Information Maximization and Stochastic Resonance in Single Neurons

Thesis by

Martin Bernard Stemmler

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

1997

(Submitted)

Acknowledgements

Caltech is as close to a research paradise as one could hope for, so it is bittersweet to leave this place for other venues. Thanks to Christof Koch for providing the environment, the facilities, and having an abiding sense of good humor. I was able to pursue the topics of my own choosing, without being set too many hurdles or asked too many questions. The freedom and independence I enjoyed here would not have been possible at other universities, so I shall always bear a special allegiance to Caltech. Of course, the Howard Hughes Medical Institute enabled me to be here in the first place, thanks to a generous predoctoral fellowship—I am probably one of the few physicists, or theorists, for that matter, they have ever funded, so I am still amazed that they decided to take a risk on me. My thanks, in no particular order, go to Dr. Ernst Niebur, Dr. Marius Usher, Dr. Frank Moss, Dr. Adi Bulsara, Dr. Heinz Schuster, Dr. Tony Zador, Gary Holt, Dr. Humbert Suarez, Liz Ho, all of the members of the thesis committee, my colleagues in the Computation and Neural Systems option, my parents, and my brothers, Bertram and Robin.

Abstract

Does the nervous system "tune" itself to perform at peak efficiency? Optimal transmission of information in a single nerve cell occurs when the response is matched to the statistics of naturally occurring stimuli, such that all firing rates are used with equal probability and that redundant temporal correlations in the input are removed. Non-Hebbian, local learning rules are developed to adapt the voltage-dependent ionic conductances in Hodgkin-Huxley models of neurons with the goal of matching the neuron's response to the statistics of natural stimuli. These learning rules allow a nerve cell to maximize the amount of information transmitted about arriving stimuli. At a more detailed level, information transmission in neurons is limited by the noise in the input, defined as the root mean square of the fluctuations in the input. Three different performance measures are shown to scale identically as a function of the noise in simple models of neurons that have both a voltage and current threshold. These performance measures are: the probability of correctly detecting a constant input in a limited time, the signal-to-noise ratio in response to sinusoidal input, and the mutual information between an arbitrarily varying input and the output spike train of the model neuron. Of these, detecting a constant signal is the simplest and most fundamental quantity. For subthreshold signals, the model exhibits stochastic resonance, a non-zero noise amplitude that optimally enhances signal detection. In this case, noise paradoxically does not limit, but instead improves performance. Even

though the noise amplitude can dwarf the signal, detection of a weak constant signal using stochastic resonance is still possible when the signal elicits on average only one additional spike. Stochastic resonance could thus play a role in neurobiological sensory systems, where speed is of the utmost importance and averaging over many individual spikes is not possible.

