2\pi} \log_2 \left[ 1 + \frac{S(\omega) + \delta S(\omega)}{N(\omega)} \right]$$

$$= \frac{1}{\ln 2} \int_{-\infty}^{\infty} \frac{d\omega}{2\pi} \ln \left[ 1 + \frac{S(\omega) + \delta S(\omega)}{N(\omega)} \right] \tag{A.344}$$

$$= \frac{1}{\ln 2} \int_{-\infty}^{\infty} \frac{d\omega}{2\pi} \ln \left[ 1 + \frac{S(\omega)}{N(\omega)} + \frac{\delta S(\omega)}{N(\omega)} \right] \tag{A.345}$$

$$= \frac{1}{\ln 2} \int_{-\infty}^{\infty} \frac{d\omega}{2\pi} \ln \left\{ \left[ 1 + \frac{S(\omega)}{N(\omega)} \right] \left[ 1 + \frac{\delta S(\omega)/N(\omega)}{1 + S(\omega)/N(\omega)} \right] \right\} \tag{A.346}$$

$$= \frac{1}{\ln 2} \int_{-\infty}^{\infty} \frac{d\omega}{2\pi} \ln \left[ 1 + \frac{S(\omega)}{N(\omega)} \right] + \frac{1}{\ln 2} \int_{-\infty}^{\infty} \frac{d\omega}{2\pi} \ln \left[ 1 + \frac{\delta S(\omega)/N(\omega)}{1 + S(\omega)/N(\omega)} \right] \tag{A.347}$$

Once again, to proceed we make use of the Taylor expansion of the natural logarithm, Eq. (A.222). In this case we expand

$$\ln \left[ 1 + \frac{\delta S(\omega)/N(\omega)}{1 + S(\omega)/N(\omega)} \right] = \frac{\delta S(\omega)/N(\omega)}{1 + S(\omega)/N(\omega)} - \frac{1}{2} \left( \frac{\delta S(\omega)/N(\omega)}{1 + S(\omega)/N(\omega)} \right)^2 + \cdots. \tag{A.348}$$

Substituting into Eq. (A.347) we have

$$\begin{aligned}
&\int_{-\infty}^{\infty}\frac{d\omega}{2\pi}\log_2\left[1+\frac{S(\omega)+\delta S(\omega)}{N(\omega)}\right]\\
&\quad=\frac{1}{\ln 2}\int_{-\infty}^{\infty}\frac{d\omega}{2\pi}\ln\left[1+\frac{S(\omega)}{N(\omega)}\right]\\
&\qquad+\frac{1}{\ln 2}\int_{-\infty}^{\infty}\frac{d\omega}{2\pi}\ln\left[1+\frac{\delta S(\omega)/N(\omega)}{1+S(\omega)/N(\omega)}\right]\\
&\quad=\frac{1}{\ln 2}\int_{-\infty}^{\infty}\frac{d\omega}{2\pi}\ln\left[1+\frac{S(\omega)}{N(\omega)}\right]\\
&\qquad+\frac{1}{\ln 2}\int_{-\infty}^{\infty}\frac{d\omega}{2\pi}\frac{\delta S(\omega)/N(\omega)}{1+S(\omega)/N(\omega)}\\
&\qquad-\frac{1}{2\ln 2}\int_{-\infty}^{\infty}\frac{d\omega}{2\pi}\left(\frac{\delta S(\omega)/N(\omega)}{1+S(\omega)/N(\omega)}\right)^2+\cdots. \qquad \text{(A.349)}
\end{aligned}$$

Now we can substitute back into the definition of $\tilde{R}[S(\omega)+\delta S(\omega)]$,

$$\begin{aligned}
&\tilde{R}[S(\omega)+\delta S(\omega)]\\
&\quad=\frac{1}{2}\int_{-\infty}^{\infty}\frac{d\omega}{2\pi}\log_2\left[1+\frac{S(\omega)+\delta S(\omega)}{N(\omega)}\right]\\
&\qquad-\lambda\int_{-\infty}^{\infty}\frac{d\omega}{2\pi}[S(\omega)+\delta S(\omega)]\\
&\quad=\frac{1}{\ln 2}\int_{-\infty}^{\infty}\frac{d\omega}{2\pi}\ln\left[1+\frac{S(\omega)}{N(\omega)}\right]\\
&\qquad+\frac{1}{\ln 2}\int_{-\infty}^{\infty}\frac{d\omega}{2\pi}\frac{\delta S(\omega)/N(\omega)}{1+S(\omega)/N(\omega)}\\
&\qquad-\frac{1}{2\ln 2}\int_{-\infty}^{\infty}\frac{d\omega}{2\pi}\left(\frac{\delta S(\omega)/N(\omega)}{1+S(\omega)/N(\omega)}\right)^2+\cdots\\
&\qquad-\lambda\int_{-\infty}^{\infty}\frac{d\omega}{2\pi}[S(\omega)+\delta S(\omega)]. \qquad \text{(A.350)}
\end{aligned}$$

we notice that all of the terms that are independent of $\delta S(\omega)$ add up to give $\tilde{R}[S(\omega)]$, as they must, and we collect the terms proportional to $\delta S(\omega)$ and $[\delta S(\omega)]^2$:



$$\begin{aligned}
\tilde{R}[S(\omega)+\delta S(\omega)] &= \tilde{R}[S(\omega)]\\
&\quad+\frac{1}{2\ln 2}\int_{-\infty}^{\infty}\frac{d\omega}{2\pi}\frac{\delta S(\omega)/N(\omega)}{1+S(\omega)/N(\omega)}\\
&\quad-\frac{1}{2\ln 2}\int_{-\infty}^{\infty}\frac{d\omega}{2\pi}\frac{1}{2}\left(\frac{\delta S(\omega)/N(\omega)}{1+S(\omega)/N(\omega)}\right)^2\\
&\quad+\cdots\\
&\quad-\lambda\int_{-\infty}^{\infty}\frac{d\omega}{2\pi}\delta S(\omega) \qquad \text{(A.351)}\\
&= \tilde{R}[S(\omega)]\\
&\quad+\int_{-\infty}^{\infty}\frac{d\omega}{2\pi}\delta S(\omega)\left[\frac{1}{2\ln 2}\cdot\frac{1/N(\omega)}{1+S(\omega)/N(\omega)}-\lambda\right]\\
&\quad-\frac{1}{4\ln 2}\int_{-\infty}^{\infty}\frac{d\omega}{2\pi}[\delta S(\omega)]^2\left(\frac{1/N(\omega)}{1+S(\omega)/N(\omega)}\right)^2\\
&\quad+\cdots \qquad \text{(A.352)}\\
&= \tilde{R}[S(\omega)]+\int_{-\infty}^{\infty}\frac{d\omega}{2\pi}\delta S(\omega)\frac{\delta\tilde{R}[S(\omega)]}{\delta S(\omega)}\\
&\quad+\frac{1}{2}\int_{-\infty}^{\infty}\frac{d\omega}{2\pi}[\delta S(\omega)]^2\frac{\delta^2\tilde{R}[S(\omega)]}{\delta S(\omega)\delta S(\omega)}+\cdots, \qquad \text{(A.353)}
\end{aligned}$$

where we have identified the functional derivatives

$$\frac{\delta\tilde{R}[S(\omega)]}{\delta S(\omega)}=\frac{1}{2\ln 2}\frac{1/N(\omega)}{1+S(\omega)/N(\omega)}-\lambda \qquad \text{(A.354)}$$

$$=\frac{1}{2\ln 2}\frac{1}{S(\omega)+N(\omega)}-\lambda, \qquad \text{(A.355)}$$

$$\frac{\delta^2\tilde{R}[S(\omega)]}{\delta S(\omega)\delta S(\omega)}=-\frac{1}{2\ln 2}\left(\frac{1/N(\omega)}{1+S(\omega)/N(\omega)}\right)^2. \qquad \text{(A.356)}$$

We see from Eq. (A.356) that the second derivative is automatically negative, which means that if we find a function $S(\omega)$ such that the first derivative vanishes this will maximize $\tilde{R}$, as required. The vanishing of the first functional derivative requires, from Eq (A.355),

$$0 = \frac{\delta \bar{R}[S(\omega)]}{\delta S(\omega)} \tag{A.357}$$

$$= \frac{1}{2 \ln 2} \frac{1}{S(\omega) + N(\omega)} - \lambda \tag{A.358}$$

$$\lambda = \frac{1}{2 \ln 2} \frac{1}{S(\omega) + N(\omega)} \tag{A.359}$$

$$S(\omega) + N(\omega) = \frac{1}{2\lambda \ln 2}. \tag{A.360}$$

So we see that to maximize information transmission we must shape the signal spectrum to complement the noise spectrum, with the sum of the two being constant or "white," as schematized in Fig. 3.10.

Obviously the condition in Eq. (A.360) cannot be satisfied at frequencies where the noise is too large, that is where $N(\omega) > 1/2\lambda \ln 2$. At these forbidden frequencies it turns out that the optimum is just to set $S(\omega) = 0$. Again the graphical implementation of these ideas is shown in Fig. 3.10. To see how things work numerically, it is worthwhile to study a simple example.

Suppose that the power spectrum of the noise is given by

$$N(\omega) = N_0[1 + (\omega\tau)^2], \tag{A.361}$$

where $\tau$ sets the time resolution of the system, since frequencies above $\tau^{-1}$ are buried in much higher noise levels. Then to optimize information transmission we should choose the signal spectrum as

$$S_{\rm opt}(\omega) = \frac{1}{2\lambda \ln 2} - N(\omega) \tag{A.362}$$

$$= \frac{1}{2\lambda \ln 2} - N_0[1 + (\omega\tau)^2]. \tag{A.363}$$

Now this works only for frequencies in some range $-\omega_c < \omega < \omega_c$, where $\omega_c$ is defined by the vanishing of the optimal spectrum:

$$0 = S_{\rm opt}(\omega_c) \tag{A.364}$$

$$= \frac{1}{2\lambda \ln 2} - N_0[1 + (\omega_c\tau)^2]. \tag{A.365}$$

we can solve this equation to determine $\omega_c$, or we could just trade our (still unknown!) parameter $\lambda$ for a new parameter $\omega_c$. Let's follow this latter path, which leads us to write



$$S_{\rm opt}(\omega) = \frac{1}{2\lambda \ln 2} - N_0[1 + (\omega\tau)^2]$$

$$= N_0[1 + (\omega_c\tau)^2] - N_0[1 + (\omega\tau)^2] \tag{A.366}$$

$$= N_0(\omega_c\tau)^2[1 - (\omega/\omega_c)^2], \tag{A.367}$$

where we remember that $S_{\rm opt}(\omega) = 0$ for $|\omega| > \omega_c$.

Now we have to impose the constant variance constraint from Eq. (3.65), that will determine $\omega_c$:

$$\langle s^2 \rangle = \int_{\infty}^{\infty} \frac{d\omega}{2\pi} S(\omega)$$

$$= \int_{-\omega_c}^{\omega_c} \frac{d\omega}{2\pi} N_0(\omega_c\tau)^2[1 - (\omega/\omega_c)^2] \tag{A.368}$$

$$= N_0(\omega_c\tau)^2 2 \int_0^{\omega_c} \frac{d\omega}{2\pi}[1 - (\omega/\omega_c)^2] \tag{A.369}$$

$$= N_0(\omega_c\tau)^2 2\omega_c \int_0^1 \frac{dx}{2\pi}(1 - x^2), \tag{A.370}$$

where in the last step we change the integration variable to $x = \omega/\omega_c$. Then

$$\langle s^2 \rangle = N_0(\omega_c\tau)^2 2\omega_c \int_0^1 \frac{dx}{2\pi}(1 - x^2)$$

$$= N_0(\omega_c\tau)^2 \frac{\omega_c}{\pi}\left(1 - \frac{1}{3}\right) \tag{A.371}$$

$$= \frac{2}{3} N_0 \omega_c (\omega_c\tau)^2. \tag{A.372}$$

To obey this condition requires that we choose $\omega_c$ correctly:

$$\langle s^2 \rangle = \frac{2}{3} N_0 \omega_c (\omega_c\tau)^2$$

$$\Rightarrow \omega_c = \left[\frac{3\langle s^2 \rangle}{2N_0\tau^2}\right]^{1/3}. \tag{A.373}$$

Thus the optimal signal spectrum is spread over a bandwidth that increases as the 2/3 power of the signal to noise ratio $\langle s^2 \rangle / N_0$.

How much information is actually transmitted when we choose the spectrum to have the optimal form? This information transmission rate is, from Eq. (3.72),

$$R_{\text{info}} = \frac{1}{2}\int_{-\infty}^{\infty}\frac{d\omega}{2\pi}\log_2\left[1+\frac{S(\omega)}{N(\omega)}\right]\ \text{bits/s}$$

$$= \frac{1}{2}\int_{-\omega_c}^{\omega_c}\frac{d\omega}{2\pi}\log_2\left[1+\frac{N_0(\omega_c\tau)^2[1-(\omega/\omega_c)^2]}{N_0[1+(\omega\tau)^2]}\right] \tag{A.374}$$

$$= \frac{1}{2}\int_{-\omega_c}^{\omega_c}\frac{d\omega}{2\pi}\log_2\left[\frac{1+(\omega_c\tau)^2}{1+(\omega\tau)^2}\right]. \tag{A.375}$$

It appears that the natural units of $\omega$ are related to $\tau$, so we introduce a new integration variable $y = \omega\tau$ and find

$$R_{\text{info}} = \frac{1}{2}\int_{-\omega_c}^{\omega_c}\frac{d\omega}{2\pi}\log_2\left[\frac{1+(\omega_c\tau)^2}{1+(\omega\tau)^2}\right]$$

$$= \frac{1}{\tau}\int_0^{\omega_c\tau}\frac{dy}{2\pi}\log_2\left[\frac{1+(\omega_c\tau)^2}{1+y^2}\right] \tag{A.376}$$

$$= \frac{1}{2\pi\tau}\int_0^{\omega_c\tau}dy\left(\log_2[1+(\omega_c\tau)^2]-\log_2[1+y^2]\right) \tag{A.377}$$

$$= \frac{1}{2\pi\tau}\left\{\int_0^{\omega_c\tau}dy\log_2[1+(\omega_c\tau)^2]-\int_0^{\omega_c\tau}dy\log_2[1+y^2]\right\}. \tag{A.378}$$

The first integral we have to do is simple, because we are just integrating a constant:

$$\int_0^{\omega_c\tau}dy\log_2[1+(\omega_c\tau)^2] = \omega_c\tau\log_2[1+(\omega_c\tau)^2]. \tag{A.379}$$

The second integral is more difficult, but the related integral with a natural log can be found in most integral tables,

$$\int_0^{\omega_c\tau}dy\ln[1+y^2] = \omega_c\tau\ln[1+(\omega_c\tau)^2]-2\omega_c\tau+2\tan^{-1}(\omega_c\tau). \tag{A.380}$$

Remembering that $\log_2 Z = \ln Z/\ln 2$ for any $Z$, the integral we want to do can be written as



$$\int_0^{\omega_c\tau}dy\log_2[1+y^2] = \frac{1}{\ln 2}\int_0^{\omega_c\tau}dy\ln[1+y^2] \tag{A.381}$$

$$= \frac{1}{\ln 2}\{(\omega_c\tau)\ln[1+(\omega_c\tau)^2]-2\omega_c\tau+2\tan^{-1}(\omega_c\tau)\} \tag{A.382}$$

$$= \omega_c\tau\log_2[1+(\omega_c\tau)^2]+\frac{2}{\ln 2}[\tan^{-1}(\omega_c\tau)-\omega_c\tau]. \tag{A.383}$$

So finally we can substitute back into the expression for the information rate:

$$R_{\text{info}} = \frac{1}{2\pi\tau}\left\{\int_0^{\omega_c\tau}dy\log_2[1+(\omega_c\tau)^2]-\int_0^{\omega_c\tau}dy\log_2[1+y^2]\right\}$$

$$= \frac{1}{2\pi\tau}\left[\omega_c\tau\log_2[1+(\omega_c\tau)^2]\right.$$

$$\left.-\left(\omega_c\tau\log_2[1+(\omega_c\tau)^2]+\frac{2}{\ln 2}[\tan^{-1}(\omega_c\tau)-\omega_c\tau]\right)\right] \tag{A.384}$$

$$= \frac{1}{2\pi\tau}\cdot\frac{2}{\ln 2}\left[\omega_c\tau-\tan^{-1}(\omega_c\tau)\right] \tag{A.385}$$

$$= \frac{1}{\tau\pi\ln 2}\left[\omega_c\tau-\tan^{-1}(\omega_c\tau)\right]\ \text{bits/s}. \tag{A.386}$$

To make sense out of this answer it is useful to study the limit as the signal to noise ratio becomes large, so that from Eq. (A.373) the cutoff frequency $\omega_c$ also becomes large. In the brackets of the last expression, this makes the first term become large, while the second term approaches a constant ($\tan^{-1}x \to \pi/2$ as $x \to \infty$). Thus we can approximate

$$R_{\text{info}}(\omega_c\tau \gg 1) \sim \frac{1}{\tau\pi\ln 2}\omega_c\tau = \frac{\omega_c}{\pi\ln 2}\ \text{bits/s}, \tag{A.387}$$

which is a rather simple result.

It is convenient to define a dimensionless signal to noise ratio, $SNR = \langle s^2\rangle\tau/N_0$, and then the expression for the cutoff frequency, Eq. (A.373), becomes

$$\omega_c = \left[\frac{3\langle s^2\rangle}{2N_0\tau^2}\right]^{1/3} = \frac{1}{\tau}(3SNR/2)^{1/3}. \tag{A.388}$$

Then the information rate becomes

$$R_{\text{info}}(SNR \gg 1) \sim \frac{\omega_c}{\pi \ln 2} = \frac{1}{\tau}\frac{(3/2)^{1/3}}{\pi \ln 2}(SNR)^{1/3} \tag{A.389}$$

$$\sim \frac{(0.53)}{\tau}(SNR)^{1/3} \text{ bits/s}. \tag{A.390}$$

The rate of information transmission is measured naturally in units of the time resolution $\tau$, and the number of bits per time $\tau$ increases with the signal to noise ratio. The interesting point is that by optimizing the spectrum we can get the information to grow as the 1/3 power of the $SNR$, where from the basic Shannon formula one might have expected that the information grows only logarithmically with $SNR$. This difference—which is large when the $SNR$ is large—is what we gain by optimization.

## A.16 MAXIMUM LIKELIHOOD

In this section we develop the optimal strategy for discrimination between two alternative signals that are presented in a background of noise. Let us call the two alternatives + and −, and assume that we must base our decisions on the observation of a single variable $x$. If the signal is +, then the observable $x$ will be drawn from the distribution $P(x|+)$, and if the signal is − then the observable $x$ will be drawn from the distribution $P(x|-)$. These distributions tell us that if we know the signal, then we can predict the statistics of the observables. But we are interested in the opposite problem: If we have just seen a particular value of $x$, can we tell which signal was presented?

We are making a choice between just two alternatives, so each value of $x$ must be assigned either to + or to −. Intuitively we might think that one of the signals is bigger than the other, so that large values of $x$—larger than some critical value $x_0$—should be assigned to the signal + (for example) and small values of $x$ to the signal −. We will come back at the end of the discussion to a case where we need a more complex decision rule.

With our simple rule we divide the $x$ axis at the critical point $x_0$; everything to the right is called + and everything to the left is called −. How should we choose the location of the dividing line or threshold $x_0$? We want to make correct decisions as often as possible, so we should compute the probability



of a correct decision $P_c(x_0)$ as a function of the threshold, then find the value of the threshold that maximizes this probability.

If the signal really was + then values of $x$ are chosen from $P(x|+)$. But we will assign $x$ to + only if $x > x_0$. Thus the probability of correctly identifying the signal + is

$$P(\text{"say +"}|\text{signal is +}) \equiv P(+|+) \tag{A.391}$$

$$= \int_{x_0}^{\infty} dx\, P(x|+). \tag{A.392}$$

On the other hand, if the signal really was − then values of $x$ are chosen from $P(x|-)$ and we assign $x$ to − only if $x < x_0$. Thus the probability of correctly identifying the signal − is

$$P(\text{"say −"}|\text{signal is −}) \equiv P(-|-) \tag{A.393}$$

$$= \int_{-\infty}^{x_0} dx\, P(x|-). \tag{A.394}$$

Now the total probability of making the correct identification is determined by these factors and by the overall probability that the signal is + or −,

$$P_c(x_0) \equiv P(\text{signal is +}) \times P(\text{"say +"}|\text{signal is +}) + P(\text{signal is −}) \times P(\text{"say −"}|\text{signal is −}) \tag{A.395}$$

$$= P(+)P(+|+) + P(-)P(-|-) \tag{A.396}$$

$$= P(+)\int_{x_0}^{\infty} dx\, P(x|+) + P(-)\int_{-\infty}^{x_0} dx\, P(x|-) \tag{A.397}$$

Now we have to find the maximum of the expression in Eq. (A.397).

We recall that to find the maximum or minimum of a function we have to find a place where the derivative of the function is zero. In this case our function is defined as an integral. So it is useful to recognize that derivatives of integrals are especially simple:

$$\frac{d}{dy}\int_{y}^{\infty} dx f(x) = -f(y),$$

$$\frac{d}{dy}\int_{-\infty}^{y} dx f(x) = +f(y). \tag{A.398}$$

To find the condition that $P_c(x_0)$ is maximized we need to solve the equation

$$0 = \frac{dP_c(x_0)}{dx_0}. \tag{A.399}$$

We substitute the expression for $P_c(x_0)$ from Eq. (A.397), and then use the rules from Eq. (A.398):

$$0 = \frac{dP_c(x_0)}{dx_0}$$

$$= \frac{d}{dx_0}\left[P(+)\int_{x_0}^{\infty} dx\, P(x|+) + P(-)\int_{-\infty}^{x_0} dx\, P(x|-)\right] \quad \text{(A.400)}$$

$$= P(+)\left[\frac{d}{dx_0}\int_{x_0}^{\infty} dx\, P(x|+)\right] + P(-)\left[\frac{d}{dx_0}\int_{-\infty}^{x_0} dx\, P(x|-)\right] \quad \text{(A.401)}$$

$$= P(+)\left[-P(x_0|+)\right] + P(-)\left[P(x_0|-)\right]. \quad \text{(A.402)}$$

So we see that to maximize the probability of a correct decision, we should choose the threshold $x_0$ to satisfy the equation

$$0 = P(+)\left[-P(x_0|+)\right] + P(-)\left[P(x_0|-)\right]$$

$$P(+)P(x_0|+) = P(-)P(x_0|-). \quad \text{(A.403)}$$

In the simple case that the signals + and − are equally likely, $P(+) = P(-)$ and then we have to choose the threshold $x_0$ so that $P(x_0|+) = P(x_0|-)$. In words, the threshold has to be set at the point where the two probability distributions cross, as in Fig. 4.8.

More generally the threshold $x_0$ has to be chosen to satisfy Eq. (A.403). To understand this condition a little better, let us recall Bayes' rule, which tells us that, for example, the probability of the signal + given that we have observed the value $x$ is

$$P(+|x) = P(x|+) \times P(+) \times \frac{1}{P(x)} \quad \text{(A.404)}$$

$$= \frac{P(+)P(x|+)}{P(x)}, \quad \text{(A.405)}$$

where $P(x)$ is the overall probability of seeing the value $x$, averaged over the two possible signals,

$$P(x) = P(+)P(x|+) + P(-)P(x|-). \quad \text{(A.406)}$$

Similarly, the probability that the signal was − given that we observed $x$ is



$$P(-|x) = \frac{P(-)P(x|-)}{P(x)}. \quad \text{(A.407)}$$

Now the equation for the threshold $x_0$ can be manipulated a bit, starting with Eq. (A.403):

$$P(+)P(x_0|+) = P(-)P(x_0|-)$$

$$\frac{P(+)P(x_0|+)}{P(x_0)} = \frac{P(-)P(x_0|-)}{P(x_0)} \quad \text{(A.408)}$$

$$P(+|x_0) = P(-|x_0) \quad \text{(A.409)}$$

Thus the threshold must be set to the point where the signals + and − are equally probable. This means that we will maximize our probability of correctly identifying the signal if we always choose that signal that has the larger probability given the data we have observed; the decision boundary between the alternatives is the point where the probabilities are equal. This rule is called "maximum likelihood," and it generalizes to choosing among multiple alternatives.

If the $P(x|+)$ and $P(x|-)$ are both Gaussians with same variance but different means, then there is only one point $x_0$ that satisfies the condition in Eq. (A.409). On the other hand, if the two distributions are Gaussians with same mean but different variances, then there are two solutions and hence two dividing lines. But these are both special cases of the general rule: Assign the observed data $x$ to the signal that is most probable.

## A.17 POISSON AVERAGES

Here we indicate the steps involved in computing an average over the Poisson distribution of spike arrival times, as in Eq. (4.5). The quantity of interest is of the general form

$$\left\langle \sum_{i=1}^{N} f(t_i) \right\rangle_+ = \sum_{N=0}^{\infty} \int_0^T dt_1 \int_0^T dt_2 \cdots \int_0^T dt_N \sum_{i=1}^{N} f(t_i) P[t_1, t_2, \cdots, t_N|+]. \quad \text{(A.410)}$$

We substitute the expression for $P[t_1, t_2, \cdots, t_N|+]$ from the Poisson model, pull out the exponential factor that is independent of the $\{t_i\}$, and expand out the sum:

$$\left\langle \sum_{i=1}^{N} f(t_i) \right\rangle_+ = \sum_{N=0}^{\infty} \int_0^T dt_1 \int_0^T dt_2 \cdots \int_0^T dt_N \sum_{i=1}^{N} f(t_i) \times \exp\left[ -\int_0^T r_+(t) \right] \frac{1}{N!} r_+(t_1) r_+(t_2) \cdots r_+(t_N)$$

$$= \exp\left[ -\int_0^T dt\, r_+(t) \right] \times \sum_{N=0}^{\infty} \frac{1}{N!} \int_0^T dt_1 r_+(t_1) \int_0^T dt_2 r_+(t_2) \cdots \int_0^T dt_N r_+(t_N) \left[ f(t_1) + f(t_2) + \cdots + f(t_N) \right]. \tag{A.411}$$

In doing the integrals for Eq. (A.411), we notice that there are $N$ terms each of which is like

$$\int_0^T dt_1 r_+(t_1) \int_0^T dt_2 r_+(t_2) \cdots \int_0^T dt_N r_+(t_N) f(t_1).$$

In doing this integral we see that there are $N-1$ integrals over $r(t)$ and one integral ($t_1$ in this case) over the product $r(t)f(t)$:

$$\int_0^T dt_1 r_+(t_1) \int_0^T dt_2 r_+(t_2) \cdots \int_0^T dt_N r_+(t_N) f(t_1) = \left[ \int_0^T dt_1 r_+(t_1) f(t_1) \right] \left[ \int_0^T dt_2 r_+(t_2) \right] \cdots \left[ \int_0^T dt_N r_+(t_N) \right] \tag{A.412}$$

$$= \left[ \int_0^T dt\, r_+(t) f(t) \right] \times \left[ \int_0^T dt\, r_+(t) \right]^{N-1}. \tag{A.413}$$

But it doesn't really matter that we focused on the term with $f(t_1)$; the answer would have been the same for any $f(t_i)$. Thus when we sum over the $N$ terms, we obtain $N$ times the previous result,

$$\int_0^T dt_1 r_+(t_1) \int_0^T dt_2 r_+(t_2) \int_0^T dt_N r_+(t_N) \left[ f(t_1) + f(t_2) + \cdots + f(t_N) \right]$$



$$= N \left[ \int_0^T dt\, r_+(t) f(t) \right] \times \left[ \int_0^T dt\, r_+(t) \right]^{N-1} \tag{A.414}$$

So now we can substitute back into Eq. (A.411):

$$\left\langle \sum_{i=1}^{N} f(t_i) \right\rangle_+ = \exp\left[ -\int_0^T dt\, r_+(t) \right] \times \sum_{N=0}^{\infty} \frac{1}{N!} \int_0^T dt_1 r_+(t_1) \int_0^T dt_2 r_+(t_2) \cdots \int_0^T dt_N r_+(t_N) \left[ f(t_1) + f(t_2) + \cdots + f(t_N) \right]$$

$$= \exp\left[ -\int_0^T dt\, r_+(t) \right] \sum_{N=0}^{\infty} \frac{1}{N!} N \left[ \int_0^T dt\, r_+(t) \right]^{N-1} \times \int_0^T dt' r_+(t') f(t')$$

$$= \exp(-Q_+) \times \sum_{N=0}^{\infty} \frac{1}{N!} N Q_+^{N-1} \times F_+, \tag{A.415}$$

where we have defined the integral of the rate by analogy with Eq. (A.111),

$$Q_+ = \int_0^T dt\, r_+(t), \tag{A.416}$$

and the factor

$$F_+ = \int_0^T dt\, r_+(t) f(t). \tag{A.417}$$

To complete our calculation we have to sum the infinite series which appears in Eq. (A.415). As discussed in section A.5, the trick is to recognize that $N Q_+^{N-1}$ is the derivative of $Q_+^N$, and then see that we have the expansion of the exponential:

$$\sum_{N=0}^{\infty} \frac{1}{N!} N Q_+^{N-1} = \sum_{N=0}^{\infty} \frac{1}{N!} \frac{\partial}{\partial Q_+} Q_+^N \tag{A.418}$$

$$= \frac{\partial}{\partial Q_+} \sum_{N=0}^{\infty} \frac{1}{N!} Q_+^N \tag{A.419}$$

$$= \frac{\partial}{\partial Q_+} \exp(Q_+) = \exp(Q_+). \tag{A.420}$$

Finally we substitute back into Eq. (A.415) and cancel the exponentials, to give

$$\left\langle \sum_{i=1}^{N} f(t_i) \right\rangle_+ = \exp(-Q_+) \times \sum_{N=0}^{\infty} \frac{1}{N!} N Q_+^{N-1} \times F_+$$

$$= \exp(-Q_+) \times \exp(Q_+) \times F_+ = F_+ \tag{A.421}$$

$$= \int_0^T dt\, r_+(t) f(t), \tag{A.422}$$

which is the result promised in the text. We leave it as an exercise for the reader to show, using these same techniques, that

$$\left\langle \left[ \sum_{i=1}^{N} f(t_i) \right]^2 \right\rangle_+ = \left[ \int_0^T dt\, r_+(t) f(t) \right]^2 + \int_0^T dt\, r_+(t) [f(t)]^2, \tag{A.423}$$

so that the variance

$$\left\langle \left[ \delta \sum_{i=1}^{N} f(t_i) \right]^2 \right\rangle_+ = \int_0^T dt\, r_+(t) [f(t)]^2, \tag{A.424}$$

as in Eq. (4.12).

The results of this section are used in section 4.1.3, where we find that the change in mean of the log–likelihood ratio, Eq. (4.10), can be written as

$$\Delta M = \int_0^T dt \left[ r_+(t) - r_-(t) \right] \ln \left[ \frac{r_+(t)}{r_-(t)} \right]$$

$$= \int_0^T dt\, |\Delta r(t)| \ln \left[ \frac{r_-(t) + \Delta r(t)}{r_-(t)} \right] \tag{A.425}$$

$$= \int_0^T dt\, |\Delta r(t)| \ln \left[ 1 + \frac{\Delta r(t)}{r_-(t)} \right]. \tag{A.426}$$

We recall that this problem is interesting in the limit where discrimination is difficult and hence $\Delta r(t)$ is very small. Then we can expand the logarithm using the Taylor series illustrated in Fig. 2.8,

$$\ln \left[ 1 + \frac{\Delta r(t)}{r_-(t)} \right] \approx \frac{\Delta r(t)}{r_-(t)} - \frac{1}{2} \left( \frac{\Delta r(t)}{r_-(t)} \right)^2 + \cdots, \tag{A.427}$$



which is valid if $\Delta r(t)$ is small. We substitute into Eq. (A.426) and keep only the first term of the series, which will be the dominant contribution when the change in rate $\Delta r(t)$ is very small:

$$\Delta M = \int_0^T dt\, |\Delta r(t)| \ln \left[ 1 + \frac{\Delta r(t)}{r_-(t)} \right]$$

$$= \int_0^T dt\, |\Delta r(t)| \left[ \frac{\Delta r(t)}{r_-(t)} - \frac{1}{2} \left( \frac{\Delta r(t)}{r_-(t)} \right)^2 + \cdots \right] \tag{A.428}$$

$$\approx \int_0^T dt\, \frac{[\Delta r(t)]^2}{r_-(t)}. \tag{A.429}$$

Now we can write the factor $r_-(t)$ in this expression in terms of the average of the two rates, $r(t) = (1/2)[r_+(t) + r_-(t)]$, that is, $r_-(t) = r(t) - \Delta r(t)/2$. We could expand once more in a Taylor series, but it is clear that all but the first term in this series—which we obtain just by replacing $r_-(t)$ with $r(t)$—will involve higher powers of $\Delta r(t)$ and hence the errors we make by dropping these terms will be negligible. So we arrive, finally, at a simple formula for the change in mean value of the log-likelihood ratio,

$$\Delta M \approx \int_0^T dt\, \frac{[\Delta r(t)]^2}{r(t)}. \tag{A.430}$$

## A.18 SIGNAL TO NOISE RATIOS WITH WHITE NOISE

When we observe one variable, the picture in Fig. 4.19 shows how to compute the detectability of changes in this variable against a background of Gaussian noise. But we would like to generalize this picture to the case where the signals to be distinguished are functions of position in space, for the discussion of visual hyperacuity in section 4.2.1, or time, for the discussion of bat echolocation in section 4.2.3. It turns out that in both cases it is useful to study the limit in which the noise is "white"—statistically independent in each pixel of the image or at each instant of time in the echo waveform. We discuss explicitly the case of signals that vary in time, and then make what we hope is an obvious generalization to the problem of images.

Let us start by discretizing time into bins of size $\Delta\tau$; at the end of the calculation we will let these bins become arbitrarily small, so this discretization is not a significant restriction. Gaussian white noise is, by definition, a Gaussian random variable that fluctuates independently in each time bin, and we

call the variance of these fluctuations $\langle \eta^2 \rangle$. Imagine that we observe a signal $s(t)$ added to this background of noise, and let's call the observable quantity in time bin $n$

$$x(t_n) = s(t_n) + \eta_n, \tag{A.431}$$

where $s(t_n)$ is the signal in time bin $n$ and $\eta_n$ is the noise in this bin. Given the signal $s(t_n)$, the probability that we observe a particular $x(t_n)$ is given by

$$P[x(t_n)|s(t_n)] = P[\eta_n = x(t_n) - s(t_n)] \tag{A.432}$$

$$= \frac{1}{\sqrt{2\pi \langle \eta^2 \rangle}} \exp\left\{ -\frac{[x(t_n) - s(t_n)]^2}{2\langle \eta^2 \rangle} \right\}. \tag{A.433}$$

Now suppose that we want to know what happens not in one bin but in an entire interval of time from $t = 0$ to $t = T$, which contains $N = T/\Delta\tau$ bins. Because the noise is statistically independent in every bin, we can compute the condtitional probability for the *set* of values $x(t_1), x(t_2), \cdots, x(t_N)$ by multiplying the probabilities in the individual bins,

$$P[\{x(t_n)\}|\{s(t_n)\}] = \prod_{n=1}^{N} P[x(t_n)|s(t_n)] \tag{A.434}$$

$$= \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi \langle \eta^2 \rangle}} \exp\left\{ -\frac{[x(t_n) - s(t_n)]^2}{2\langle \eta^2 \rangle} \right\} \tag{A.435}$$

$$= \left[ \frac{1}{\sqrt{2\pi \langle \eta^2 \rangle}} \right]^N \prod_{n=1}^{N} \exp\left\{ -\frac{[x(t_n) - s(t_n)]^2}{2\langle \eta^2 \rangle} \right\} \tag{A.436}$$

$$= \left[ \frac{1}{\sqrt{2\pi \langle \eta^2 \rangle}} \right]^N \exp\left\{ -\sum_{n=1}^{N} \frac{[x(t_n) - s(t_n)]^2}{2\langle \eta^2 \rangle} \right\} \tag{A.437}$$

$$= \left[ \frac{1}{\sqrt{2\pi \langle \eta^2 \rangle}} \right]^{T/\Delta\tau} \exp\left\{ -\frac{1}{2\langle \eta^2 \rangle} \sum_{n=1}^{T/\Delta\tau} [x(t_n) - s(t_n)]^2 \right\}. \tag{A.438}$$

Now we said at the start that we would let the bins size $\Delta\tau$ become small. Then we know that sums over the discrete bins approximate integrals over continuous time, that is

$$\sum_{n=1}^{T/\Delta\tau} F(t_n) \to \frac{1}{\Delta\tau} \int_0^T dt\, F(t) \tag{A.439}$$



for any reasonably smooth function $F(t)$. In Eq. (A.438) the relevant sum is

$$\sum_{n=1}^{T/\Delta\tau} [x(t_n) - s(t_n)]^2 \to \frac{1}{\Delta\tau} \int_0^T dt\, |x(t) - s(t)|^2, \tag{A.440}$$

so that

$$P[\{x(t_n)\}|\{s(t_n)\}] = \left[ \frac{1}{\sqrt{2\pi \langle \eta^2 \rangle}} \right]^{T/\Delta\tau} \exp\left\{ -\frac{1}{2\langle \eta^2 \rangle} \sum_{n=1}^{T/\Delta\tau} [x(t_n) - s(t_n)]^2 \right\}$$

$$\to \left[ \frac{1}{\sqrt{2\pi \langle \eta^2 \rangle}} \right]^{T/\Delta\tau} \exp\left[ -\frac{1}{2\langle \eta^2 \rangle \Delta\tau} \int_0^T dt\, |x(t) - s(t)|^2 \right]. \tag{A.441}$$

Now as we pass to the limit of small bins, the probability distribution describing the values of the $x(t_n)$ in each bin becomes the probability distribution for the function $x(t)$, which is sometimes called a probability distribution functional (Feynman and Hibbs 1965). We write this schematically as

$$P[x(t_1), x(t_2), \cdots, x(t_N)|s(t_1), s(t_2), \cdots, s(t_N)] \to P[x(t)|s(t)], \tag{A.442}$$

and from Eq. (A.441) we have

$$P[x(t)|s(t)] \propto \exp\left[ -\frac{1}{2\langle \eta^2 \rangle \Delta\tau} \int_0^T dt\, |x(t) - s(t)|^2 \right]. \tag{A.443}$$

The quantity that seems to arise naturally in this discussion is $N_0 = \langle \eta^2 \rangle \Delta\tau$, and we notice that this has the units of (noise)$^2$ · (time), or (noise)$^2$/frequency, which are also the units for the spectral density of the noise $\eta(t)$. Indeed, $N_0$ is exactly the power spectrum of the noise as we have defined it in section 3.1.4, and the fact that the entire spectrum comes out as one number is because we have assumed that the noise is white, so the spectrum is constant.