Contents

Acknowledgements i

Abstract ii

1 Prologue 1

2 Information and Entropy Maximization 6

2.1 Overview 6

2.2 The Kinetics of Ion Channels as a Substrate for Adaptation 12

2.3 Information Maximization 29

2.4 The Adaptation or Learning Rule 41

2.5 The Set of Basis Functions 51

2.6 Feedforward Learning 59

2.7 Computing with the Steady State–The Role of Feedback 71

2.8 Stability of the Steady State* 78

2.9 Simulations Gallery 84

2.10 The Hodgkin-Huxley Spiking Model 97

2.11 Effective Conductance and Current-Discharge Relationships for Hodgkin-Huxley Spiking Models with Adaptation* 105

2.12 Hodgkin-Huxley Spiking Model Simulations 115

2.13 Time-Varying Input Signals 130

3 Stochastic Resonance **147**

3.1 Noise and the Speed of Information Processing 147

3.2 Fisher Information and Signal Detection 151

3.3 The Spike Count Distribution and the First Passage Time 155

3.4 The Stochastic Resonance 159

3.5 Resonance versus Linearization 164

3.6 Discriminability, Mutual Information, Detection Probability, Information Capacity, and Coherence 173

3.7 Reducing the Threshold 180

3.8 Spike Times or Firing Rate? 183

3.9 Conclusion 188

Bibliography **200**

A Appendix **212**

A.1 Simple Model for Ca^{2+}-dependent Learning 212

A.2 Notes on the Hodgkin-Huxley Spiking Model 215

A.3 Mean and Variance of the First Passage Time Distribution 220

A.4 Note on the Effect of Firing Rate Saturation 222

A.5 The Poisson Estimate of the Fisher Information 223

List of Figures

1.1 Model orientation tuning curve 2

1.2 Optimal orientation tuning curve 3

2.1 Matching a neuron's response to the input 7

2.2 Two-state ion channel energy diagram 15

2.3 The steady-state activation curve and the frequency of channel events 19

2.4 Different models of ion channel kinetics 22

2.5 Regulation of channel properties 24

2.6 Model of transition/state sensor using a G-protein mechanism 29

2.7 Analogy to simulated annealing 48

2.8 The basis functions for learning 54

2.9 Function approximation by sigmoidal basis functions 56

2.10 Nonlinear term in learning rule on a linear scale 66

2.11 Nonlinear term in learning rule on a logarithmic scale 67

2.12 Single compartment model 72

2.13 Target probability distribution of firing rates 83

2.14 Numerical computation of the steady state 86

2.15 Fixed current-discharge relationship 87

2.16 Learning in the context of a nonlinear dynamical system 88

2.17 Varying the nonlinearity in the $f(I)$ curve 90

2.18 Synaptic feedback 91

2.19 Firing rate distributions with and without synaptic feedback 93

2.20 Entropy of the firing rate distribution in time 94

2.21 Evolution of the peak conductances 96

2.22 Schematic of a spiking cell's voltage trace 106

2.23 Unadapted current-discharge relationship 108

2.24 Current injection *in vitro* 112

2.25 Adaptation of f-I curves and the effective conductance 113

2.26 The spiking Hodgkin-Huxley model 118

2.27 Learning in the Hodgkin-Huxley spiking model 123

2.28 Matching the response to the input statistics 125

2.29 Varying the Coupling Conductance 126

2.30 Readaptation 128

2.31 Venn diagram of entropies 131

2.32 Power spectra of conductance input and membrane voltage 142

2.33 Sample traces 143

3.1 Signal Detection 153

3.2 Stochastic resonance in the Fisher information 194

3.3 Validity of asymptotic expansion 195

3.4 Noise linearization of the input-output relationship 195

3.5 Mutual information between the signal and spike count 196

3.6 Probability of correct detection in a two-alternative forced choice test 197

3.7 Mutual information and detection probability as a function of mean current μ and its standard deviation σ 198

3.8 Information capacity of leaky I&F unit 198

3.9 Mutual information as a function of the bias 199

3.10 The number of spikes required for signal detection 199

List of Tables

2.1 Nonlinear term in learning rule 65

2.2 Adapted parameters of the voltage-dependent Ca^{2+} conductances for fig. 2.27 121

2.3 Adapted parameters of the voltage-dependent K^{+} conductances for fig. 2.27 121

3.1 Small Signal Relationships 177

Chapter 1 Prologue

What can optimality principles teach us about the functioning of the nervous system? We will investigate from a purely theoretical perspective how single neurons could adjust the number of ion channels in the membrane or even the level of noise to be more efficient and transmit more information. But before even assuming that the nervous system is in any sense optimal, it is instructive to begin with a cautionary tale.

The nervous system is sometimes far from optimal, as can be seen from a careful consideration of the mathematical approach taken by two physicists, Seung and Sompolinsky (1993), that tackles the problem of encoding visual information in networks of neurons. We set the stage of this tale as follows:

A characteristic feature of neurons in primate primary visual cortex (V1) is that they respond to oriented stimuli within restricted regions of the visual field. So-called simple cells in V1 have a distinct preference for one particular orientation over others. One can measure an orientation tuning curve for such cells, plotting the neuronal firing rate response as a function of stimulus orientation.

Seung and Sompolinsky (1993) used the following function to characterize the

orientation tuning curve of simple cells in V1

$$f(\theta) = \begin{cases} f_{\min} + (f_{\max} - f_{\min}) \cos^2(\frac{\pi\theta}{2w}) & |\theta| < w, \\ f_{\min} & |\theta| \geq w, \end{cases} \tag{1.1}$$

where the parameter w measures the width of the tuning curve. This parameterization, displayed in fig. 1.1, yields a fair representation of actual orientation tuning curves for neurons in visual cortex.

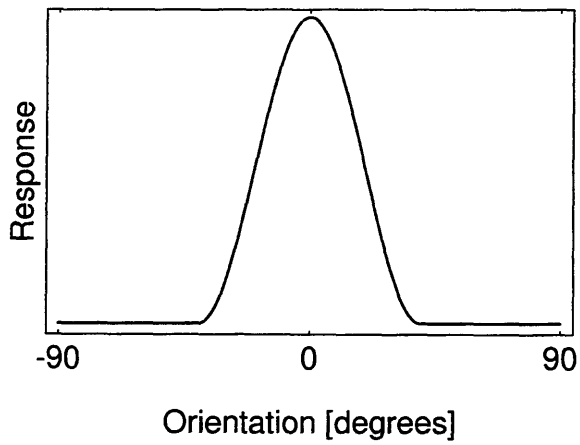


Figure 1.1: Neurons in primary visual cortex respond preferentially to particular orientations of an input stimulus. Displayed here in the graph is a parameterization of an orientation tuning curve given by eqn. 1.1.

The optimal tuning curve depends on the "noise" present in the representation of information. Typically, the variance of the firing rate σ_f^2 will behave as some power law of the firing rate itself:

$$\sigma_f^2 \sim f^\alpha,$$

where $\alpha = 1$ if the timing of spikes is completely random. The technical term for such

a random spike train is a Poisson point process. Suppose now that the orientation ϕ of the stimulus is uniformly distributed over all possible angles. As shall be discussed in chapters 2 and 3, the information about the stimulus orientation contained in the orientation tuning curve of firing rates $f(\phi)$ can be quantified by the following expression:

$$\text{information} \propto \frac{1}{2}\int \ln\left\{\frac{[f'(\phi)]^2}{[f(\phi)]^\alpha}\right\} d\phi + \text{constant}$$

The ideal orientation tuning curve that maximizes the information has the shape displayed in fig. 1.2.

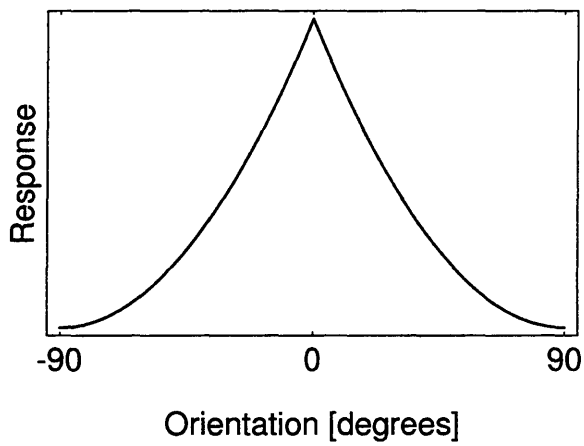


Figure 1.2: Optimal firing rate tuning curve for orientation in a Poisson neuron. In the Poisson case, the discrimination between two nearby angles θ and $\theta + d\theta$ does not depend on θ if the tuning curve is optimal.

Such a cusp-like shape to the orientation curve is *never* observed in experimental recordings. Whether "maximum likelihood estimation ... possesses a plausible biological implementation" that can make use of the "optimized" information is an ancillary question. Primary visual cortex is and remains suboptimal in the encoding

of stimulus orientation information.