Thus, if the background noise is Gaussian and white, with spectral density $N_0$, then the probability of observing a particular waveform $x(t)$ is

$$P[x(t)|s_1(t)] \propto \exp\left[ -\frac{1}{2N_0} \int dt\, |x(t) - s_1(t)|^2 \right], \tag{A.444}$$

where $s_1(t)$ is one possible signal. Similarly, if the signal is $s_2(t)$, the distribution has the same form but with the mean waveform $s_1(t)$ replaced by $s_2(t)$:

$$P[x(t)|s_2(t)] \propto \exp\left[-\frac{1}{2N_0}\int dt\,|x(t)-s_2(t)|^2\right]. \tag{A.445}$$

Now in a standard forced choice experiment, if we observe a particular waveform $x(t)$, we can decide whether the signal is $s_1(t)$ or $s_2(t)$ by computing the logarithm of the relative likelihood, as in the discussion of Eq. (A.258):

$$\lambda[x(t)] = \ln\left(\frac{P[x(t)|s_1(t)]}{P[x(t)|s_2(t)]}\right) \tag{A.446}$$

$$= \frac{1}{2N_0}\int dt\,[|x(t)-s_1(t)|^2 - |x(t)-s_2(t)|^2] + \text{constant} \tag{A.447}$$

$$= \frac{1}{2N_0}\int dt\,[x^2(t) - 2x(t)s_1(t) + s_1^2(t)] - \frac{1}{2N_0}\int dt\,[x^2(t) - 2x(t)s_2(t) + s_2^2(t)] + \text{constant} \tag{A.448}$$

$$= \frac{1}{2N_0}\int dt\,\{-2x(t)[s_1(t)-s_2(t)] + s_1^2(t) - s_2^2(t)\} + \text{constant} \tag{A.449}$$

$$= \frac{1}{N_0}\int dt\,x(t)[s_2(t)-s_1(t)] + \text{stuff}. \tag{A.450}$$

In these manipulations we have introduced "constant" to keep track of the proportionality constants in Eq's. (A.444) and (A.445), and we collect in "stuff" all the terms in $\lambda[x(t)]$ that don't depend on $x(t)$.

We see that the log-likelihood ratio, $\lambda[x(t)]$, which is all we need to make optimal discriminations between the two stimuli $s_1(t)$ and $s_2(t)$, is a linear functional of $x(t)$. But the waveform $x(t)$ consists of the signal plus Gaussian white noise, which means that $x(t)$ is a Gaussian random variable, and any linear combination of the $x(t)$ at different times, such as $\lambda[x(t)]$, is also a Gaussian random variable. But this means that our original problem of discriminating between functions has been reduced to the problem of discrimination based on one variable, namely $\lambda$ itself, and this discrimination problem is exactly the problem described in Fig. 4.19. The "signal" for discrimination comes from the fact that the average values of $\lambda[x(t)]$ are different in the two distributions $P[x(t)|s_1(t)]$ and $P[x(t)|s_2(t)]$. To calculate this difference we



note that the average value of $x(t)$ always the signal, so that

$$\langle\lambda[x(t)]\rangle_1 = \frac{1}{N_0}\int dt\,\langle x(t)\rangle_1[s_2(t)-s_1(t)] + \text{stuff} \tag{A.451}$$

$$= \frac{1}{N_0}\int dt\,s_1(t)[s_2(t)-s_1(t)] + \text{stuff}, \tag{A.452}$$

and similarly

$$\langle\lambda[x(t)]\rangle_2 = \frac{1}{N_0}\int dt\,\langle x(t)\rangle_2[s_2(t)-s_1(t)] + \text{stuff} \tag{A.453}$$

$$= \frac{1}{N_0}\int dt\,s_2(t)[s_2(t)-s_1(t)] + \text{stuff}, \tag{A.454}$$

so that the difference in means is given by

$$\langle\lambda[x(t)]\rangle_1 - \langle\lambda[x(t)]\rangle_2 = \frac{1}{N_0}\int dt\,s_1(t)[s_2(t)-s_1(t)] + \text{stuff} - \frac{1}{N_0}\int dt\,s_2(t)[s_2(t)-s_1(t)] - \text{stuff} \tag{A.455}$$

$$= \frac{1}{N_0}\int dt\,[s_1(t)-s_2(t)][s_2(t)-s_1(t)] \tag{A.456}$$

$$= -\frac{1}{N_0}\int dt\,[s_1(t)-s_2(t)]^2. \tag{A.457}$$

This determines the "signal" for the discrimination, and to find the noise we need to compute the variance of $\lambda[x(t)]$.

To find the variance of $\lambda[x(t)]$ we can throw away any terms that do not fluctuate. Thus we can discard "stuff," which is independent of $x(t)$, and we can subtract from $x(t)$ its mean value to leave just the noise $\eta(t)$. Then the fluctuation in $\lambda[x(t)]$ is given by

$$\delta\lambda[x(t)] = \frac{1}{N_0}\int dt\,\eta(t)[s_2(t)-s_1(t)], \tag{A.458}$$

and the variance is

$$\langle(\delta\lambda[x(t)])^2\rangle = \left\langle\left[\frac{1}{N_0}\int dt\,\eta(t)[s_2(t)-s_1(t)]\right]^2\right\rangle \tag{A.459}$$

$$= \frac{1}{N_0^2}\left\langle \int dt\, \eta(t)[s_2(t) - s_1(t)] \int dt'\, \eta(t')[s_2(t') - s_1(t')] \right\rangle \tag{A.460}$$

$$= \frac{1}{N_0^2} \int dt \int dt' \langle \eta(t)\eta(t') \rangle [s_2(t) - s_1(t)][s_2(t') - s_1(t')], \tag{A.461}$$

Now white noise has a very simple correlation function, proportional to the Dirac delta function defined in section A.1, and the constant of proportionality is just the power spectrum, so that

$$\langle \eta(t)\eta(t') \rangle = N_0 \delta(t - t'). \tag{A.462}$$

Substituting into Eq. (A.461), we find that

$$\left\langle (\delta\lambda[x(t)])^2 \right\rangle = \frac{1}{N_0^2} \int dt \int dt' \langle \eta(t)\eta(t') \rangle [s_2(t) - s_1(t)][s_2(t') - s_1(t')]$$

$$= \frac{1}{N_0^2} \int dt \int dt' N_0 \delta(t - t')[s_2(t) - s_1(t)][s_2(t') - s_1(t')] \tag{A.463}$$

$$= \frac{1}{N_0} \int dt \int dt' \delta(t - t')[s_2(t) - s_1(t)][s_2(t') - s_1(t')] \tag{A.464}$$

$$= \frac{1}{N_0} \int dt \int dt' \delta(t - t')[s_2(t) - s_1(t)]^2, \tag{A.465}$$

where in the last step we use the fact that $\delta(t - t') = 0$ for $t \neq t'$ so we know that for any smooth function in the integral we can set $t = t'$. But now we can do the integral over $t'$, since, as in Eq. (A.10),

$$\int dt' \delta(t - t') = 1, \tag{A.466}$$

so that

$$\left\langle (\delta\lambda[x(t)])^2 \right\rangle = \frac{1}{N_0} \int dt \int dt' \delta(t - t')[s_2(t) - s_1(t)]^2$$

$$= \frac{1}{N_0} \int dt\, [s_2(t) - s_1(t)]^2 \int dt' \delta(t - t') \tag{A.467}$$



$$= \frac{1}{N_0} \int dt\, [s_2(t) - s_1(t)]^2. \tag{A.468}$$

Finally we can put the pieces together and find the signal to noise ratio for discrimination between signals $s_1(t)$ and $s_2(t)$ in a white noise background:

$$SNR = \left[\langle \lambda[x(t)] \rangle_1 - \langle \lambda[x(t)] \rangle_2\right]^2 \times \left[\left\langle (\delta\lambda[x(t)])^2 \right\rangle\right]^{-1} \tag{A.469}$$

$$= \left[-\frac{1}{N_0} \int dt\, [s_1(t) - s_2(t)]^2\right]^2 \left[\frac{1}{N_0} \int dt\, [s_1(t) - s_2(t)]^2\right]^{-1} \tag{A.470}$$

$$= \frac{1}{N_0} \int dt\, [s_1(t) - s_2(t)]^2 \tag{A.471}$$

$$= \frac{1}{N_0} \int dt\, [\Delta s(t)]^2, \tag{A.472}$$

where we write the answer in terms of the difference in signals, $\Delta s(t) = s_1(t) - s_2(t)$.

The result in Eq. (A.472) has a clear generalization to the discrimination between two images. We imagine that these two images generate patterns of light intensity $I_1(x)$ and $I_2(x)$, and that the white noise is equivalent to noise in the measurement of these intensities. Then the signal to noise ratio for discrimination is essentially the same as for the time dependent signals, with the pixels in the image playing the role of the discrete time bins, so that

$$SNR = \frac{1}{N_0} \int d^2x\, [\Delta I(x)]^2. \tag{A.473}$$

This is the formula we need in section 4.2.1.

For the discussion of bat echolocation in section 4.2.3, the two waveforms that the bat must discriminate between are both stereotyped echo waveforms $s_0(t - \tau)$, with $\tau$ the echo delay. The difference between $s_1(t)$ and $s_2(t)$ is only the value of the delay, so that

$$s_1(t) = s_0(t - \tau) \tag{A.474}$$

$$s_2(t) = s_0(t - \tau - \delta\tau), \tag{A.475}$$

where $\delta\tau$ is the small target jitter. Now the difference waveform is

$$\Delta s(t) = s_1(t) - s_2(t)$$

$$= s_0(t - \tau) - s_0(t - \tau - \delta\tau) \tag{A.476}$$

$$\approx s_0(t-\tau) - \left[ s_0(t-\tau) - \delta\tau \frac{ds_0(t-\tau)}{dt} + \cdots \right] \tag{A.477}$$

$$= \delta\tau \frac{ds_0(t-\tau)}{dt}, \tag{A.478}$$

where we use the Taylor expansion [see Eq. (A.37)] for small $\delta\tau$. Substituting into Eq. (A.472), we find the signal to noise ratio for jitter discrimination,

$$SNR = \frac{1}{N_0} \int dt\, [\Delta s(t)]^2$$

$$= \frac{1}{N_0} \int dt \left[ \delta\tau \frac{ds_0(t-\tau)}{dt} \right]^2 \tag{A.479}$$

$$= (\delta\tau)^2 \frac{1}{N_0} \int dt \left[ \frac{ds_0(t-\tau)}{dt} \right]^2, \tag{A.480}$$

as promised in the text.

## A.19 OPTIMAL FILTERS

In this section we consider the problem of maximizing the information transmission through a filter. Signals are presented in a background of noise, the output of the filter has limited dynamic range, and there is also noise at the output. Our task is to shape the filter characteristic both to protect the signal and to make maximum use of the available dynamic range. For simplicity we take the signals and noises to come from Gaussian distributions, so they are described completely once we know their power spectra. Furthermore we assume that the filter is linear, which also simplifies the calculations enormously. It turns out that this isn't really an assumption, since with Gaussian signals and noise one can show that our optimization problem is solved by a linear filter — given our phrasing of the constraints, nonlinearities won't help to transmit more information.

The setup of our problem is from section 5.3: We have a signal $s(t)$ which is added to a background of noise $\eta_1(t)$, then this combination is filtered by a device with impulse response $F(\tau)$, and finally a noise $\eta_2(t)$ is added at the output to produce a voltage $V(t)$:

$$V(t) = \int dt'\, F(t-t')[s(t') + \eta_1(t')] + \eta_2(t). \tag{A.481}$$

It is useful to think about how this works in the frequency domain, so we Fourier transform both sides of the equation, as in the discussion of impedance in section 2.1.3, from Eq. (2.7) to Eq. (2.10):



$$\int dt\, \exp(+i\omega t) V(t) = \int dt\, \exp(+i\omega t) \int dt'\, F(t-t')[s(t') + \eta_1(t')]$$

$$+ \int_{-\infty}^{\infty} dt\, \exp(+i\omega t)\eta_2(t) \tag{A.482}$$

$$\tilde{V}(\omega) = \int dt\, \exp(+i\omega t) \int dt'\, F(t-t')[s(t') + \eta_1(t')]$$

$$+ \int dt\, \exp(+i\omega t)\eta_2(t) \tag{A.483}$$

$$= \int dt \int dt'\, \exp(+i\omega t) F(t-t')[s(t') + \eta_1(t')]$$

$$+ \tilde{\eta}_2(\omega) \tag{A.484}$$

$$= \int dt \int dt'\, \exp[+i\omega(t-t')]$$

$$\times F(t-t') \exp(i\omega t')[s(t') + \eta_1(t')] + \tilde{\eta}_2(\omega). \tag{A.485}$$

To finish the calculation it is convenient to change variables in the integral. Instead of integrating over $t$ and $t'$, we switch to $t'$ and $\tau$, where the new variable is defined as $\tau = t - t'$. Then we have

$$\tilde{V}(\omega) = \int dt \int dt'\, \exp[+i\omega(t-t')] F(t-t') \exp(i\omega t')[s(t') + \eta_1(t')]$$

$$+ \tilde{\eta}_2(\omega)$$

$$= \int d\tau \exp(i\omega\tau) F(\tau) \int dt'\, \exp(i\omega t')[s(t') + \eta_1(t')]$$

$$+ \tilde{\eta}_2(\omega) \tag{A.486}$$

$$= \left[ \int d\tau \exp(i\omega\tau) F(\tau) \right]$$

$$\times \left[ \int dt'\, \exp(i\omega t') s(t') + \int dt'\, \exp(i\omega t')\eta_1(t') \right]$$

$$+ \tilde{\eta}_2(\omega) \tag{A.487}$$

$$= \tilde{F}(\omega)[\tilde{s}(\omega) + \tilde{\eta}_1(\omega)] + \tilde{\eta}_2(\omega). \tag{A.488}$$

So we see that, frequency component by frequency component, the output voltage consists of one term proportional to the signal $\tilde{s}(\omega)$ and two terms which reflect the noises at the input and output, $\tilde{\eta}_1$ and $\tilde{\eta}_2$, respectively.