Given that the primary visual cortex is not optimal, Seung and Sompolinsky (1993) pose the following question: If the orientation tuning curve is given by eqn 1.1 and orientation is encoded as the vector sum of the preferred orientations of all neurons weighted by their respective firing rates, is there an optimal width w to the orientation tuning curve? Seung and Sompolinsky derive the conclusion that the optimal width depends on the ratio of the peak firing rate $f_{\max}$ to the spontaneous firing rate $f_{\min}$:

$$w_{\text{opt}} \propto \left(\frac{f_{\min}}{f_{\max}}\right)^{\frac{1}{3}}$$

However, the crucial point to note is that this ratio is on the order of 10^{-2} and that it varies from cell to cell over a wide range. Raised to the one-third power, this ratio can easily lead to optimal orientation tuning curve widths at half-maximum that are 10°, 30°, or even 60°. From the converse perspective, the simple method of encoding orientation in the vector sum of preferred orientations is robust: discrimination performance changes little as the orientation tuning width changes, since the maxima in any performance criterion are extremely broad as a function of tuning width w.

In short,

The width of the orientation tuning curve doesn't matter.

Thus the theory of optimality does not allow one to predict the response of cells to oriented stimuli. (However, similar theoretical considerations do lead to valuable insights into stereo vision (Lehky and Sejnowski, 1990): how does the visual system decide whether an object is in front of or behind the plane of focus? Tuning

curves for the depth of an object in three dimensions are much closer to optimal for discrimination purposes than orientation tuning curves.)

So, instead of asking the wrong question, we ask:

How does a single neuron represent *any* information in the first place?

Note that there is no such thing as a prototypical cortical neuron, at least not in mammalian neocortex. Given the wide variation in the properties of neurons that belong to the same class, neurons should have some capability of regulating themselves, adapting their overall structure and constituent membrane properties, even adjusting the level of fluctuations in the transmembrane voltage. By considering what would be optimal, we will gain some insight into how such variable neurons learn to represent stimulus information.

Chapter 2 Information and Entropy Maximization

2.1 Overview

For the brain to compute, cortex must first find a neuronal representation of the environment, converting the sights, sounds, and smells that arrive into neuronal events, such as spikes traveling down axons to carry information from one location to the next in the brain. The encoding and decoding of neuronal information about the outside world will depend on the representation used: information can be represented in the timing of individual spikes of neurons, the timing of spike burst patterns, the rate with which packets of neurotransmitter are released, or the time-averaged firing rate. A further distinction can be made between information encoded in the output of individual cells or in the distributed patterns of activity over many different neurons.

Possibly the simplest representation of information is the time-averaged firing rate of a single cell, i.e., the number of action potentials generated by a cell in response to a stimulus over a set time window. Given such a representation, how could a neuron adapt itself to best encode the stimulus amplitude, such as visual contrast, sound pressure, or odor intensity?

Stimuli that arrive through the senses are neither completely predictable nor com-

pletely random. If a stimulus were completely predictable, then the explicit or implicit representation of that stimulus is not required for computation. On the other hand, if the stimulus amplitude were allowed to take on any real-numbered value at any instant in time, then this form of complete randomness would overwhelm the limited dynamic range with which neurons can respond—the firing rate of a cell cannot exceed some maximal value, whereas a negative number of spikes is not possible.

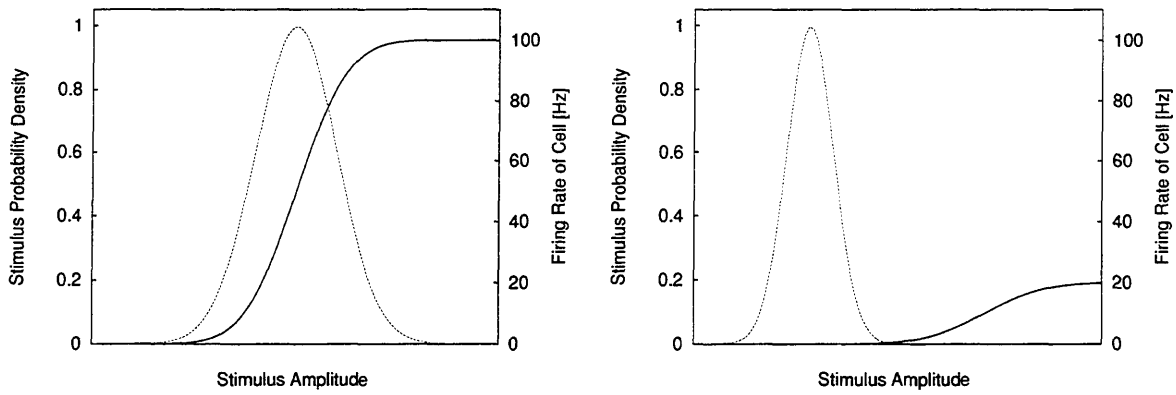


Figure 2.1: Suppose that stimuli are drawn from a hypothetical bell-shaped probability distribution of amplitudes (dotted lines). On the left, the neuron responds differentially to varying stimulus amplitudes by changing its firing rate (solid line). On the right, the neuron does not respond at all to most of the input range. In this latter case, the neuron clearly does not accurately represent the input.

By histogramming the frequency with which particular visual contrasts, spatial orientations, or spectral reflectances (colors) are encountered in a natural environment (Baddeley and Hancock, 1991; Atick, 1992; Ruderman, 1995), a model for the probability distribution of stimulus amplitudes reaching a photoreceptor in the retina can be developed. An efficient encoding of information would match the limited dynamic range of a neuron to the probability distribution of stimulus amplitudes, as

illustrated in fig. 2.1. Naturally occurring stimuli will exhibit spatiotemporal correlations, so an efficient and compact encoding of information would also eliminate superfluous correlations. Using concepts borrowed from statistical mechanics, it is possible to define the 'best' encoding of a continuous, time-varying input variable in the firing rate of a neuron. But even before insisting that a neuron be optimal, the central and real question is: does the firing rate encode the stimulus *at all*? Suppose that one had the ability to change the subcellular parameters that govern a neuron's behavior: even minor random tampering with these parameters could well result in a neuron whose firing rate no longer represents any stimulus information whatsoever.

The mathematical framework we will be working with is a generalization of the Hodgkin-Huxley equations that describe the generation of action potentials in the squid giant axon (Hodgkin and Huxley, 1952). The voltage V of the neuron evolves as

$$C\frac{dV}{dt} = \sum_i \bar{g}_i \, m_i^{p_i} \, h_i^{q_i} \, (E_i - V), \tag{2.1}$$

where C is the membrane capacitance, $\bar{g}_i$ is the peak value of the i-th conductance, p_i and q_i are integers, and E_i are the ion-specific reversal potentials. The variables m_i and h_i describe the activation or inactivation of voltage-dependent conductances, which change the current flowing across the membrane. These variables obey first order kinetics of the type

$$\tau(V)\frac{dm}{dt} = m_\infty(V) - m, \tag{2.2}$$

where $\tau(V)$ is a voltage-dependent time constant and $m_\infty(V)$ denotes the steady state activation when the voltage is clamped to V. Most often, the steady state activation functions $m_\infty(V)$ and $h_\infty(V)$ are S-shaped or reverse S-shaped functions of voltage. Together, the equations for V, m_i, and h_i form coupled set of nonlinear differential equations.

The challenge is to let the neuron 'develop' through experience, adapting its parameters so that it actually forms a representation of incoming stimuli in its response. Adaptation in highly nonlinear systems is, of course, a nontrivial matter.

This chapter is organized into two separate, but interdigitating tracks for readers with different interests in the subject. Each 'track' is fairly self-contained. Starred sections contain technical material that can be omitted on a first reading.

Track 1 is intended for readers who are familiar with the subject of 'Hebbian' learning and want more of an intuitive understanding of the non-Hebbian adaptation mechanisms proposed in this chapter. The possible biophysical substrates for adapting to the statistics of natural stimuli form one recurring theme. This track also proposes some simple experimental tests of the theory.

Track 1

$$2.2 \rightarrow 2.5 \rightarrow 2.7 \rightarrow 2.9 \rightarrow 2.12$$

2.2. At the microscopic level, the stochastic behavior of ion channels opening and closing in the cell membrane of the neuron gives rise to the macroscopic behavior of the Hodgkin-Huxley equations. The potential substrates for ion-channel specific modification of conductance properties are explored.

2.5. Investigation of possible supervised and unsupervised learning rules that change the parameters of voltage-dependent conductances. These learning or adaptation rules are based on functions that the membrane biophysics computes naturally, as explored in section 2.2.

2.7. Specific adaptation rules that match a neuron's range of firing rates to the range of inputs it receives are developed.

2.9. Numerical simulations using a simplified model neuron are used to demonstrate some of the properties of the adaptation. The learning rule of section 2.7 is shown to lead to a uniform distribution of firing rates, so that each firing rate is used equally often. The dynamical evolution of the neuron subject to information maximization learning rules is studied.

2.12 To test the learning rules on a true spiking model, the Hodgkin-Huxley model of Connor *et al.* (1977) with an additional adaptation current is implemented in a two-compartment model. Some experimental predictions are made.

Track 2 covers the more formal approach to information maximization and is intended for readers with an interest in nonlinear dynamical systems, or adaptation and learning algorithms in general.

Track 2

$$2.3 \rightarrow 2.4 \rightarrow 2.6 \rightarrow 2.7 \rightarrow 2.10 \rightarrow 2.13$$

2.3. The concepts of mutual information and entropy are used to provide a theoretical foundation for optimizing a neuron's firing rate response to input.

2.4. Derivation of a stochastic learning rule to maximize the mutual information or entropy by changing the parameters of the Hodgkin-Huxley equations (eqs. 2.1 and 2.2).

2.6. We map the information maximization problem onto an equivalent problem in the standard artificial neural network literature by ignoring the the nonlinear feedback implicit in the Hodgkin-Huxley equations. This corresponds to a zeroth-order model of a neuron. The resulting learning rules for changing the parameters of the Hodgkin-Huxley equations are shown to be horrendously complicated and biophysically implausible.

2.7. We show how nonlinear feedback and the physical constraint of charge conservation lead to much simpler learning rules. In this first-order model, we consider the mean firing rate of a neuron in response to a steady state input, ignoring any transient dynamics.

2.10. The mean firing rate model is replaced by a a periodically spiking Hodgkin-Huxley neuron. Learning rules are extended to cover periodic inputs to the adaptation mechanism for the model parameters.

2.13. Not only will information maximization match the amplitude of the neuron's response to the underlying statistics of the input, it will also remove statistical correlations in the input as a function of time. We here treat fully time-varying input stimuli.

2.2 The Kinetics of Ion Channels as a Substrate for Adaptation

At the microscopic level, sensation, perception, action, and thought are mediated by the flicker of ion channels switching between open and closed states in the cell membrane of nerve cells. The mammalian central nervous system possesses a panoply of diverse ion channel types that allow selected species of ions to cross the membrane. One major class of channel proteins carries a characteristic sequence of positively charged amino-acid residues along one of six transmembrane segments in each domain of the ion channel protein: called voltage-gated channels, these proteins change their conformation under the application of a voltage across the membrane, changing between open states that allow current to flow and closed states. These voltage-gated channels are the basis for the nerve impulse and also help shape the subthreshold response to stimuli.