We are interested in knowing how much information the output voltage $V(t)$ provides about the input signal $s(t)$. If we look at frequency component $\omega$, it is clear that the piece of $\tilde{V}(\omega)$ corresponding to the signal is $\tilde{F}(\omega)\tilde{s}(\omega)$, while the piece corresponding to noise is $\tilde{F}(\omega)\eta_1(\omega) + \eta_2(\omega)$. Recall from Eq. (3.66) that we can find the power spectrum of the signal $s(t)$ by normalizing the variances of the Fourier coefficients, so that

$$\langle \tilde{s}(\omega)\tilde{s}(-\omega)\rangle = \langle|\tilde{s}(\omega)|^2\rangle = TS(\omega), \text{ or} \tag{A.489}$$

$$S(\omega) = \frac{1}{T}\langle|\tilde{s}(\omega)|^2\rangle. \tag{A.490}$$

In the present case the effective signal has Fourier components $\tilde{F}(\omega)\tilde{s}(\omega)$, so the effective signal power spectrum is given by

$$S_{\text{eff}}(\omega) = \frac{1}{T}\langle|\tilde{F}(\omega)\tilde{s}(\omega)|^2\rangle \tag{A.491}$$

$$= |\tilde{F}(\omega)|^2\frac{1}{T}\langle|\tilde{s}(\omega)|^2\rangle = |\tilde{F}(\omega)|^2 S(\omega). \tag{A.492}$$

The effective noise spectrum is found by the same manipulations:

$$N_{\text{eff}}(\omega) = \frac{1}{T}\langle|\tilde{F}(\omega)\eta_1(\omega) + \eta_2(\omega)|^2\rangle \tag{A.493}$$

$$= |\tilde{F}(\omega)|^2\frac{1}{T}\langle|\eta_1(\omega)|^2\rangle + \frac{1}{T}\langle|\eta_2(\omega)|^2\rangle, \tag{A.494}$$

where in the last step we use the fact that the two noise sources $\eta_1(t)$ and $\eta_2(t)$ are statistically independent. We see that the effective noise spectrum can be written in terms of the power spectra for the individual noise components, which we call $N_1$ and $N_2$, and we assume for simplicity that these noises are white. Thus

$$N_{\text{eff}}(\omega) = |\tilde{F}(\omega)|^2\frac{1}{T}\langle|\eta_1(\omega)|^2\rangle + \frac{1}{T}\langle|\eta_2(\omega)|^2\rangle = |\tilde{F}(\omega)|^2 N_1 + N_2. \tag{A.495}$$

Finally we put these expressions together to obtain the signal to noise ratio at each frequency,

$$SNR(\omega) \equiv \frac{S_{\text{eff}}(\omega)}{N_{\text{eff}}(\omega)} \tag{A.496}$$

$$= \frac{|\tilde{F}(\omega)|^2 S(\omega)}{|\tilde{F}(\omega)|^2 N_1 + N_2}. \tag{A.497}$$



The rate at which the output $V(t)$ provides information about the input signal $s(t)$ can be calculated from Shannon's formula in Eq. (3.72),

$$R_{\text{info}} = \frac{1}{2}\int_{-\infty}^{\infty}\frac{d\omega}{2\pi}\log_2[1 + SNR(\omega)]$$

$$= \frac{1}{2}\int_{-\infty}^{\infty}\frac{d\omega}{2\pi}\log_2\left[1 + \frac{|\tilde{F}(\omega)|^2 S(\omega)}{|\tilde{F}(\omega)|^2 N_1 + N_2}\right] \tag{A.498}$$

we want to find the filter characteristic $\tilde{F}(\omega)$ which maximizes this information flow subject to the constraint that the variance of the output is fixed. The output variance is related to the power spectra by a generalization of Eq. (3.65),

$$\langle V^2\rangle = \int_{-\infty}^{\infty}\frac{d\omega}{2\pi}[S_{\text{eff}}(\omega) + N_{\text{eff}}(\omega)] \tag{A.499}$$

$$= \int_{-\infty}^{\infty}\frac{d\omega}{2\pi}[|\tilde{F}(\omega)|^2 S(\omega) + |\tilde{F}(\omega)|^2 N_1 + N_2]. \tag{A.500}$$

As in previous sections, we use the method of Lagrange multipliers to maximize $R_{\text{info}}$ while holding $\langle V^2\rangle$ fixed.

We define a new quantity $\tilde{R}$ and a Lagrange multiplier $\lambda$,

$$\tilde{R} \equiv R_{\text{info}} - \lambda\langle V^2\rangle \tag{A.501}$$

$$= \frac{1}{2}\int_{-\infty}^{\infty}\frac{d\omega}{2\pi}\log_2\left[1 + \frac{|\tilde{F}(\omega)|^2 S(\omega)}{|\tilde{F}(\omega)|^2 N_1 + N_2}\right]$$

$$- \lambda\int_{-\infty}^{\infty}\frac{d\omega}{2\pi}[|\tilde{F}(\omega)|^2 S(\omega) + |\tilde{F}(\omega)|^2 N_1 + N_2]. \tag{A.502}$$

We want to examine how $\tilde{R}$ changes when we make a small change in the filter $\tilde{F}(\omega)$, isolate the functional derivatives and find the condition for a maximum. We notice that $\tilde{R}$ depends on $|\tilde{F}(\omega)|^2$, so we will use this as the independent variable. So we begin with

$$\tilde{R}[|\tilde{F}(\omega)|^2 + \delta|\tilde{F}(\omega)|^2]$$

$$= \frac{1}{2}\int_{-\infty}^{\infty}\frac{d\omega}{2\pi}\log_2\left[1 + \frac{[|\tilde{F}(\omega)|^2 + \delta|\tilde{F}(\omega)|^2]S(\omega)}{[|\tilde{F}(\omega)|^2 + \delta|\tilde{F}(\omega)|^2]N_1 + N_2}\right]$$

$$- \lambda\int_{-\infty}^{\infty}\frac{d\omega}{2\pi}[|\tilde{F}(\omega)|^2 S(\omega) + |\tilde{F}(\omega)|^2 N_1 + N_2]$$

$$- \lambda\int_{-\infty}^{\infty}\frac{d\omega}{2\pi}\delta|\tilde{F}(\omega)|^2[S(\omega) + N_1]. \tag{A.503}$$

As in previous calculations, all of the difficulty comes from the logarithmic term, which we work on first, starting by converting to natural logarithms:

$$\frac{1}{2}\int_{-\infty}^{\infty}\frac{d\omega}{2\pi}\log_2\left[1+\frac{[|\tilde{F}(\omega)|^2+\delta|\tilde{F}(\omega)|^2]S(\omega)}{[|\tilde{F}(\omega)|^2+\delta|\tilde{F}(\omega)|^2]N_1+N_2}\right]$$
$$=\frac{1}{2\ln 2}\int_{-\infty}^{\infty}\frac{d\omega}{2\pi}\ln\left[1+\frac{[|\tilde{F}(\omega)|^2+\delta|\tilde{F}(\omega)|^2]S(\omega)}{[|\tilde{F}(\omega)|^2+\delta|\tilde{F}(\omega)|^2]N_1+N_2}\right]. \tag{A.504}$$

Now since $\delta|\tilde{F}(\omega)|^2$ is small, it is useful to recall that

$$\frac{1}{A+\delta B}=\frac{1}{A}-\frac{1}{A^2}\delta B+\frac{1}{A^3}[\delta B]^2-\cdots \tag{A.505}$$
$$=\frac{1}{A}\left(1-\frac{1}{A}\delta B+\frac{1}{A^2}[\delta B]^2-\cdots\right). \tag{A.506}$$

for any small $\delta B$. The steps should now be familiar: We use this expansion, together with the Taylor expansion of the natural logarithm, then collect terms which have the same powers of $\delta|\tilde{F}(\omega)|^2$. This allows us to identify the functional derivative of $\tilde{R}$:

$$\frac{\delta\tilde{R}}{\delta|\tilde{F}(\omega)|^2}=\frac{1}{2\ln 2}\cdot\frac{S(\omega)}{|\tilde{F}(\omega)|^2N_1+N_2}\cdot\frac{N_2}{|\tilde{F}(\omega)|^2[S(\omega)+N_1]+N_2}$$
$$-\lambda[S(\omega)+N_1]. \tag{A.507}$$

We can use this to find the optimal filter, setting the functional derivative to zero and solving for $|\tilde{F}(\omega)|^2$. But is more enlightening to look at some limiting cases.

If the signal to noise ratio is large, we can take $S(\omega)\to\infty$ in Eq. (A.507), so that

$$\frac{\delta\tilde{R}}{\delta|\tilde{F}(\omega)|^2}\approx\frac{1}{2\ln 2}\cdot\frac{N_2}{|\tilde{F}(\omega)|^2N_1+N_2}\cdot\frac{1}{|\tilde{F}(\omega)|^2}-\lambda S(\omega). \tag{A.508}$$

Then the condition that the functional derivative equal zero, optimizing information transmission, is just

$$0=\frac{\delta\tilde{R}}{\delta|\tilde{F}(\omega)|^2} \tag{A.509}$$
$$\approx\frac{1}{2\ln 2}\cdot\frac{N_2}{|\tilde{F}(\omega)|^2N_1+N_2}\cdot\frac{1}{|\tilde{F}(\omega)|^2}-\lambda S(\omega) \tag{A.510}$$



$$\lambda S(\omega)=\frac{1}{2\ln 2}\cdot\frac{N_2}{|\tilde{F}(\omega)|^2N_1+N_2}\cdot\frac{1}{|\tilde{F}(\omega)|^2} \tag{A.511}$$

$$|\tilde{F}(\omega)|^2=\frac{1}{S(\omega)}\cdot\frac{1}{2\lambda\ln 2}\cdot\frac{N_2}{|\tilde{F}(\omega)|^2N_1+N_2}. \tag{A.512}$$

Now with $S(\omega)$ very large, this equation is telling us that the filter $|\tilde{F}(\omega)|^2$ should be very small, so we can neglect $|\tilde{F}(\omega)|^2N_1$ compared to $N_2$ on the right hand side of Eq. (A.512). Thus we have the simple result

$$|\tilde{F}(\omega)|^2=\frac{1}{S(\omega)}\cdot\frac{1}{2\lambda\ln 2}. \tag{A.513}$$

This is the idea discussed in section 5.3—at high signal to noise ratios, the optimal encoding filter has a frequency dependence that cancels the input power spectrum, so that the output voltage of the filter has a power spectrum

$$S_{\rm out}(\omega)=|\tilde{F}(\omega)|^2S(\omega) \tag{A.514}$$

that is constant or white. It is clear from the more general Eq. (A.507) that this cannot continue once we reach a frequency range where the signal to noise ratio is smaller, but the details of how the optimal filter "rolls over" to exclude the noise $N_1$ depends on the precise value of $N_2$ and the exact shape of the signal spectrum $S(\omega)$. For a more expanded view of these optimization arguments see Atick (1992).

# References


Abeles, M., H. Bergman, E. Margalit, and E. Vaadia (1993). Spatiotemporal firing patterns in the frontal cortex of behaving monkeys, *J. Neurophysiol.* 70, 1629–1638.

Abragam, A. (1983). *Principles of Nuclear Magnetism,* paperback edition (Oxford University Press, Oxford).

Adrian, E. D. (1926). The impulses produced by sensory nerve endings: Part I, *J. Physiol. (Lond.)* 61, 49–72.

Adrian, E. D. (1928). *The Basis of Sensation: The Action of the Sense Organs* (W. W. Norton, New York).

Adrian, E. D. (1932). *The Mechanism of Nervous Action: Electrical Studies of the Neurone* (University of Pennsylvania Press, Philadelphia).

Adrian, E. D. (1947). *The Physical Background of Perception; being the Waynflete Lectures delivered in the College of St. Mary Magdalen, Oxford, in Hilary term 1946* (Oxford University Press, Oxford).

Adrian, E. D., and Y. Zotterman (1926a). The impulses produced by sensory nerve endings: Part II: The response of a single end organ, *J. Physiol. (Lond.)* 61, 151–171.

Adrian, E. D., and Y. Zotterman (1926b). The impulses produced by sensory nerve endings: Part III: Impulses set up by touch and pressure, *J. Physiol. (Lond.)* 61, 465–483.

Aho, A.-C., K. Donner, C. Hydén, L. O. Larsen, and T. Reuter (1988). Low retinal noise in animals with low body temperature allows high visual sensitivity, *Nature* 334, 348–350.

Aidley, D. J., (1989). *The Physiology of Excitable Cells, Third Edition* (Cambridge University Press, Cambridge).

Allen, C., and C. F. Stevens (1994). An evaluation of causes for unreliability of synaptic transmission, *Proc. Nat. Acad. Sci. USA* 91, 10380–10383.

Altes, R. A. (1989). Ubiquity of hyperacuity, *J. Acoust. Soc. Am.* 85, 943–952.

Aronson, D. G., and H. F. Weinberger (1978). Multidimensional nonlinear diffusion arising in population genetics, *Adv. Math.* 30, 33–76.

Atema, J. (1995). Chemical signals in the marine environment: Dispersal, detection and temporal signal analysis, *Proc. Nat. Acad. Sci. USA* 92, 62–66.

Atick, J. J. (1992). Could information theory provide an ecological theory of sensory processing?. In *Princeton Lectures on Biophysics*, W. Bialek, ed., pp. 223–289 (World Scientific, Singapore).

Atick, J. J., Z. Li, and A. N. Redlich (1992). Understanding retinal color coding from first principles, *Neural Comp.* 4, 559–572.

Atick, J. J., and A. N. Redlich (1990). Towards a theory of early visual processing, *Neural Comp.* 2, 308–320.

Bair, W. (1995). The analysis of temporal structure in spike trains of visual cortical area MT. Dissertation, California Institute of Technology.

Bair, W., and C. Koch (1996). Temporal precision of spike trains in extrastriate cortex of the behaving macaque monkey, *Neural Comp.* 8, 1184–1202.

Bair, W., E. Zohary, and C. Koch (1996). Correlated neuronal response: Time scales and mechanisms. In *Advances in Neural Information Processing Systems 8*, D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, eds., pp. 68–74 (MIT Press, Cambridge MA).

Barlow, H. B. (1952). The size of ommatida in apposition eyes, *J. Exp. Biol.* 29, 667–674.

Barlow, H. B. (1953a). Action potentials from the frog's retina, *J. Physiol. (Lond.)* 119, 58–68.

Barlow, H. B. (1953b). Summation and inhibition in the frog's retina, *J. Physiol. (Lond.)* 119, 69–88.

Barlow, H. B. (1956). Retinal noise and absolute threshold, *J. Opt. Soc. Am.* 46, 634–639.

Barlow, H. B. (1961). Possible principles underlying the transformation of sensory messages. In *Sensory Communication*, W. Rosenblith, ed., pp. 217–234 (MIT Press, Cambridge MA).

Barlow, H. B. (1972). Single units and sensation: A neuron doctrine for perception, *Perception* 1, 371–394.

Barlow, H. B. (1980). The absolute efficiency of perceptual decisions, *Philos. Trans. R. Soc. Lond. Ser. B* 290, 71–82.

Barlow, H. B. (1981). Critical limiting factors in the design of the eye and visual cortex, *Proc. R. Soc. Lond. Ser. B* 212, 1–34.

Barlow, H. B. (1982). What causes trichromacy? A theoretical analysis using comb-filtered spectra, *Vision Res.* 22, 635-643.

Barlow, H. B. (1988). The thermal limit to seeing, *Nature* 334, 296–297.

Barlow, H. B., R. FitzHugh, and S. W. Kuffler (1957). Change of organization in the receptive fields of the cat's retina during dark adaptation *J. Physiol.* 137, 338–354.

Barlow, H. B., and W. Levick (1969). Three factors limiting the reliable detection of light by the retinal ganglion cells of the cat, *J. Physiol. (Lond.)* 200, 1–24



Barlow, H. B., W. R. Levick, M. Yoon (1971). Responses to single quanta of light in retinal ganglion cells of the cat, *Vision Res. Suppl.* 3, 87–101.

Barth, F. G., U. Wastl, J. A. C. Humphrey, and R. Devarkonda (1993). Dynamics of arthropod filiform hairs. II: Mechanical properties of spider trichobothria (*Cupiennius salei* Keys.), *Phil. Trans. R. Soc. Ser. B* 340, 445–461.

Baylor, D. A., and A. L. Hodgkin (1974). Changes in time course and sensitivity in turtle photoreceptors, *J. Physiol. (Lond.)* 242, 729–758.

Baylor, D. A., T. D. Lamb, and K.–W. Yau (1979a). The membrane current of single rod outer segments, *J. Physiol. (Lond.)* 288, 589–634.

Baylor, D. A., T. D. Lamb, and K.–W. Yau (1979b). Responses of retinal rods to single photons, *J. Physiol. (Lond.)* 288, 613–634.

Baylor, D. A., G. Matthews, and K.–W. Yau (1980). Two components of electrical dark noise in toad retinal rod outer segments, *J. Physiol. (Lond.)* 309, 591–621.

Baylor, D. A., B. J. Nunn, and J. F. Schnapf (1984). The photocurrent, noise and spectral sensitivity of rods of the monkey *Macaca fascicularis*, *J. Physiol. (Lond.)* 357, 575–607.

Becker, S. (1996). Mutual information maximization: Models of cortical self-organization, *Network* 7, 7–31.

Bekkers, J. M., and C. F. Stevens (1994). The nature of quantal transmission at central excitatory synapses, *Advances in Second Messenger and Phosphoprotein Research* 29, 261–273.

Bell, A. J., and T. J. Sejnowski (1995). An information maximization approach to blind separation and blind deconvolution, *Neural Comp.* 6, 1129–1159.

Berg, H. C., and E. M. Purcell (1977). Physics of chemoreception, *Biophys. J.* 20, 193–219.

Bevensee, R. M. (1993). *Maximum Entropy Solutions to Scientific Problems* (Prentice Hall, Englewood Cliffs).

Bialek, W. (1987). Physical limits to sensation and perception, *Ann. Rev. Biophys. Biophys. Chem.* 16, 455–478.

Bialek, W. (1990). Theoretical physics meets experimental neurobiology. In *1989 Lectures in Complex Systems, SFI Studies in the Sciences of Complexity, Lecture Vol. II*, E. Jen, ed., pp. 513–595 (Addison–Wesley, Menlo Park, CA).

Bialek, W. (1992). Optimal signal processing in the nervous system. In *Princeton Lectures on Biophysics*, W. Bialek, ed., pp. 321–401 (World Scientific, Singapore).

Bialek, W., and M. DeWeese (1995). Random switching and optimal processing in the perception of ambiguous signals, *Phys. Rev. Lett.* 74, 3077–3080.

Bialek, W., M. DeWeese, F. Rieke, and D. Warland (1993). Bits and brains: Information flow in the nervous system, *Physica A* 200, 581–593.

Bialek, W., and W. G. Owen (1990). Temporal filtering in retinal bipolar cells: Elements of an optimal computation?, *Biophys. J.* 58, 1227–1233.

Bialek, W., F. Rieke, R. R. de Ruyter van Steveninck, and D. Warland (1990). Reading a neural code. In *Advances in Neural Information Processing Systems 2*, D. Touretzky, ed., pp. 36–43 (Morgan Kaufmann, San Mateo CA).

Bialek, W., F. Rieke, R. R. de Ruyter van Steveninck, and D. Warland (1991). Reading a neural code, *Science* 252, 1854–1857.

Bialek, W., D. L. Ruderman, and A. Zee (1991). Optimal sampling of natural images: A design principle for the visual system?, in *Advances in Neural Information Processing Systems 3*, R. P. Lippman, J. E. Moody, and D. S. Touretzky, eds., pp. 363–369 (Morgan Kaufmann, San Mateo CA).

Bialek, W., and A. Zee (1990). Coding and computation with neural spike trains, *J. Stat. Phys.* 59, 103–115.

Blake, A., H. H. Bülthoff, and D. Sheinberg (1993). Shape from texture—Ideal observers and human psychophysics, *Vision Res.* 33, 1723–1737.

Blum, K. I., and L. Abbott (1996). A model of spatial map formation in the hippocampus of the rat, *Neural Comp.*, 8, 85-93.

de Boer, E. (1976). On the residue and auditory pitch perception. In *Handbook of Sensory Physiology V/3: Auditory System. Clinical and Special Topics*, ed. W. D. Keidel and W. D. Neff, pp. 479–583 (Springer-Verlag, Berlin).

de Boer, E., and P. Kuyper (1968). Triggered correlation, *I. E. E. E. Trans. Biomed. Eng.* 15, 169–179.

Boring, E. G. (1942). *Sensation and Perception in the History of Experimental Psychology* (Appelton-Century, New York).

Born, M. (1949). *Natural Philosophy of Cause and Chance; being the Waynflete Lectures delivered in the College of St. Mary Magdalen, Oxford, in Hilary term, 1948* (Oxford University Press, Oxford).

Borst, A., and M. Egelhaaf (1989). Principles of visual motion detection, *Trends. Neurosci.* 12, 297–306.

Bouman, M. A. (1961). History and present status of quantum theory in vision. In *Sensory Communication*, W. Rosenblith, ed., pp. 377–401 (MIT Press, Cambridge MA).

Brillouin, L. (1962). *Science and Information Theory* (Academic Press, New York).

Britten, K. H. et al. (1996). [K. H. Britten, W. T. Newsome, M. N. Shadlen, S. Celebrini, and J. A. Movshon] A relationship between behavioral choice and the visual responses of neurons in macaque MT, *Vis. Neurosci.* 13, 87–100.

Britten, K. H. et al. (1992). [K. H. Britten, M. N. Shadlen, W. T. Newsome, and J. A. Movshon] The analysis of visual motion: A comparison of neuronal and psychophysical performance, *J. Neurosci.* 12, 4745–4765.

Buchner, E. (1984). Behavioural analysis of spatial vision in insects. In *Photoreception and Vision in Invertebrates*, M. Ali, ed., pp. 561–622 (Plenum, New York).

Buck, B., and V. A. Macaulay, eds. (1991). *Maximum Entropy in Action: A Collection of Expository Essays* (Oxford University Press, Oxford).



Bullock, T. H. (1970). The reliability of neurons, *J. Gen. Physiol.* 55, 584–656.

Bullock, T. H. (1976). Redundancy and noise in the nervous system: Does the model based on unreliable neurons sell nature sort?. In *Electrobiology of Nerve, Synapse, and Muscle*, ed. J. Reuben, D. Purpura, M. V. L. Bennett, and E. Kandel, pp. 179–185 (Raven Press, New York).

Cajal, S. Ramón y (1909–11). *Histologie du système nerveux de l'homme et des vertébrés*. Edition francaise traduite de l'espagnol par L. Azoulay. (A. Maloine, Paris). Translated by N. Swanson and L. W. Swanson, *Histology of the Nervous System of Man and Vertebrates* (Oxford University Press, New York, 1995).

Capranica, R. R. (1965). *The Evoked Vocal Response of the Bullfrog* (MIT Press, Cambridge, MA).

Capranica, R. R. (1968). The vocal repertoire of the bullfrog (*Rana catesbiana*), *Behavior* 31, 302–325.

Carlin, G. (1978). Seven words you can never say on television. In *Indecent Exposure: Some of the Best of George Carlin*, Little David Records (Atlantic Recording, New York).

Carr, C. E. (1993). Processing of temporal information in the brain, *Ann. Rev. Neurosci.* 16, 223–243.

Carr, C. E., W. Heiligenberg and G. J. Rose (1986). A time-comparison circuit in the electric fish midbrain. I. Behavior and physiology, *J. Neurosci.* 6, 107–119.

Carr, C. E., and M. Konishi (1990). A circuit for detection of interaural time differences in the brain stem of the barn owl, *J. Neurosci.* 10, 3227–3246.

Castaing, B., G. Gunaratne, F. Heslot, L. Kadanoff, A. Libchaber, S. Thomae, X.–Z. Wu, S. Zaleski, and G. Zanetti (1989). Scaling of hard thermal turbulence in Rayleigh–Bernard convection, *J. Fluid Mech.* 209, 1–30.

Chittka, L., and R. Menzel (1992). The evolutionary adaptation of flower colours and the insect pollinators' colour vision, *J. Comp. Physiol. A* 171, 171–181.

Corbière-Tichané, G., and R. Loftus (1983). Antennal thermal receptors of the cave beetle, *Speophyes lucidulus* Delar.: II. Cold receptor response to slowly changing temperature, *J. Comp. Physiol.* 153, 343–351.

DeAngelis, G. C., I. Ohzawa, and R. D. Freeman (1995). Receptive field dynamics in the central visual pathways, *Trends Neurosci.* 18, 451–458.

Dear, S. P., J. Fritz, T. Haresign, M. Ferragamo, and J. A. Simmons (1993). Tonotopic and functional organization in the auditory cortex of the big brown bat, *Eptesicus fuscus*, *J. Neurophys.* 70, 1988–2009.

Dear, S. P., J. A. Simmons, J. Fritz (1993). A possible neuronal basis for representation of acoustic scenes in the auditory cortex of the big brown bat, *Nature* 364, 620–623.

DeFelice, L. J. (1981). *Introduction to Membrane Noise* (Plenum Press, New York).

DeVries, S. H., and D. A. Baylor (1996). Mosaic design of ganglion cell receptive fields in rabbit retina, preprint.

DeWeese, M. (1995). Optimization Principles for the Neural Code. Dissertation, Princeton University.

Dickinson, M. H. (1994). The effects of wing rotation on unsteady aerodynamic performance at low Reynolds numbers, *J. Exp. Biol.* 192, 179–206.

Dickinson, M. H., and K. G. Götz (1993). Unsteady aerodynamic performance of model wings at low Reynolds numbers, *J. Exp. Biol.* 174, 45–64.

Dill, M., R. Wolf, and M. Heisenberg (1993). Visual pattern recognition in Drosophila involves retinotopic matching, *Nature* 365, 751–753.

Dill, M., and M. Heisenberg (1995). Visual pattern memory without shape recognition, *Phil. Trans. R. Soc. Lond. Ser. B* 349, 143–152.

Donchin, O., and W. Bialek (1995). Notes on the reliability of coding in motor cortex, unpublished.

Donner, K. (1989). The absolute sensitivity of vision: Can a frog become a perfect detector of light induced and dark rod events?, *Phys. Scr.* 39, 133–140.

Earman, J. (1992). *Bayes or Bust?: A Critical Examination of Bayesian Confirmation Theory* (MIT Press, Cambridge MA).

Ebeling, W., and T. Pöschel (1994). Entropy and long-range correlations in literary english, *Europhys. Lett.* 26, 241–246.

Eckhorn, R., and B. Pöpel (1974). Rigorous and extended application of information theory to the afferent visual system of the cat I: Basic concepts, *Kybernetik* 16, 191–200.

Eckhorn, R., and B. Pöpel (1975). Rigorous and extended application of information theory to the afferent visual system of the cat II: Experimental results, *Biol. Cybern.* 17, 7–17.

Eggermont, J. J., P. I. M. Johannesma, and A. M. H. J. Aertsen (1983). Reverse-correlation methods in auditory research, *Q. Rev. Biophys.* 16, 341–414.

Eskandar, E. N., B. J. Richmond, and L. M. Optican (1992). Role of inferior temporal neurons in visual memory: I. Temporal encoding of information about visual images, recalled images, and behavioral context, *J. Neurophys.* 68, 1277–1295.

Evans, E. F. (1982). Functional anatomy of the auditory system. In *The Senses*, H. B. Barlow and J. D. Mollon, eds., pp. 251–306 (Cambridge University Press, Cambridge).

Fee, M., and D. Kleinfeld (1994). Neuronal responses in rat vibrissa cortex during behavior, *Soc. Neurosci. Abs.* 20, 26.

Feynman, R. P., and A. R. Hibbs (1965). *Path Integrals and Quantum Mechanics.* (McGraw Hill, New York).

Feynman, R. P., R. Leighton, and M. Sands (1963). *The Feynman Lectures on Physics* (Addison–Wesley, Reading MA).

Field, D. (1987). Relations between the statistics of natural images and the response properties of cortical cells, *J. Opt. Soc. Am. A* 4, 2379–2394.

FitzHugh, R. (1958). A statistical analyzer for optic nerve messages, *J. Gen. Physiol.* 41, 675–692.



Franceschini, N., A. Riehle, and A. le Nestour (1989). Directionally selective motion detection by insect neurons, in *Facets of Vision*, R. C. Hardie and D. G. Stavenga, eds., pp. 360–390 (Springer-Verlag, Berlin).

Frost, D. R., ed. (1985) *Amphibian Species of the World. A Taxonomic and Geographic Reference* (Allen Press, Lawrence KS).

Fuortes, M. G. F., and S. Yeandle (1964). Probability of occurrence of discrete potential waves in the eye of *Limulus*. *J. Gen. Physiol.* 47, 443–463.

Gabbiani, F., and C. Koch (1996). Coding of time-varying signals in spike trains of integrate-and-fire neurons with random threshold, *Neural Comp.* 8, 44–66.

Gallant, J., C. E. Connor, and D. C. van Essen (1994). Responses of visual cortical neurons in a monkey freely viewing natural scenes, *Soc. Neurosci. Abs.* 20, 838.

Geisler, W. (1984). Physical limits of acuity and hyperacuity, *J. Opt. Soc. Am. A* 1, 775–782.

Geisler, W. S., and K. D. Davila (1985). Ideal discriminators in spatial vision: Two-point stimuli, *J. Opt. Soc. Am. A* 2, 1483–1497.

Georgopoulos, A. P., J. T. Lurito, M. Petrides, A. Schwartz and J. T. Massey (1989). Mental rotation of the population vector, *Science* 243, 234–236.

Georgopoulos, A. P., A. Schwartz and R. E. Kettner (1986). Neuronal population coding of movement direction, *Science* 233, 1416–1419.

Georgopoulos, A. P., M. Taira and A. Lukashin (1993). Cognitive neurophysiology of the motor cortex, *Science* 260, 47–52.

von Gersdorff, H., and G. Matthews (1994). Dynamics of synaptic vesicle fusion and membrane retrieval in synaptic terminals, *Nature* 367, 735–739.

Gielen, C. C. A. M., G. H. F. M. Hesselmans, and P. I. M. Johannesma (1988). Sensory interpretation of neural activity patterns, *Math. Biosci.* 88, 15–35.

Goldberg, J. M., and C. Fernandez (1971). Physiology of peripheral neurons innervating semicircular canals of the squirrel monkey. III: Variations among units in their discharge properties, *J. Neurophys.* 34, 676–684.

Goldstein, J. L. (1967). Auditory non–linearity, *J. Acoust. Soc.Am.* 41, 676–689.

Goldstein, J. L. (1973). An optimum processor theory for the central formation of the pitch of complex tones, *J. Acoust. Soc. Am.* 54, 1496–1516.

Goldstein, J. L., and P. Srulovicz (1977). Auditory–nerve spike intervals as an adequate basis for aural spectrum analysis. In *Psychophysics and Physiology of Hearing*, E. F. Evans and J. P. Wilson, eds., pp. 337–346.

Goldstein, J. L., A. Gerson, P. Srulovicz, and M. Furst (1978). Verification of the optimal probabilistic basis for aural processing in pitch of complex tones, *J. Acoust. Soc. Am.* 63, 486–497.

Gollub, J. P., J. Clarke, M. Gharib, B. Lane, and O. N. Mesquita (1991). Fluctuations and transport with a mean gradient, *Phys. Rev. Lett.* 67, 3507–3510.

Golomb, D., D. Kleinfeld, R. C. Reid, R. M. Shapley and B. I. Shraiman (1994). On temporal codes and the spatiotemporal response of neurons in the lateral geniculate nucleus, *J. Neurophys.* 72, 2990–3003.

Gozani, S. N., and J. P. Miller (1994). Optimal discrimination and classification of neuronal action potential waveforms from multiunit, multichannel recordings using software–based linear filters, *I. E. E. E. Trans. Biomed. Eng.* 41, 358–372.

Green, D. M., and J. A. Swets (1966). *Signal Detection Theory and Psychophysics.* (Wiley, New York).

Griffin, D. R. (1958). *Listening in the Dark: The Acoustic Orientation of Bats and Men* (Yale University Press, New Haven). Dover edition, 1974 (Dover, New York).

Gross, C. G., and J. Sergent (1992). Face recognition, *Curr. Opin. Neurobiol.* 2, 156–161.

Grzywacz, N. M., F. R. Amthor, and D. K. Merwine (1994). Directional hyperacuity in ganglion cells of the rabbit retina, *Vis. Neurosci.* 11, 1019–1025.

Gull, S. F., and G. J. Daniell (1978). Image reconstruction from incomplete and noisy data, *Nature* 272, 686–690.

Gutnick, M. J., and I. Mody, eds. (1995) *The Cortical Neuron* (Oxford University Press, Oxford).

Hassenstein, S., and W. Reichardt (1956). Systemtheoretische Analyse der Zeit-, Reihenfolgen-, und Vorzeichenauswertung bei der Bewegungsperzeption des Rüsselkäfers *Chlorophanus, Z. Naturforsch.* 11b, 513–524.

van Hateren, J. H. (1992). Real and optimal neural images in early vision, *Nature* 360, 68–70.

Hausen, K. (1984). The lobular complex of the fly: Structure, function, and significance in behavior. In *Photoreception and vision in invertebrates,* M. Ali, ed., pp. 523–559 (Plenum, New York).

Hausen, K. and M. Egelhaaf (1989). Neural mechanisms of visual course control in insects. In *Facets of Vision,* D. G. Stavenga and R. C. Hardie, eds., pp. 391–424 (Springer-Verlag, Berlin).

Hausen, K. and C. Wehrhahn (1983). Microsurgical lesion of horizontal cells changes optomotor yaw responses in the blowfly *Calliphora erythrocephala, Proc. R. Soc. Lond. B* 21, 211–216.

Hecht, S., S. Shlaer, and M. H. Pirenne (1942). Energy, quanta, and vision, *J. Gen. Physiol.* 25, 819–840.

Heisenberg, M., and R. Wolf (1984). *Vision in Drosophila: Genetics of Microbehavior* (Springer-Verlag, Berlin).

von Helmholtz, H. L. F. (1885). *On the Sensation of Tone as a Physiological Basis for the Theory of Music,* translated from the last German edition (1877) by A. J. Ellis (Longmans, London). Reprint, with an introduction by H. Margenau (Dover, New York, 1954).

Hille, B. (1992). *Ionic Channels of Excitable Membranes,* 2d ed. (Sinauer Associates, Sunderland MA).

Himstedt, W., and U. Grüsser–Cornehls (1976). The urodele visual system. In *The Amphibian Visual System,* K. V. Fite and W. F. Blair, eds., pp. 203–266 (Academic Press, New York).



Hodgkin, A. L., and A. F. Huxley (1952a). Currents carried by sodium and potassium ions through the membrane of the giant axon of *Loligo, J. Physiol.* 116, 449–472.

Hodgkin, A. L., and A. F. Huxley (1952b). The components of membrane conductance in the giant axon of *Loligo, J. Physiol.* 116, 473–496.

Hodgkin, A. L., and A. F. Huxley (1952c). The dual effect of membrane potential on sodium conductance in the giant axon of *Loligo, J. Physiol.* 116, 497–506.

Hodgkin, A. L., and A. F. Huxley (1952d). A quantitative description of membrane current and its application to conduction and excitation in nerve, *J. Physiol.* 117, 500–544.

Hodgkin, A. L., A. F. Huxley, W. Feldberg, W. A. H. Rushton, R. A. Gregory, and R. A. McCance (1977). *The Pursuit of Nature: Informal Essays on the History of Physiology* (Cambridge University Press, Cambridge).

Hodgkin, A. L., and W. A. H. Rushton (1946). The electrical constants of a crustacean nerve fibre, *Proc. R. Soc Lond. Ser. B* 133, 444–479.

Hopfield, J. J. (1995). Pattern recognition computation using action potential timing for stimulus representation, *Nature* 376, 33–36.

Horowitz, P., and W. Hill (1980). *The Art of Electronics* (Cambridge University Press, Cambridge).

Hubel, D. H., and T. N. Wiesel (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex, *J. Physiol. (Lond.)*160, 106–154.

Hubel, D. H., and T. N. Wiesel (1977). Functional architecture of macaque monkey visual cortex, *Proc. R. Soc. Lond. Ser. B* 198, 1–59.

Huber, F., T. E. Moore, and W. Loher, eds. (1989). *Cricket Behavior and Neurobiology* (Cornstock, Ithaca).

Humphrey, J. A. C., R. Devarkonda, I. Ingelssia, and F. G. Barth (1993). Dynamics of arthropod filiform hairs. I: Mathematical modelling of the hair and air motion, *Phil. Trans. R. Soc. Ser. B* 340, 423–444.

Jacobs, G. A., and R. Nevin (1991). Anatomical relationships between sensory afferent arborizations in the cricket cercal system, *Anat. Rec.* 231, 563–572.

Jan, L. Y., and Y. N. Jan (1994). Potassium channels and their evolving gates, *Nature* 371, 119–122.

Jaramillo, F., V. S. Markins and A. J. Hudspeth (1993). Auditory illusions and the single hair cell, *Nature* 364, 527–529.

Jaynes, E. T. (1983). *E. T. Jaynes: Papers on Probability, Statistics, and Statistical Physics.* Edited by R. D. Rosenkrantz (Kluwer Academic, Boston).

Johannesma, P. I. M. (1981). Neural representation of sensory stimuli and sensory interpretation of neural activity, *Adv. Physiol. Sci.* 30, 103–126.

Johnson, D. H. (1974). The response of single auditory nerve fibers in the cat to single tones: Synchrony and average discharge rate. Dissertation, Massachusetts Institute of Technology.

Kadanoff, L. P. (1966). Scaling laws for Ising models near $T_c$, *Physics,* 2, 263–272.

Katz, B. (1966). *Nerve, Muscle, and Synapse.* (McGraw–Hill, New York)

Kiang, N. Y.–S., T. Watanabe, E. C. Thomas, and L. F. Clark (1965) *Discharge Patterns of Single Fibers in the Cat's Auditory Nerve* (MIT Press, Cambridge MA).

Kittel, C., and H. Kroemer (1980). *Thermal Physics, 2d ed.* (Freeman, San Francisco).