The properties of voltage-gated channels will change if the channel protein is phosphorylated, the intracellular calcium concentration is raised, or the cell expresses additional 'subunits' that associate themselves with the main channel protein, or the channel protein interacts with a neuromodulatory enzyme. Molecular biology techniques have revealed more than ten different genes coding for various subunits of calcium channels, allowing for a combinatorial number of functionally different channels (Hofmann *et al.*, 1994). These channel types can be differentially expressed in development. Since all of the changes in ion channel properties can occur naturally

in vivo, without artificial intervention by experiment, the tools required to adaptively compress information from the senses are thus available to single neurons at the subcellular level.

To explore how a nerve cell might change the properties or membrane density of voltage-dependent ion channels in response to input, we must first establish how the occupancy of channel states and the transitions between states relate to the voltage across the cell membrane. These relationships will indicate how ion channels could initiate a process of self-modification. We begin with a description of a simple model of a voltage-dependent ion channel based on equilibrium physics:

The simplest and most generally applicable model of a voltage-gated ion channel has a single open and closed state. A dipole moment or gating charge on the S4 transmembrane segment in each domain of the ion channel protein biases the protein to flip between states at different voltages. Under the influence of a transmembrane voltage, the charged protein residues move across the membrane, changing the tertiary structure of the ion channel protein to allow selected species of ions to pass through the pore region linking the four protein domains (Yang and Horn, 1995; French *et al.*, 1996; Seoh *et al.*, 1996; Yang *et al.*, 1996; Aggarwal and MacKinnon, 1996; Larsson *et al.*, 1996).

The probability of an ion channel being in any given state at physical equilibrium is given by its Gibbs free energy, which is a function of the internal electrostatic and van der Waals interactions between protein residues, the entropy of protein folding, and the external electric field, if present. For state i with free energy E_i, this probability

is given by the ratio of exponential terms:

$$p(E_i) = \frac{\exp(-E_i/kT)}{\sum_j \exp(-E_j/kT)}, \tag{2.3}$$

where T is the absolute temperature (in Kelvin), k is Boltzmann's constant, and the sum in the denominator is over all accessible states.

Following the nomenclature of Almers (1978) and Hille (1992), the energy difference between two states can be divided into a conformational energy change W of the protein and an energy change $-z_g V$ due to the movement of an effective gating charge z_g across the membrane potential V. The effective charge depends on the position of positive charges in the S4 segment, the distance these move through the membrane, and the structure of the electric field within the membrane.

In some instances, a literal interpretation of the gating charge as the true charge is possible. Evidence from cysteine substitution experiments in Na^+ and K^+ ion channels suggests that the S4 segment traverses a narrowing of the membrane (reviewed in Goldstein, 1996), such that most of the translocated charge moves entirely across the membrane. In such a case, the effective gating charge is nearly equal to the true gating charge. In corroboration, there are relatively few negative residues available elsewhere to counteract the positively charged residues on the S4 segment within the membrane, pointing once again to a membrane narrowing. The transition between open and closed state could also correspond to a conformational change in the secondary structure of the S4 segment (such as a transition from α-helix to β-sheet); while such a change still implies the movement of charges, the dipole moment will

change, leading to higher order corrections in the membrane voltage V to the electrostatic energy. Whether such a conformational change underlies the gating of channels or whether the S4 α-helix translates rigidly under an applied voltage is still subject to debate. For simplicity, we will pretend that the latter scenario holds true.

We plot the two energy states of the simple ion channel model in a one-dimensional 'reaction-coordinate' system (fig. 2.2). The reaction coordinate is simply a measure

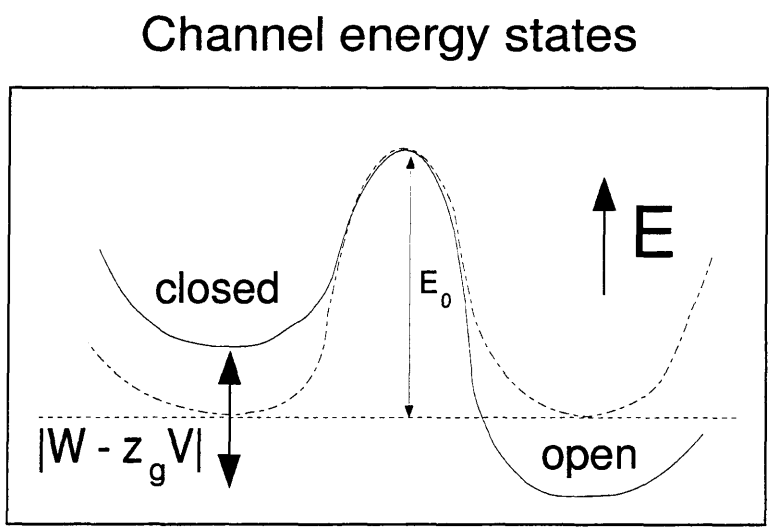


Figure 2.2: The simplest model of an voltage-dependent ion channel has two states, open and closed. Since such channels carry a charge, changing the voltage shifts the relative energies of the two states, biasing the probability for the channel to dwell in the open versus the closed state.

of the protein's folded shape as it changes from closed to open state or vice versa. When the two states are populated with equal probability (dashed-dotted line), the equilibrium rate r of transitions from the closed to open state is simply the Kramer's escape rate across the activation energy barrier

$$r \propto \sqrt{E''(x_{\min})|E''(x_{\max})|}\ \exp(-E_0/kT),$$

where x is the reaction coordinate and $E''(x_{\min})$ and $|E''(x_{\max})|$ are the curvatures of the solid curve in fig. 2.2 at the minimum and maximum.