Kjaer, T. W., J. A. Hertz, and B. J. Richmond (1994). Decoding cortical neuronal signals: Network models, information estimation, and spatial tuning, *J. Comp. Neurosci.* 1, 109–139.

Klein, S. A., and D. M. Levi (1985). Hyperacuity thresholds of 1 sec: Theoretical predictions and empirical validation, *J. Opt. Soc. Am. A* 2, 1170–1190.

Knierem, J. J., and D. C. van Essen (1992). Neuronal responses to static textures in area V1 of the alert Macaque monkey, *J. Neurophys.* 67, 961–980.

Knudsen, E. I., S. du Lac, and S. D. Esterly (1987). Computational maps in the brain, *Ann. Rev. Neurosci.* 10, 41–65.

Kolmogoroff, A. (1939). Sur l'interpolation et extrapolations des suites stationnaires, *C. R. Acad. Sci. Paris* 208, 2043–2045.

Kolmogorov, A. N. (1941). Interpolation and extrapolation of stationary random sequences (in Russian), *Izv. Akad. Nauk. SSSR Ser. Mat.*5, 3–14. English translation in *Selected Works of A. N. Kolmogorov, Volume II.* Edited by A. N. Shiryagev, pp. 272–280 (Kluwer Academic Publishers, Dordrecht, The Netherlands).

Kroese, A. B., J. M. van der Zalm, and J. van den Bercken (1978). Frequency response of the lateral-line organ of *Xenopus laevis, Pflugers Arch.* 375, 167–175.

Kuffler, S. W. (1953). Discharge patterns and functional organization of mammalian retina, *J. Neurophys.* 16, 37–68.

du Lac, S., and S. G. Lisberger (1995). Cellular processing of temporal information in medial vestibular nucleus neurons, *J. Neurosci.* 15, 8000–8010.

Land, M. F., and T. S. Collett (1974). Chasing behavior of houseflies (*Fannia canicularis*): A description and analysis, *J. Comp. Physiol.* 89, 331–357.

Landau, L. D., and E. M. Lifshitz (1969). *Statistical Physics,* Second revised and enlarged edition translated from the Russian by J. B. Sykes and M. J. Kearsley (Pergamon Press, Oxford).

Lass, Y., and M. Abeles (1975). Transmission of information by the axon. I: Noise and memory in the myelinated nerve fiber of the frog, *Biol. Cybern.* 19, 61–67.

Laughlin, S. B. (1981). A simple coding procedure enhances a neuron's information capacity, *Z. Naturforsch.* 36c, 910–912.

Lawson, J. L., and G. E. Uhlenbeck (1950). *Threshold Signals.* Massachusetts Institute of Technology Radiation Laboratory Series, vol. 24. (McGraw–Hill, New York).

Lee, B. B., C. Wehrhahn, G. Westheimer, and J. Kremer (1993). Macaque ganglion cell responses to stimuli which elicit hyperacuity in man: Detection of small displacements, *J. Neurosci.* 13, 1001–1009.

Lettvin, J. Y., H. R. Maturana, W. S. McCulloch, and W. H. Pitts (1959). What the frog's eye tells the frog's brain, *Proc. I. R. E.* 47, 1940–1951.



Lewis, E. R., E. L. Leverenz, and W. S. Bialek (1985). *The Vertebrate Inner Ear* (CRC Press, Boca Raton FL).

Lighthill, J. (1958). *An Introduction to Fourier Analysis and Generalized Functions* (Cambridge University Press, Cambridge).

Linsker, R. (1990). Perceptual neural organization: Some approaches based on network models and information theory, *Ann. Rev. Neurosci.* 13, 257–281.

Liu, Z. L., D. C. Knill, and D. Kersten (1995). Object classification for human and ideal observers, *Vision Res.* 35, 549–568.

Loftus, R., and G. Corbière-Tichané (1981). Antennal warm and cold receptors of the cave beetle, *Speophyes lucidulus* Delar., in sensilla with a lamellated dendrite: I. Response to sudden temperature change, *J. Comp. Physiol.* 143, 443–452.

Loftus, R., and G. Corbière-Tichané (1987). Response of antennal cold receptors of the catopid beetles, *Speophyes lucidulus* Delar. and *Choleva augustata* Fab. to very slowly changing temperature, *J. Comp. Physiol.* 161, 399–405.

Lucas, K. (1917). *The Conduction of the Nervous Impulse* (Longmans, London).

Ma, S.–K. (1976). *Modern Theory of Critical Phenomena* (W. A. Benjamin, Reading MA).

MacKay, D., and W. S. McCulloch (1952). The limiting information capacity of a neuronal link, *Bull. Math. Biophys.* 14, 127–135.

Maddess, T., and S. B. Laughlin (1985). Adaptation of the movement sensitive neuron H1 is generated locally and governed by contrast frequency, *Proc. R. Soc. Lond. Ser. B* 225, 251–275.

Mafra–Neto, A., and R. T. Cardé (1994). Fine-scale structure of pheromone plumes modulates upwind orientation of flying moths, *Nature* 369, 142–144.

Mainen, Z. F., and T. J. Sejnowski (1995). Reliability of spike timing in neocortical neurons, *Science* 268, 1503–1506.

Mallock, A. (1894). Insect sight and the defining power of compound eyes, *Proc. R. Soc. Lond. Ser. B* 55, 85–90.

Marmarelis, P. Z., and V. Z. Marmarelis (1978). *Analysis of physiological systems: The white-noise approach* (Plenum Press, New York).

Mathews, J., and R. L. Walker (1964). *Mathematical Methods of Physics* (W. A. Benjamin, New York).

McClurkin, J. W., et al. (1991a). [J. W. McClurkin, T. J. Gawne, L. M. Optican, and B. J. Richmond] Lateral geniculate neurons in behaving primates. II: Encoding of visual information in the temporal shape of the response, *J. Neurophys.* 66, 794–808.

McClurkin, J. W., et al. (1991b). [J. W. McClurkin, T. J. Gawne, B. J. Richmond, L. M. Optican, and D. L. Robinson] Lateral geniculate neurons in behaving primates. I: Responses to two dimensional stimuli, *J. Neurophys.* 66, 777–793.

McClurkin, J. W., et al. (1991c). [J. W. McClurkin, L. M. Optican, B. J. Richmond, and T. J. Gawne] Concurrent processing and complexity of temporally encoded messages in visual perception, *Science* 253, 675–677.

McKee, S. P. (1991). The physical constraints on visual hyperacuity. In *Limits of Vision: Vision and Visual Dysfunction 5*, J. J. Kulikowski, V. Walsh, and I. J. Murray, eds., pp. 221–233 (CRC Press, Boca Raton FL).

Meinertzhagen, I. (1993). The synaptic populations of the fly's optic neuropil and their dynamic regulation—parallels with the vertebrate retina, *Prog. Retinal Res.* 12, 13–39.

Meister, M., L. Lagnado, and D. A. Baylor (1995). Concerted signaling by retinal ganglion cells, *Science* 270, 1207–1210.

Meister, M., J. Pine, and D. A. Baylor (1994). Multi–neuronal signals from the retina: Acquisition and analysis, *J. Neurosci. Meth.* 51, 95–106.

Menne, D., and H. Hackbarth (1986). Accuracy of distance measurement in the bat *Eptesicus fuscus*: Theoretical aspects and computer simulations, *J. Acoust. Soc. Am.* 79, 386–397.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information, *Psychol. Rev.* 63, 81–97.

Miller, J. P., G. A. Jacobs, F. E. Theunissen (1991). Representation of sensory information in the cricket cercal sensory system. I: Response properties of the primary interneurons, *J. Neurophys.* 66, 1680–1703.

Miller, M. I., and K. E. Mark (1992). A statistical study of cochlear nerve discharge patterns in response to complex speech stimuli, *J. Acoust. Soc. Am.* 92, 202–209.

Miller, M. I., and M. B. Sachs (1983). Representation of stop consonants in the discharge patterns of auditory–nerve fibers, *J. Acoust. Soc. Am.* 74, 502–517.

Miller, M. I., and M. B. Sachs (1984). Representation of voice pitch in discharge patterns of auditory–nerve fibers, *Hearing Res.* 14, 257–279.

Mollon, J. D., Estévez, O., and Cavonius, C. R. (1990). The two subsystems of colour vision and their rôles in wavelength discrimination. In *Vision: Coding and Efficiency*, C. Blakemore, ed., pp. 119–131 (Cambridge University Press, Cambridge)

Mountcastle, V. B. (1957). Modality and topographic properties of single neurons of cat's somatic sensory cortex, *J. Neurophys.* 20, 408–434.

Murlis, J., J. S. Elkinton, and R. T. Cardé (1992). Odor plumes and how insects use them, *Ann. Rev. Entomol.* 37, 505–532.

Narins, P. M., and E. R. Lewis (1984). The vertebrate inner ear as an exquisite seismic sensor, *J. Acoust. Soc. Am.* 76, 1384–1387.

Nelkin, I. (1995). On the structure of natural sounds, unpublished.

van Ness, F. L., and M. A. Bouman (1967). Spatial modulation transfer in the human eye, *J. Opt. Soc. Am.* 57, 401–406.

von Neumann, J. (1956). Probabilistic logics and the synthesis of reliable organisms from unreliable components, In *Automata Studies*, C. E. Shannon and J. McCarthy eds., pp. 43–98 (Princeton University Press, Princeton).

von Neumann, J. (1958). *The Computer and the Brain.* (Yale University Press, New Haven CT).



Newsome, W. T., K. H. Britten, C. D. Salzman, and J. A. Movshon (1990). Neuronal mechanisms of motion perception, *Cold Spring Harbor Symp. Quant. Biol.* 55, 697–705.

Newsome, W. T., and E. B. Paré (1988). A selective impairment of motion perception following lesions of the middle temporal visual area (MT), *J. Neurosci.* 8, 2201–2211.

Newsome, W. T., M. N. Shadlen, E. Zohary, K. H. Britten, and J. A. Movshon (1995). Visual motion: Linking neuronal activity to psychophysical performance. In *The Cognitive Neurosciences*, M. Gazzaniga, ed., pp. 401–414 (MIT Press, Cambridge MA).

Nunn, B. J., J. L. Schnapf, and D. A. Baylor (1984). Spectral sensitivity of single cones in the retina of *Macaca fascicularis*, *Nature* 309, 264–267.

O'Carroll, D. (1993). Feature–detecting neurones in dragonflies, *Nature* 362, 541–543.

O'Keefe, J., and L. Nadel (1978). *The Hippocampus as a Cognitive Map* (Oxford University Press, New York).

O'Keefe, J., and M. Recce (1993). Phase relationship between hippocampal place units and the EEG theta rhythm, *Hippocampus* 3, 317–330.

Optican, L. M., and B. J. Richmond (1987). Temporal encoding of two dimensional patterns by single units in primate inferior temporal cortex. III: Information theoretic analysis, *J. Neurophys.* 57, 162–178.

Panzeri, S., G. Biella, E. T. Rolls, W. E. Skaggs, and A. Treves (1996). Speed, noise, information and the graded nature of neuronal responses, preprint.

Papoulis, A. (1965). *Probability, random variables and stochastic processes* (McGraw–Hill, New York).

Parker, A. J., and M. J. Hawken (1985). Capabilities of monkey cortical cells in spatial–resolution tasks, *J. Opt. Soc. Am. A* 2, 1101–1114.

Pera, M. (1986). *La rana ambigua* (Giuilio Einaudi editore, Torino). Translated by J. Mandelbaum, *The Ambiguous Frog: The Galvani–Volta Controversy on Animal Electricity* (Princeton University Press, Princeton, 1992).

Perkel, D. H., and T. H. Bullock (1968). Neural coding, *Neurosci. Res. Prog. Sum.* 3, 405–527.

Pippard, A. B. (1985). *Response and Stability: An Introduction to the Physical Theory* (Cambridge Unviersity Press, Cambridge).

Poggio, T., and W. Reichardt (1976). Visual control of orientation behavior in the fly. Part II. Towards the underlying neural interactions, *Q. Rev. Biophys.* 9, 377–438.

Potters, M., and W. Bialek (1994). Statistical mechanics and visual signal processing, *J. Phys. I France* 4, 1755–1775.

Press, W. H., B. P. Flanney, S. A. Teukolsky, and W. T. Vetterling (1992). *Numerical Recipes in C: The Art of Scientific Computing, 2d ed.* (Cambridge University Press, Cambridge).

Pumphrey, R. J. (1940). Hearing in insects, *Biol. Reviews* 15, 107–132.

Ratliff, F., ed. (1974). *Studies on Excitation and Inhibiton in the Retina* (Rockefeller University Press, New York).

Reichardt, W., and T. Poggio (1976). Visual control of orientation behavior in the fly. Part I: A quantitative analysis. *Q. Rev. Biophys.* 9, 311–375.

Reid, R. C., and R. M. Shapley (1992). Spatial structure of cone inputs to receptive fields in primate lateral geniculate nucleus, *Nature* 356, 716–718.

Reid, R. C., R. E. Soodak, and R. M. Shapley (1991). Directional selectivity and spatiotemporal structure of receptive fields of simple cells in cat striate cortex, *J. Neurophys.* 66, 505–529.

Rice, S. O. (1944–45). Mathematical analysis of random noise, *Bell Sys. Tech. J.* 23, 1–51; 24, 52–162. (reprinted in Wax 1954).

Richmond, B. J., and L. M. Optican (1987). Temporal encoding of two dimensional patterns by single units in primate inferior temporal cortex. II: Quantification of response waveform, *J. Neurophys.* 57, 147–161.

Richmond, B. J., and L. M. Optican (1990). Temporal encoding of two dimensional patterns by single units in primate primary visual cortex. II: Information transmission, *J. Neurophys.* 64, 370–380.

Richmond, B. J., L. M. Optican, and H. Spitzer (1990). Temporal encoding of two dimensional patterns by single units in primate primary visual cortex. I: Stimulus-response relations, *J. Neurophys.* 64, 351–369.

Richmond, B. J., L. M. Optican, M. Podell, and H. Spitzer (1987). Temporal encoding of two dimensional patterns by single units in primate inferior temporal cortex. I: Response characteristics, *J. Neurophys.* 57, 132–146.

Rieke, F. (1991). Physical Principles Underlying Sensory Processing and Computation. Dissertation, University of California at Berkeley.

Rieke, F., D. Bodnar, and W. Bialek (1992). Coding of natural sound stimuli by the bullfrog auditory nerve: Phase, amplitude and information rates. In *Proceedings of the Third International Congress of Neuroethology*, abstract 153.

Rieke, F., D. Bodnar, and W. Bialek (1995). Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory neurons, *Proc. R. Soc. Lond. Ser. B*, 262, 259–265.

Rieke, F., W. G. Owen, and W. Bialek (1991). Optimal filtering in the salamander retina, in *Advances in Neural Information Processing Systems 3*, J. E. Moody, S. J. Hanson, and R. P. Lippmann, eds., pp. 377–383 (Morgan Kaufmann, San Mateo CA).

Rieke, F., D. Warland, and W. Bialek (1993). Coding efficiency and information rates in sensory neurons, *Europhys. Lett.*, 22, 151–156.

Rieke, F., et al. (1996). [F. Rieke, D. Warland, R. R. de Ruyter van Steveninck, and W. Bialek] Optimal processing of visual movement signals: Theory and experiments in the blowfly, unpublished.

Rieke, F., et al. (1992). [F. Rieke, W. Yamada, K. Moortgat, E. R. Lewis, and W. Bialek] Real-time coding of complex signals in the auditory nerve, in *Auditory Physiology and Perception: Proceedings of the 9th International Symposium on Hearing*, Y. Cazals, L. Demany, and K. Horner, eds., pp. 315–322 (Elsevier, Amsterdam).



Roberts, A., and B. M. H. Bush, eds. (1981). *Neurones Without Impulses: Their significance for vertebrate and invertebrate nervous systems* (Cambridge University Press, Cambridge).

Roeder, K. D. (1963). *Nerve Cells and Insect Behavior* (Harvard University Press, Cambridge MA).

Roeder, K. D., and R. S. Payne (1966). Acoustic orientation of a moth in flight by means of two sense cells, *Symp. Soc. Exp. Biol.* 20, 251–272.

Roeder, K. D., and A. E. Treat (1957). Ultrasonic reception by the tympanic organ of noctuid moths, *J. Exp. Zool.* 134, 127–157.

Roeder, K. D., and A. E. Treat (1961). The detection and evasion of bats by moths, *Am. Scientist* 49, 135–148.

Rolls, E. T., and M. J. Tovee (1994). Processing speed in the cerebral cortex and the neurophysiology of visual masking, *Proc. R. Soc. Lond. Ser. B* 257, 9–15.

Rose, A. (1948). The sensitivity performance of the human eye on an absolute scale, *J. Opt. Soc. Am.* 38, 196–208.

Rose, J. E., J. F. Brugge, D. J. Anderson, and J. E. Hind (1967). Phase-locked response to low-frequency tones in single auditory nerve fibers of the squirrel monkey, *J. Neurophys.* 30, 769–793.

Rose, G., and W. Heiligenberg (1985). Temporal hyperacuity in the electric sense of fish, *Nature* 318, 178–180.

Ruderman, D. L. (1993). Natural Ensembles and Sensory Signal Processing. Dissertation, University of California at Berkeley.