Suppose that the conformational energy difference W between the two states is constant and independent of voltage and that the electric field is linear across the membrane. As in a see-saw, changing the transmembrane potential V shifts the Gibbs free energies of the open and closed states by opposite but equal amounts. Note that we will *assume* the energy barrier to be symmetric. The total energy change is $\Delta E = W - z_g V$. The forward transition rate $\alpha(V)$ from closed to open state will be:

$$\alpha(V) = r \exp\left[-\left(\frac{(W - z_g V)}{2kT}\right)\right], \tag{2.4}$$

and the complementary backward transition rate $\beta(V)$ will be

$$\beta(V) = r \exp\left[+\left(\frac{(W - z_g V)}{2kT}\right)\right]. \tag{2.5}$$

Here the rate r denotes the Kramer's escape rate when the two states are equally balanced.

At the macroscopic scale of ion flow across thousands of channels, $m(V)$ represents the fraction of channels of one type that are open at a given time. The quantity $m(V)$ obeys the first-order differential equation

$$\frac{dm}{dt} = \alpha(V)(1 - m) - \beta(V)m. \tag{2.6}$$

In the Hodgkin-Huxley equations (Hodgkin and Huxley, 1952), this is the equation for a single gating particle.

For each steady state transmembrane voltage V, the channel will be in the open state with a probability $m_\infty(V)$ as a function of voltage and in the closed state with probability $(1 - m_\infty(V))$. As long as the ion channel possesses only two states that are in energetic equilibrium, $m_\infty(V)$ must be an S-shaped function of voltage given by the Boltzmann equation:

$$m_\infty(V) = 1/\{1 + \exp[-s(V - V_{\frac{1}{2}})]\},$$

where $s = z_g/(kT)$ and $V_{\frac{1}{2}} = W/z_g$. This is simply the solution of eq. 2.6 when $dm/dt = 0$. $V_{\frac{1}{2}}$ is the midpoint voltage of half-activation, and s is the slope of the S-shaped function at its maximum inflection.

We can, of course, rewrite eq. 2.6 in terms of $m_\infty(V)$ as

$$\tau(V)\frac{dm}{dt} = m_\infty(V) - m, \tag{2.7}$$

with $\tau(V) = [\alpha(V) + \beta(V)]^{-1}$.

While the steady state activation curve $m_\infty(V)$ must obey eq. 2.3, the forward rate $\alpha(V)$ and backward rate $\beta(V)$ are by no means unique. As a consequence, ionic channels can have the same steady-state properties, but differ in their fluctuation properties and time constants of activation. To remove this degeneracy, we must specify both functions, $\alpha(V)$ and $\beta(V)$, precisely.

Even at steady state, individual channels will make frequent transitions between open and closed states at a rate proportional to $\left[\frac{d}{dV}m_\infty(V)\right]^\gamma$, where the exponent γ is typically 1/2 or 1, depending on the model for the kinetics, as we shall show below. The frequency of transitions follows a bell-shaped curve of voltage, with the highest frequency of transitions typically occurring when the channel is balanced between open and closed states at $m_\infty(V) = 1/2$.

The number of channel openings f_O (closings f_C) per unit time at equilibrium is simply the product of the opening (closing) rate with the number of closed (open) channels given by $m_\infty(V)$. This can be shown to be

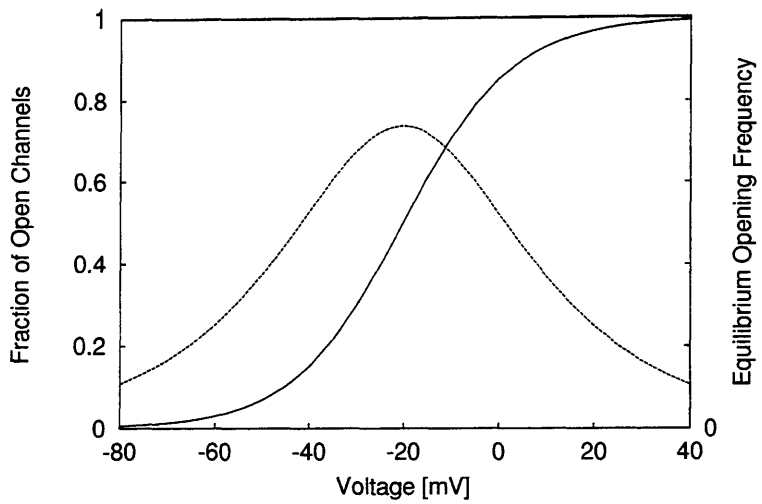
$$\begin{aligned} f_O, f_C = \; & r\Big\{\exp[s/2\,(V - V_{\frac{1}{2}})] + \exp[-s/2\,(V - V_{\frac{1}{2}})]\Big\}^{-1} \\ = \; & \left[\tfrac{r^2}{s}\,\tfrac{d}{dV}m_\infty(V)\right]^{\frac{1}{2}}, \end{aligned}$$

which peaks at $V = V_{\frac{1}{2}}$ (centered at the inflection point of the S-shaped function $m_\infty(V)$), falling off to either side in the fashion of a bell-shaped curve as in fig. 2.3.

We point out several other quantities which scale as powers of the first derivative of $m_\infty(V)$. From eq. 2.7, the voltage-dependent time constant is

$$\tau(V) = \frac{1}{\alpha(V) + \beta(V)} = \left[\frac{1}{s}\,\frac{d}{dV}m_\infty(V)\right]^{\frac{1}{2}}.$$

Suppose the voltage fluctuates by a small amount ΔV. The fluctuation-dissipation theorem predicts that this voltage fluctuation will decay with the time constant $\tau(V)$. The additional charge $\Delta Q(V)$ that flows through the channel in that time will scale



The bell-shaped curve represents the number of channel openings (closings) at equilibrium, while the sigmoid represents the steady-state activation curve of the conductance, corresponding to the average fraction of channels in the open state. The bell-shaped curve is proportional to the square root of the sigmoid's derivative.