Ruderman, D. L., and W. Bialek (1994). Statistics of natural images: Scaling in the woods, *Phys. Rev. Lett.* 73, 814–817.

de Ruyter van Steveninck, R. R. (1986). Real-time Performance of a Movement-Sensitive Neuron in the Blowfly Visual System. Academisch Proefschrift, Rijksuniversiteit Groningen.

de Ruyter van Steveninck, R., and W. Bialek (1988). Real-time performance of a movement sensitive neuron in the blowfly visual system: Coding and information transfer in short spike sequences, *Proc. R. Soc. Lond. Ser. B* 234, 379–414.

de Ruyter van Steveninck, R. R., and W. Bialek (1992). Statistical reliability of a blowfly movement-sensitive neuron, in *Advances in Neural Information Processing Systems 4*, R. Lippmann, J. Moody, and D. Touretzky, eds., pp. 27–34 (Morgan Kaufmann, San Mateo CA).

de Ruyter van Steveninck, R. R., and W. Bialek (1995). Reliability and statistical efficiency of a blowfly movement-sensitive neuron, *Phil. Trans. R. Soc. Lond. Ser. B* 348, 321–340.

de Ruyter van Steveninck, R. R., W. Bialek, and W. H. Zaagman (1984). Vernier movement discrimination with three spikes from one neuron, *Perception* 13, A47–48.

de Ruyter van Steveninck, R. R., W. Bialek, M. Potters, and R. H. Carlson (1994). Statistical adaptation and optimal estimation in movement computation by the blowfly

visual system, *Proceedings of the 1994 I. E. E. E. Conference on Systems, Man and Cybernetics*, pp. 302–307.

de Ruyter van Steveninck, R. R., W. Bialek, M. Potters, R. H. Carlson, and G. D. Lewen (1996). Adaptive movement computation by the blowfly visual system. In *Natural and Artificial Parallel Computation: Proceedings of the Fifth NEC Research Symposium*, D. L. Waltz, ed., pp. 21–41 (SIAM, Philadelphia)

de Ruyter van Steveninck, R. R., and S. B. Laughlin (1996a). The rate of information transfer at graded-potential synapses, *Nature* 379, 642–645.

de Ruyter van Steveninck, R. R., and S. B. Laughlin (1996b). Light adaptation and reliability in blowfly photoreceptors, *Int. J. Neural Sys.* in press.

de Ruyter van Steveninck, R. R., W. H. Zaagman, and H. Mastebroek (1986). Adaptation of transient responses of a movement-sensitive neuron in the visual system of the blowfly *Calliphora erythrocephela*, *Biol. Cybern.* 54, 223–236.

Sachs, M. B., and E. D. Young (1980). Effects of nonlinearities on speech encoding in the auditory nerve. *J. Acoust. Soc. Am.* 69, 858–875.

Sakai, H. M. (1992). White-noise analysis in neurophysiology, *Physiol. Rev.* 72, 491–505.

Sakitt, B. (1972). Counting every quantum, *J. Physiol. (Lond.)* 223, 131–150.

Sakmann, B., and E. Neher, eds. (1983). *Single Channel Recording* (Plenum, New York).

Salinas, E., and L. Abbott (1994). Vector reconstruction from firing rates, *J. Comp. Neurosci.* 1, 89–107.

Salzman, C. D., K. H. Britten, and W. T. Newsome (1990). Cortical microstimulation influences perceptual judgements of motion direction, *Nature* 346, 174–177. Erratum 346, 589.

Salzman, C. D., C. M. Murasagi, K. H. Britten, and W. T. Newsome (1992). Microstimulation in visual area MT: Effects on direction discrimination performance, *J. Neurosci.* 12, 2331–2355.

Schnapf, J. L., B. J. Nunn, M. Meister, and D. A. Baylor (1990). Visual transduction in cones of the monkey *Macaca fascicularis*, *J. Physiol. (Lond.)* 427, 681–713.

Schwartz, J. J., and A. M. Simmons (1990). Encoding of a spectrally-complex communication sound in the bullfrog's auditory nerve, *J. Comp. Physiol. A* 166, 489–499.

Segundo, J. P., G. P. Moore, L. J. Stensaas, and T. H. Bullock (1963). Sensitivity of neurones in *Aplysia* to temporal pattern of arrving impulses, *J. Exp. Biol.* 40, 643–667.

Seung, H. S., and H. Sompolinsky (1993). Simple models for reading neuronal population codes, *Proc. Nat. Acad. Sci. USA* 90, 10749–10753.

Shannon, C. E. (1948). A mathematical theory of communciation, *Bell Sys. Tech. J.* 27, 379–423, 623–656 (reprinted in Shannon and Weaver 1949).

Shannon, C. E. (1949). Communication in the presence of noise, *Proc. I. R. E.* 37, 10–21.

Shannon, C. E. (1951). Prediction and entropy of printed English, *Bell Sys. Tech. J.* 30, 50–64.



Shannon, C. E., and W. Weaver (1949). *The Mathematical Theory of Communication* (University of Illinois Press, Urbana).

Shapley, R. M., and J. D. Victor (1986) Hyperacuity in cat retinal ganglion cells, *Science* 231, 999–1002.

Shraiman, B. I., and E. D. Siggia (1994). Lagrangian path integrals and fluctuations in random flow, *Phys. Rev. E* 49, 2912–2927.

Siebert, W. M. (1965). Some implications of the stochastic behavior of primary auditory neurons, *Kybernetik* 2, 206–215.

Siebert, W. M. (1970). Frequency discrimination in the auditory system: Place or periodicity mechanisms?, *Proc. I. E. E. E.* 58, 723–730.

Simmons, A. M., and M. Ferragamo (1993). Periodicity extraction in the anuran auditory nerve. I: "Pitch-shift" effects, *J. Comp. Physiol. A* 172, 57–69.

Simmons, A. M., G. Reese, M. Ferragamo (1993). Periodicity extraction in the anuran auditory nerve. II: Phase and temporal fine structure, *J. Acoust. Soc. Am.* 93, 3374–3389.

Simmons, J. A. (1979). Perception of echo phase information in bat sonar, *Science* 204, 1336–1338.

Simmons, J. A. (1989). A view of the world through the bat's ear: The formation of acoustic images in echolocation, *Cognition* 33, 155–199.

Simmons, J. A., M. Ferragamo, C. F. Moss, S. B. Stevenson, and R. A. Altes (1990). Discrimination of jittered sonar echoes by the echolocating bat, *Eptesicus fuscus*: The shape of target images in echolocation, *J. Comp. Physiol. A* 167, 589–616.

Skilling, J., ed. (1989). *Maximum Entropy and Bayesian Methods: Proceedings of the Eighth Maximum Entropy Workshop at St. John's College, Cambridge, 1988* (Kluwer Academic, Boston).

Smakman, J. G. J., J. H. van Hateren, and D. G. Stavenga (1984). Angular sensitivity of blowfly phoreceptors: Intracellular measurements and wave-optical predictions, *J. Comp. Physiol. A* 155, 239–247.

Smirnakis, S., D. Warland, W. Bialek, and M. Meister (1995). Tiger salamander retina adapts to temporal contrast modulation to improve coding efficiency, *Invest. Ophthamol. Vis. Sci. (Suppl.)* 36, 624.

Smirnakis, S., M. Berry, D. Warland, W. Bialek, and M. Meister (1996). Retinal processing adapts dynamically to image contrast, in preparation.

Snyder, A. W., D. S. Stavenga, and S. B. Laughlin (1977). Spatial information capacity of compound eyes, *J. Comp. Physiol.* 116, 183–207.

Srinivasan, M. V., S. B. Laughlin, and A. Dubs (1982). Predictive coding: A fresh view of inhibition in the retina *Proc. R. Soc. Lond. Ser. B* 216, 427–459.

Srulovicz, P., and J. L. Goldstein (1983). A central spectrum model: A synthesis of auditory-nerve timing and place cues in monoaural communication of frequency spectrum, *J. Acoust. Soc. Am.* 73, 1266–1276.

Strong, S. P., R. Koberle, R. R. de Ruyter van Steveninck, and W. Bialek (1996). Entropy and information in neural spike trains, preprint.

Strutt, J. W., Baron Rayleigh (1877–78). *The Theory of Sound* (Macmillan, London); Reprint, from the 2d ed., revised and enlarged, with an introduction by R. Bruce Lindsay (Dover, New York, 1945).

Surlykke, A. (1984). Hearing in notodontid moths: A tympanic organ with a single auditory neurone. *J. Exp. Biol.* 113, 323–335.

Swindale, N. V., and M. S. Cynader (1986). Vernier acuity in cat visual cortex, *Nature* 319, 591–593.

Teich, M. C. (1989). Fractal character of the auditory neural spike train, *I. E. E. E. Trans. Biomed. Eng.* 36, 150–160.

Teich, M. C., D. H. Johnson, A. R. Kumar, and R. G. Turcott (1990). Rate fluctuations and fractional power law noise recorded from cells in the lower auditory pathway of the cat, *Hearing Res.* 46, 41–52.

Teich, M. C., and S. M. Khanna (1985). Pulse number distribution for the neural spike train in the cat's auditory nerve, *J. Acoust. Soc. Am.* 77, 1110–1128.

Teich, M. C., P. R. Prucnal, G. Vannucci, M. E. Breton, and W. J. McGill (1982a). Multiplication noise in the human visual system at threshold. I: Quantum fluctuations and the minimum detectable energy, *J. Opt. Soc. Am.* 72, 419–431.

Teich, M. C., P. R. Prucnal, G. Vannucci, M. E. Breton, and W. J. McGill (1982b). Multiplication noise in the human visual system at threshold. III: The role of non–Poisson quantum fluctuations, *Biol. Cybern.* 44, 157–165.

Theunissen, F. (1993). An Investigation of Sensory Coding Principles Using Advanced Statistical Techniques. Dissertation, University of California at Berkeley.

Theunissen, F., and J. P. Miller (1991). Representation of sensory information in the cricket cercal sensory system. II: Information theoretic calculation of system accuracy and optimal tuning curve widths of four primary interneurons, *J. Neurophys.* 66, 1690–1703.

Thorpe, S. J. (1990). Spike arrival times: A highly efficient coding scheme for neural networks. In *Parallel Processing in Neural Systems*, R. Eckmiller, G. Hartman, and G. Hauske, eds., pp. 91–94 (Elsevier, Amsterdam).

Thorpe, S., D. Fize, and C. Marlot (1996). Speed of processing in the human visual system, *Nature* 381, 520–522.

Tolhurst, D. J., Y. Tadmor, and T. Chao (1992). Amplitude spectra of natural images, *Ophthal. Physiol. Opt.* 12, 229–232.

Tovee, M. J., E. T. Rolls, A. Treves and R. P. Bellis (1993). Information encoding and the responses of single neurons in the primate temporal visual cortex, *J. Neurophys.* 70, 640–654.

Trappe, M. (1982). Verhalten und Echoortung der Grossen Hufeisennase (*Rhinolophus ferrumequinum*) beim Insektenfang. Dissertation, Universität Marburg.

Treves, A., and S. Panzeri (1995). The upward bias in information derived from limited data samples, *Neural Comp.* 7, 399–407.



Troy, J. B., and J. G. Robson (1992). Steady discharges of X and Y retinal ganglion cells of cat under photopic illuminance, *Vis. Neurosci.* 9, 535–553.

Valbo, A. B. (1995). Single afferent neurons and somatic sensation in humans. In *The Cognitive Neurosciences*, M. Gazzaniga, ed., pp. 237–252 (MIT Press, Cambridge, MA).

van der Velden, H. A. (1944). Over het aantal lichtquanta dat nodig is voor een lichtprikkel bij het menselijk oog, *Physica* 11, 179–189.

Verveen, A. A. (1961). Fluctuation in excitability: Research report on signal transmission in nerve fibers. Disseration, Netherlands Central Institute for Brain Research, Amsterdam.

Vickers, N. J., and T. C. Baker (1994). Reiterative responses to single strands of odor promote sustained upwind flight and odor source location by moths, *Proc. Nat. Acad. Sci. USA* 91, 5756–5760.

Volterra, V. (1930). *Theory of Functionals and of Integral and Integro–differential Equations* (Blackwell Scientific, London). Reprint, with a preface by G. C. Evans (Dover, New York, 1959).

Voss, R. F., and J. Clarke (1977). 1/f noise in music and speech, *Nature* 258, 317–318.

de Vries, Hl. (1943). The quantum character of light and its bearing upon threshold of vision, the differential sensitivity and visual acuity of the eye, *Physica* 10, 553–564.

Wagner, H. (1986a). Flight performance and visual control of flight in the free–flying house fly (*Musca domestica L.*). I: Organization of the flight motor, *Phil. Trans. R. Soc. Lond. Ser. B* 312, 527–551.

Wagner, H. (1986b). Flight performance and visual control of flight in the free–flying house fly (*Musca domestica L.*). II: Pursuit of targets, *Phil. Trans. R. Soc. Lond. Ser. B* 312, 553–579.

Wagner, H. (1986c). Flight performance and visual control of flight in the free–flying house fly (*Musca domestica L.*). III: Interactions between angular movement induced by wide- and small-field stimuli, *Phil. Trans. R. Soc. Lond. Ser. B* 312, 581–595.

Warland, D. (1991). Reading Between the Spikes: Real–Time Processing in Neural Systems. Dissertation, University of California at Berkeley.

Warland, D., M. Landolfa, J. P. Miller, and W. Bialek (1992). Reading between the spikes in the cercal filiform hair receptors of the cricket. In *Analysis and Modeling of Neural Systems*, F. Eeckman, ed., pp. 327–333 (Kluwer Academic, Boston).

Warland, D., and M. Meister (1993). The decoding of multi–neuronal signals from the retina, *Soc. Neurosci. Abs. 1993*, 1258.

Warland, D., and M. Meister (1995). Multi–neuronal firing patterns among retinal ganglion cells encode spatial information, *Invest. Opthalmol. Vis. Sci. Suppl.* 36, 932.

Wax, N., ed. (1954). *Selected Papers on Noise and Stochastic Processes* (Dover, New York).

Weinberg, S. (1983). *The Discovery of Subatomic Particles* (W. H. Freeman, San Francisco).

Weiss, T. F. (1966). A model of the auditory periphery, *Kybernetik* 4, 153–175.

Werner, G., and V. B. Mountcastle (1965). Neural activity in mechanoreceptive cutaneous afferents: Stimulus–response relations, Weber functions, and information transmission, *J. Neurophys.* 28, 359–397.

Wessel, R., C. Koch, and F. Gabbiani (1996). Coding of time–varying electric field amplitude modulations in a wave–type electric fish, *J. Neurophys.* 75, 2280–2293.

Westheimer, G. (1981). Visual hyperacuity, *Prog. Sens. Physiol.* 1, 1–30.

Wever, E. G. (1949). *Theory of Hearing* (John Wiley and Sons, New York).

Wiener, N. (1949). *Extrapolation, Interpolation and Smoothing of Time Series* (Wiley, New York).

Wiener, N. (1958). *Nonlinear Problems in Random Theory* (MIT Press, Cambridge MA).

Wilson, K. G. (1975). The renormalization group, critical phenomena, and the Kondo problem, *Revs. Mod. Phys.* 47, 773–840.

Wilson, K. G. (1983). The renormalization group and critical phenomena, *Revs. Mod. Phys.* 55, 583–600.

Wilson, M. A., and B. L. McNaughton (1993). Dynamics of the hippocampal code for space, *Science* 261, 1055–1058.

Winslow, R. L., P. E. Barta and M. B. Sachs (1987). Rate coding in the auditory nerve. In *Auditory Processing of Complex Sounds*, W. A. Yost and C. S. Watson, eds., pp. 212–224 (Erlbaum, Hillsdale NJ).

Wolf, R., and M. Heisenberg (1990). Visual control of straight flight in *Drosophila melanogaster*, *J. Comp. Physiol. A* 167, 269–283.

Young, E. D., and M. B. Sachs (1979). Representation of steady–state vowels in the temporal aspects of the discharge patterns of auditory–nerve fibers, *J. Acoust. Soc. Am.* 66, 1381–1403.

Zaagman, W. H., H. A. K. Mastebroek, and J. W. Kuiper (1978). On the correlation model: Performance of a movement–detecting neural element in the fly visual system, *Biol. Cybern.* 31, 163–168.

Zaagman, W. H., H. A. K. Mastebroek, and R. R. de Ruyter van Steveninck (1983). Adaptive strategies in fly vision: On their image processing qualities, *I. E. E. E. Trans. Sys. Man Cybern.* 13, 900–906.

Zohary, E., P. Hillman, and S. Hochstein (1990). Time course of perceptual discrimination and single neuron reliability, *Biol. Cybern.* 62, 475–486.

Zohary, E., M. N. Shadlen, and W. T. Newsome (1994). Correlated neuronal discharge rate and its implication for psychophysical performance, *Nature* 370, 140–143.