Figure 2.3:

as:

$$\Delta Q(V) \propto \left[\frac{d}{dV} m_\infty(V)\right] \Delta V \, \tau(V) = \frac{1}{\sqrt{s}} \left[\frac{d}{dV} m_\infty(V)\right]^{\frac{3}{2}}.$$

If N channels are present, each with a probability $m_\infty(V)$ of being open and $1 - m_\infty(V)$ of being closed, then the number of open channels will be binomially distributed with variance $N m_\infty(V)(1 - m_\infty(V))$ (Johnston and Wu, 1995). Thus the current noise through N summed channels of a particular type will be

$$\sigma_I^2 = N\gamma^2 (V - E_{\text{rev}})^2 \frac{\alpha(V)\beta(V)}{(\alpha(V) + \beta(V))^2},$$

where the constant of proportionality $N\gamma^2$ is the product of the number of channels and the squared single-channel peak conductance. The current noise, therefore, obeys the scaling

$$\sigma_I^2 \propto (E_{\text{rev}} - V)^2 \left[\frac{d}{dV} m_\infty(V)\right]^1.$$

Thus, we have the general scaling:

$$f_O, f_C, \tau(V), \Delta Q(V), \sigma_I^2(V) \propto \left[\frac{d}{dV} m_\infty(V)\right]^\gamma,$$

where γ is some power ($\gamma = 1/2, 1, 3/2$) .

To obtain a scaling of f_O and f_C with an exponent of $\gamma = 1$ and without changing the steady state activation function, it is straightforward to show that we need

$$\alpha(V) = r \left\{1 + \exp\left[-\left(\frac{(W - z_g V)}{kT}\right)\right]\right\}^{-1} = r\, m_\infty(V),$$
$$\beta(V) = r \left\{1 + \exp\left[\left(\frac{(W - z_g V)}{kT}\right)\right]\right\}^{-1} = r\,(1 - m_\infty(V)).$$

(Just solve $m_\infty(V) = \alpha(V)/[\alpha(V) + \beta(V)]$ and $f_C = m_\infty(V)\beta(V) = r m_\infty(V)[1 - m_\infty(V)]$ using the assumption of Boltzmann statistics. The second equation is simply the statement that the forward rate should be some multiple of the derivative of $m_\infty(V)$.) In this case, the forward rate is proportional to the steady-state fraction of open channels. Physically, this corresponds to rate saturation in eq. 2.6. The time

constant $\tau(V)$ is constant!

$$\tau(V) = \frac{1}{\alpha(V) + \beta(V)} = \Big\{ r\big[m_\infty(V) + (1 - m_\infty(V))\big] \Big\}^{-1} = \frac{1}{r}$$

Also, the current noise remains

$$\sigma_I^2(V) \propto \left[\frac{d}{dV} m_\infty(V) \right]^1 .$$

Hence, the steady state properties of this model are indistinguishable from the previous model.

In practice, many *ad hoc* empirical models are used to describe channel kinetics. For instance, one standard parametrization of the forward and backward rates of ion channel opening is:

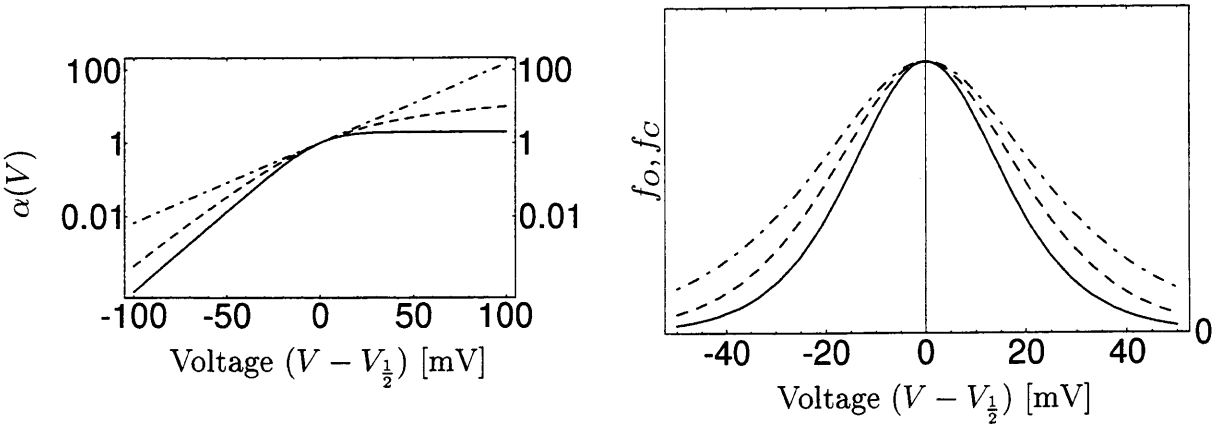
$$\alpha(V) = r\,(V - V_{\frac{1}{2}}) \left\{ 1 - \exp\Big[-s\,(V - V_{\frac{1}{2}}) \Big] \right\}^{-1}$$

$$\beta(V) = -\,r\,(V - V_{\frac{1}{2}}) \left\{ 1 - \exp\Big[+s\,(V - V_{\frac{1}{2}}) \Big] \right\}^{-1}$$

as used by Mainen *et al.*, 1995. The steady-state fraction $m_\infty(V)$ of open channels is still given by the Boltzmann equation; only the kinetics change. Hodgkin and Huxley, 1952 originally proposed this phenomenological function to describe the saturation in the forward rates of potassium and sodium gating particles at higher transmembrane voltages, as illustrated in fig. 2.4a. Rate saturation arises when multiple states of the ion channel exist (as in eq. 2.3), each one of which must be traversed for the ion

channel to move into the open configuration. The phenomenological kinetic model proposed by Hodgkin and Huxley leads to a steady-state frequency of channel openings and closings that is more closely proportional to $d/dV\left[m_\infty(V)\right]$ in the tails than the original kinetic scheme:

$$\begin{aligned} f_O, f_C &\sim \left|V - V_{\frac{1}{2}}\right| \left[\frac{d}{dV} m_\infty(V)\right]^1 \qquad \text{as } \left|V - V_{\frac{1}{2}}\right| \to \infty \\ &\sim \left[\frac{d}{dV} m_\infty(V)\right]^{\frac{1}{2}} \qquad \text{as } \left|V - V_{\frac{1}{2}}\right| \to 0. \end{aligned}$$



(a) (b)

Figure 2.4: Comparison of forward rates and steady-state transition frequencies for three models of ion channel kinetics. Solid line: voltage-independent time constant; dashed line: two-state energy-barrier model; dot-dashed line: Hodgkin-Huxley phenomenological model.