# Index

1/f noise, 52, 266. *See also* Natural signals

Absence of spikes, information conveyed by, 65
Action potential. *See also* Spikes
  all-or-none law, 5
  deterministic description of, 35
  generation of, 6
  noise in the propagation of, 36–37
Adaptation, 7–9
  and invariant representation, 81
  optimal coding and computation, 275
Adaptive nonlinearity, poorly described by power series, 48
Adrian, 3–8, 31–32
All-or-none law, 4–5
Ambiguity
  amplitude/frequency, 37–38
  in coding, 37, 77–78, 81, 96
  Kolmogorov, 82n (*see also* Kolmogoroff)
Aplysia, 33
Atick, 268–272
Auditory neurons
  phase locking of, 30–32
  synchronous activity in, 38
  transient response, 153
Auditory system
  amplitude/frequency ambiguity, 37–38
  bat, 36, 55, 191–192, 231–235 (*see also* echolocation)
  cat, 37, 204
  frequency discrimination, 204–214
  frog, 30, 181–185
  moth, 191–193
  owl, 31
Autocorrelation function, 35
Available information. *See also* Information
  captured by linear reconstruction, 170
  conditions for it to be well-defined, 110
Average information, 122. *See also* Information
Averaging over many neurons. *See* Pooling of Spike trains

Barlow, 9, 153. *See also* Levick; Discrimination; Dark noise
  photon counting, 197–201
  redundancy reduction, 268
Barn owls, sound localization, 31
Bat
  auditory cortex, 36
  echolocation, 55, 231–235
  vs moth, 191–192
Bayes' rule
  definition, 23
  encoding/decoding, 21–28
  requirements for decoding, 61
Behavioral time scales
  sensory noise, 238
  structure of code, 95–96
Bit, as unit of entropy or information, 106
de Boer, 45
Born, xvii
Brillouin, 114

Carlin, dirty words, 23
Cat
  auditory nerve, 37, 204
  visual system, 197–201
Causality, 89, A.3, A.8
Cave beetle, 120–121
Center-surround, 9–11. *See also* Receptive fields
Cercal system
  cricket, 166–167
  population coding, 257–258
Coding
  asymmetry, 90
  with second order statistics, 37–38

Coding *(cont.)*
and smoothness in the stimulus domain, 72
Coding efficiency, 17
auditory afferents, 185
for cricket mechanoreceptors, 173–174
definition, 170
for naturalistic sounds, 185–186
and reproducibility 186–187
for retinal ganglion cells, 179
for sacculus, 176–177
and timing precision of individual spikes, 172–173
and upper bound to spike train entropy, 172
Coding strategy. *See also* Encoding/decoding
and moments in the conditional distribution of spike trains, 38
Color discrimination, 227
Communication channel, and information theory, 102–104. *See also* Shannon
Compound eye design, 276
Conditional distribution, 28, 32, 61
experimental characterization of, 63
of spike trains, and coding strategy, 38
Conditional entropy. *See* Entropy
Conditional mean, as optimal estimator, 79, 83, 306–310, A.7
Conditional probability, 21
Conditional rate. *See* Correlation function
Conditional stimulus ensemble, information rate, 162
Context, relevance of, 62, 81
Continuous stimulus estimate. *See* Decoding
Correlated firing
and reliability of motion detection, 217–218
role in coding, 259–260
Correlation function, 33, A.2
relation to power spectrum, 137–139, A.14
and time dependent rate, 53
time translation, 159
Correlation time, 138
of frog calls, 181
and success of decoding, 86–87
Cortex
auditory, 36
IT, 153
motor, 256
somatosensory, 12, 57
visual, 57, 214–221, 229
Covariance matrix, 137, 159
Cricket cercal system
information transmission by, 166–175
population coding in, 257–258
Crosscorrelation function, 288

Dark noise, 196, 202
Decoding, 17. *See also* Estimation
ambiguity, 77–78, 81
causality, 89, A.3, A.8
conditional probabilities, 61
criterion for quality of estimate, 81
dependence on input statistics, 80, 88
difficulties in, 76–77
dynamics differ from encoding dynamics, 80
error metrics, 89
estimation of continuous stimulus, 62, 65
with an exponential prior, 84–85
limits to linear decoding, 85
with linear filters, 81, 91–93
linearity of, vs. nonlinearity of encoding, 92–94
for Poisson model, 77
reading the neural code, 76
reliability and decision time, 95
requirements for success, 76–78, 81
Deconvolution, 225
Decorrelation, and redundancy reduction, 270
Delay, effect on reconstruction, 95
Detectability, and effective input noise, 124
DeWeese, 272–273
Diffraction, physical limits, 224–225
Dirac delta function, A.1
Direction selectivity. *See also* H1
ganglion cells, 230
Discriminability
$d'$, 239–247
and effective input noise, 124
response conditional ensembles, 70–71
Discrimination
acuity and hyperacuity, 225–227
based on spike counts, 245
color, 227–228
displacement, 236–247
and interspike intervals, 71–72
temporal, 231–233

Echolocation. *See* Bat
Eckhorn and Pöpel, 150–151
Effective input noise. *See also* Noise
detectability and discriminability, 124
as a measure of reconstruction quality, 164, 249–250
Effective noise level, and natural scale of signals, 47
Eigenvalues of covariance matrix, 137
Eigenvectors of covariance matrix, 72–73
Electrical activity of nerves, 2
Electric fish, 231
Electrolocation, 231
Encoding
described by conditional probability, 61



dynamics differ from decoding dynamics, 80
nonlinearity of, vs. linearity of decoding, 92–94
and small expansion parameter, 87
Encoding/decoding, 27, 85–86
Ensemble
of inputs and outputs, 44
of stimuli, 62
Entropy
and available information, 104
calculation of observable quantities, 110
conditional, 121
of continuous variables, 105, 109
extensive quantity, 115
as function of timing precision, 116, 172–174
of Gaussian distribution, 108, A.9
maximizing, 116–119, 118n, 125
measurement of, 109
and number of possible states, 105
and physical limits, 113–121, 127
of probability distribution, 104n
rate of, 116
of spike count distribution, 116, 154
of spike trains, 113–121, 172–174, A.10
Envelope of sound pressure waveform, 81
Ergodicity, 44
Error metric and decoding, 88, 97–98
Errors, random and systematic, 179–180
Estimation, A.8. *See also* Decoding
difference with conventional input/output analysis, 82
and displacement resolution, 247–253
of Gaussian signal in Gaussian noise, 83
nonlinearities, 170, 176
optimal linear filtering, 82
Expansion parameter, 46–47, 87
Explosion 49, 74–76
and U.S. budget, 156
Exponential
maximum entropy distribution, 118–119
series expansion of, A.4
Extensive quantity. *See* Entropy

Fano factor, 52–53. *See also* Spike count distribution
Feature selectivity, 9–12
Field, 263–264. *See also* Natural images
Firing
probability of, 30
temporal precision of, 35
Firing rate
defined, A.1
first moment of conditional distribution, 28
time-dependent, 28, 30 (*see also* Rate code)
Firing statistics
approximate descriptions of, 49–54
bursting, 37
cortex, 33
FitzHugh, 62–63
Flight stabilization, 236
Fly
motion processing, 63
neuroanatomy, 64
visual system, 236
Flynculus, 13–14. *See also* Homunculus
Fourier analysis, 128–133
Fourier series, 128–129
Fourier coefficients, variance of, 131
Frequency discrimination in auditory system, 204–214
Frequency of seeing
psychophysics, 194–196
in retinal ganglion cells, 199–201
Frog
auditory system, 30, 181–185
calls and decoding, 181–186
retina, 9
sacculus, 176–177
Functional, 40, A.7
Functional derivative, A.7, A.11, A.13, A.15

Gaussian channel and mutual information, 123–127, A.12
Gaussian distribution
entropy of, 108, A.9
and maximum entropy, 125, A.13
Gaussian input distribution, vs. nongaussian output distribution, 93
Gaussian random functions, 130
Geisler, 225
Golomb, 153
Green and Swets, 14, 196n, 239–241

H1 neuron, 63
Hamlet, 110
Hartline, 3, 9
Hecht, 194–196
Hippocampus, 58, 258–259
Hobgoblin, xvii. *See also* Homunculus
Hodgkin, 6
Hodgkin-Huxley equations, 6, 35
Homunculus, 13–15, 22, 89, 213. *See also* Flynculus
impoverished, 16, 60, 255
providing a running commentary, 15
statistically sophisticated, 16
Horseshoe crab, 9. *See also* Limulus
Hubel, 10–12
Huxley, 6

Hyperacuity, 221–235
  echolocation, 234–235
  motion detection, 235–253
  and single neurons, 228–231

Impedance, 41–42
Information
  available amount of, 104
  carried by spike count, 116–119
  conveyed by absence of spikes, 65
  linear reconstruction, 170
  lower bound, 160–162
  mutual, 122, 123
  upper bound, 126
Information capacity, 126
  of photoreceptors and LMCs, 146–147
Information content
  of natural images, 112
  of rare events, 112–113
  and sparseness of spike train, 120
Information rate
  computed from the conditional stimulus ensemble, 162
  and decoding, 166–186
  Poisson model, 273
  relation to signal to noise, 140
  strategy for estimating, 156–163
Information theory
  and language, 111–112
  Shannon's formulation of, 102
Information transmission
  chemical synapse, 147
  coding efficiency, 173
  dependence on integration time, 150, 155
  for Gaussian signal and noise, 138–140
  maximizing, A.15
  measured by stimulus reconstruction, 156–165
  measurement for graded potential cells, 141–148
  optimal filters, A.19
  optimal signal spectrum for, 140
  in rate code, 116–119
  by sensory neurons, 102–103
  in spiking cells, 150–156
Input/output analysis, 38–48
Input/output relation, nonlinearity of, 28
Integration time and temporal precision. *See* Temporal precision
Intensity discrimination, 197–199
Interspike interval distribution, 33, 35
Interval coding and small numbers of spikes, 55
Interval distribution. *See* Interspike interval distribution
Invariance of information theoretic quantities, 108–109

Johannesma, 63

Kolmogoroff, 82
Kolmogoroff-Wiener estimation, 82. *See also* Estimation
Kuffler, 10, 153

Lagrange multiplier, A.11, A.13, A.15
Lateral geniculate, 153
Lettvin, 10
Levick, 197–210
Limulus, 9, 197
Linear decoding, 91–95. *See also* Decoding
  conditions for, 87
  and nonlinear encoding, 85
Linear response, 41
Log-likelihood function, 206–207
Logarithm
  concavity of, 161
  expansion of, A.4, A.13, A.15
Lucas, 3

Mach bands, 9–10
MacKay and McCulloch, 113, 149, 168, 172
Maps
  cortical, 12
  directional, 167
Matched filtering, 274
Maximum entropy, 116–118, 118n. *See also* Entropy
  distribution, A.13
  information theory, A.11
Maximum likelihood, 77, 205, A.16
Mean effective stimulus, 44
Motion detection
  fly vision, 63
  monkey cortex, 214–221
Mountcastle, 12, 154–155
Multineuron coding, 255–260
  decoding, 178–179
  cricket, 258
  hippocampus, 258–259
  retina, 259–260
Mutual information, 122–123. *See also* Information
  as difference between output and noise entropies, 126
  Gaussian channel, 123, 126, A.12

Natural images. *See also* Statistics of natural images
  contrast distribution, 47, 146



  exponential tails in distribution, 265–266
Natural signals, 261–266
  olfactory processing, 262
  probability distributions of, 113, 261–266
  statistics of natural images, 262–263
  and time averages, 15, 119–120
  time scales compared to interspike intervals, 56–59
  visual processing, 263–264
Natural sounds
  1/f spectrum, 266
  and coding efficiency, 181–186
  and reliability, 53
Natural stimuli, 15–16
  decoding, 62
  statistics, 112–113
  and time constants in decoding, 80
Neural code
  dictionary for, 1, 65
  performance on an absolute scale, 120
Neural response, complete description of, 28
von Neumann, 190
Newsome, 214–221
Noctoid moths, 191–193
Noise
  1/f, 52
  in action potential propagation, 36–37
  and effect on estimation, 84–85
  in synaptic transmission, 36–37, 145
Noise level and temporal precision. *See* Temporal precision
Noise whitening and information transmission, 141, A.15
Nonlinear encoding and linear decoding, 85, 92–94
Nonlinearities
  adaptive, 48
  and auditory coding, 186
  in auditory system, 47
  in decoding, 170, 176
  in encoding, 27–28,70, 92–94
Normalization of Poisson model, A.4

Odors, dynamics of, 261
Optican, 151–152
Optimal coding, 272–274
Optimal coding and computation, 267–277
Optimal computation
  adaptation, 275
  motion estimation, 274–275
  photon detection, 274
Optimal estimate. *See also* Decoding
  expansion of, 87
Optimal estimation and conditional mean, 79, A.7
Optimal filters, information transmission, A.19
Orthogonalization in Wiener analysis, A.3

Parseval's theorem, 136
Phase locking, 30–32
Photon counting, 193–204
Photoreceptor
  fly, 141–145
  noise and photon counting, 202–204
Physical limits, 17
  coding efficiency, 173
  diffraction, 224–225
  displacement discrimination, 246–248, 251–253
  echolocation, 234
  and optimal computation, 267
  and small number of spikes, 59
  and spike train entropy, 120, 127, 155
  synaptic transmission, 147–148
Pirenne, 194–196
Place cells, 59, 258–259
Poisson distribution, 51
Poisson process
  averages over, A.17
  coding in lateral geniculate, 153
  and decoding, 77–79
  and frequency discrimination, 205
  as model of firing statistics, 49
  photon counting, 193–204
  and probability of spike sequence, 50–51, A.4
  and spike-count distributions, 51, A.5
Pooling of spike trains from many cells, 59
Population coding, 256
Population vector, 256–257
Post-stimulus time histogram, 20, 29
Power series and adaptive nonlinearity, 48
Power spectral density, 133–136
Power spectrum, 134–136
  and correlation function, A.14
  units, 136
  and variance of Fourier components, 131–132
Precision of spike timing, 70–71
Primary visual cortex and small number of spikes, 57
Principal components
  and estimation of transmitted information, 151
  and power spectrum, 137–138
Prior distribution, 22, 271
Prior knowledge and information theory, 110–113
Probability distribution
  conditional, 22

Probability distribution *(cont.)*
finite data, 150–152, 155–156
joint, 22
learning of, 111
of natural signals, 113, 261–266
prior, 22
of random functions, 130
Probability theory, 21
Pulse number distribution. *See* Spike count distribution

Quantum bumps, 197

Random and systematic errors, 164–165
Random function
and Fourier series, 130
entropy, 127
probability distribution of, 130–131
relation to the variance of its Fourier components, 133
Rare events, information content of, 112
Rate code, 7–8
as distinct from timing code, 12, 29, 54, 60, 119, 173
and information transfer, 116–119
smooth transfer to timing code, 119
and time dependent firing rates, 31–32
two meanings of, 29, 32
Rate of entropy. *See* Entropy
Rayleigh, 30
Receptive fields
center-surround, 9, 153
organization, 260
Reconstruction. *See* Decoding
Redundancy reduction, 268–272
Refractoriness 53
Reichardt model, 251
Reliability
of computation, 17
and decision time, 95
of nervous system, 189
of neurons and perception, 154
and reproducibility, 21, 34
and unreliable components, 190
Renewal process, 54
Renormalization, 264
Representation, of stimulus, invariant, 81, 97
Reproducibility
coding efficiency, 186–187
cortex, 220
Response functions, 38
measurement of, 43
Response-conditional ensembles, 22, 63, 66–70
and discriminability, 70–71
and stimulus estimation, 74–75, A.6
Retina
frog, 9
ganglion cells, 229
multineuron recording, 259–260
salamander, 178–180
Reverse correlation, 38
Reverse correlation function, 44–45
and stimulus estimation, 88
Richmond, 151–152
Robust estimation, conditions for, 77
Roeder, 191–193
Ruderman, 264–266

Sacculus, 176–177
Salamander retina, 178–180
Scale invariance of natural images, 263–266
Second order statistics and coding, 37–38
Sensory neurons as communication channel, 102–103
Serial correlation in the spike train, 74
Shannon, 102–113, 140, 148
Shlaer, 194–196
Siebert, 204–214, 256
Sigmoidal input/output relation. *See* Input/output relation
Signal detection theory, 14, 196n, 239–241
Signal to noise
and auditory discrimination, 209–212
cercal system, 168–171
and correlated firing, 217–218
in echolocation, 233–234
and information, 124, 140, 165, 169–170, 184
retinal ganglion cells, 179–180
sacculus, 176–177
for white noise, A.18
Signals
ensemble of, 22
in the real world, 102
Signals and spike trains, joint distribution of, 22
Simmons, 55–56, 231–235
Single spike, importance of, 60, 279
Small numbers of spikes, 55–61
Somatosensory cortex and small number of spikes, 55
Sound localization, 30–31
Sound pressure waveform, 81
Sparse coding in the time domain, 60
Sparse events, information content of, 120
Specific nerve energies, 2
Spike count
information carried by, 116–119, 154
maximum entropy distribution, A.11
mean and variance of, 51–54



Spike generating mechanism and refractoriness, 53
Spike statistics. *See* Firing statistics
Spike trains
conditional probability of, 21
ensemble of, 29
entropy proportional to their length, 115
meaning depends on stimulus ensemble, 80–81
upper bound to spike train entropy and timing precision, 172–173, A.10
Spike triggered average, 26, 44–45, 87–88
Spikes, significance of small numbers of, 55–59
Spontaneous rate, 30
Squid giant axon, 6
States, number of states and entropy, 105
Stationarity
and correlation function, 137
random functions, 130, 133
Statistics of natural images, 262–266. *See also* Natural images/signals
Field, 263–264
Ruderman, 264–265
Stimulus
covariance matrix of conditional stimulus ensemble, 159
invariant representation of, 81, 97
mean conditional stimulus waveform, 159
reconstruction and rate of information transmission, 156–165
Stimulus ensemble and adaptation, 81
Stimulus reconstruction. *See* Decoding, estimation
Stirling's approximation, 115
Stretch receptor, rate coding, 6–7
Synaptic transmission
and information theory, 147–148
noise in, 36–37
Systematic errors
and information rate, 157
and prior knowledge, 84
Systems identification, 38

Taylor series, 39–40
Teich, 51–52
Temperature detection in *Speophyes lucidulus*, 120–121
Temporal precision, 36
Time dependent firing rate (*see* Firing rate)
Time scales of natural signals, 55–57
Time translation invariance (*see* Stationarity)
Timing code, 29
as distinct from rate code, 12, 29, 54, 60, 119, 173
Timing errors, robustness to, 72, 91
Timing jitter, 36
Timing of spikes, 151–152
Timing precision
coding efficiency, 172–173
and entropy (*See* Entropy)
of neural code, 72, 91–92
Tovee, 153
Transfer function, 41
Triggered correlation function, 44
Turbulence and olfactory stimuli, 261–262

Units and effective noise, 124

Vestibular system, firing statistics, 33
Vibration sensors, 176. *See also* Sacculus
Visual acuity, 221–227. *See also* Hyperacuity
Visual system
cat, 197–201
fly, 236
monkey, 151–152
Volterra, 40
Volterra series, 43, A.3

Werner, 154–155, 170
White noise, A.3
White noise analysis, 38, 43–46
Wiener, 39
approach to systems identification, 43–46, A.3
and estimation, 82
kernel, 44–45, A.3
Wiener-Khinchine theorem, 137–138, A.14
Wiener-Volterra series
and adaptation, 48
convergence of, 47
Wiesel, 10–12

Xenopus, phase locking, 31

**SPIKES**

Exploring the Neural Code

Fred Rieke, David Warland, Rob de Ruyter van Steveninck, and William Bialek

What does it mean to say that a certain set of spikes is the right answer to a computational problem? In what sense does a spike train convey information about the sensory world? *Spikes* begins by providing precise formulations of these and related questions about the representation of sensory signals in neural spike trains. The answers to these questions are then pursued in experiments on sensory neurons.

Intended for neurobiologists with an interest in mathematical analysis of neural data as well as the growing number of physicists and mathematicians interested in information processing by "real" nervous systems, *Spikes* provides a self-contained review of relevant concepts in information theory and statistical decision theory.

Fred Rieke is Assistant Professor in the Department of Physiology and Biophysics at the University of Washington. David Warland is Research Associate in the Department of Molecular and Cellular Biology, Harvard University. Rob de Ruyter van Steveninck is Research Scientist and William Bialek is Senior Research Scientist, both at the NEC Research Institute.

"A joy to read. . . . This book will undoubtedly become a classic. The ideas presented in it have already begun (in no small part through the work of the authors) to reshape our views of the neural code. This book will make them accessible to a much wider audience."
—Anthony Zador, *Science*

Computational Neuroscience series
A Bradford Book

9 780262 681087 90000

cover design by James McWethy

THE MIT PRESS | Massachusetts Institute of Technology | Cambridge, Massachusetts 02142 | http://mitpress.mit.edu | 0-262-68108-0 RIESP