Now that several possible functional relationships between the voltage and the

transition rate between different states of voltage-dependent ion channels have been enumerated, we turn to the question of how the density of channels or the channel characteristics could be regulated in a "voltage-dependent" manner.

Arrayed against the backdrop of ion channels is an armada of second messenger enzymes involved in regulating ion channel activity. These enzymes can catalyze successive enzymes in multiple-step reactions known as cascades. A common element to most second-messenger cascades is a class of heterotrimeric GTP-binding proteins, also called G-proteins, known to modulate ion channel function both directly (Ma *et al.*, 1994; Herlitze *et al.*, 1996) and indirectly by activating protein kinases that phosphorylate channel proteins or initiating gene transcription of new channel proteins.

In general, channel proteins will go through several intermediate, meta-stable states en route from the more stable open and closed states. We propose that the channel is subject to modification (e.g., by being phosphorylated) as it passes through these intermediate states, as illustrated in the cartoon schematic of fig. 2.5. Alternatively, these intermediate states can interact with second messengers, such as a G-protein, to elicit a cascade leading eventually to gene expression of the channel protein and a change in the channel density.

For the sake of argument, suppose the metastable state O^* is subject to modification. If the dwell time in the state O^* is τ_{O^*}, the "concentration" of channels in this

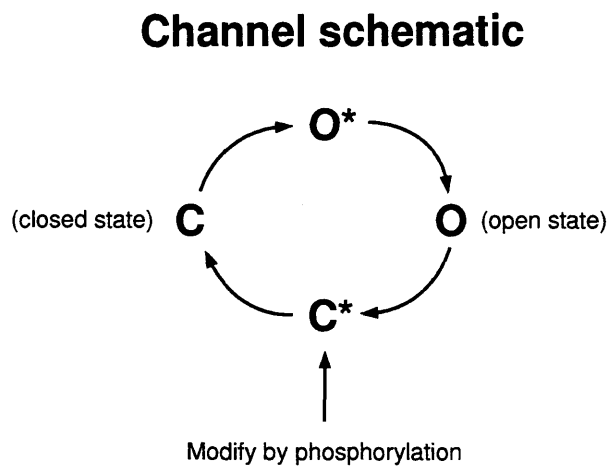


Figure 2.5: The channel passes through several short-lived meta-stable states in switching between closed open states. These intermediate meta-stable states are indicated by an asterisk on the diagram above. Assume that no modification occurs while the channel rests in the stable open or closed state. If there is only a finite probability per unit time that these transient meta-stable states are phosphorylated, then the degree of phosphorylation depends on the frequency of channel state transitions. As illustrated in fig. 2.3, the transition frequency is a bell-shaped curve centered along the voltage axis at the inflection point of the conductance activation curve.

state at equilibrium is:

$$\begin{aligned}[O^*] &= f_O \, \tau_{O^*} \\ &= \left[\frac{r^2}{2\beta} \frac{d}{dV} m_\infty(V)\right]^{\frac{1}{2}} \tau_{O^*}.\end{aligned}$$

The rate of modification will thus depend on the frequency of channel openings/closings.

Most second-messenger systems that have been explored to date are initiated by ligands binding to receptors, such as an odorant binding to a β-adrenergic receptor in a chemosensor or a photon "binding" to and isomerizing a rhodopsin molecule in the photoreceptor (Lamb and Pugh Jr., 1992a, 1992b). Here, the rate of G-protein activation will be directly proportional to the amount of bound ligand.

By proposing that voltage-gated ion channels initiate second-messenger cascades as a form of self-regulation. we assume that the voltage, rather than a ligand concentration, sets the rate of G-protein activation. The second-messenger system can inherit its voltage-dependency from that of the ion channels in one of two ways: if the G-protein acts as a *state* sensor, the activation rate will be an S-shaped function of voltage; if instead it acts as a *transition* sensor, the rate will be a bell-shaped function of voltage. A state sensor corresponds to a single-step G-protein activation process in which the G-protein complex associates with only one of the two states of the channel. A transition sensor requires an additional, intermediate step, which could consist of either of the following:

1. A short-lived, metastable state O* that lies between the stable open and closed states. If the G-protein binds only to the metastable state, then activation of the second-messenger cascade will be proportional to the "concentration" of channels in that state, as given by the equation above.

2. Since both the α subunit and $\beta\gamma$ complex of the G-protein can act as catalyzers for second-messenger, the unbinding of the $\beta\gamma$ complex could function as a transition sensor, as illustrated in fig. 2.6).

Figure 2.6 is a purely hypothetical 'Rube-Goldberg'-type mechanism that involves a second-messenger cascade initiated by a GTP-binding protein. Four homologous domains form the α-subunit of an ion channel protein, at the center of which is the pore or P-region which allows selected species of ions to cross the membrane. The $\beta\gamma$ complex of certain heterotrimeric GTP-binding protein binds to the calcium ion

channel protein (see, for example, De Waard *et al.*, 1997). One direct, known effect of this binding is to shift the voltage-dependency of the calcium channel being in the open state.

By postulating that indirect effects of G-protein binding are also possible, we now show how the G-protein could conceivably act as both a state *and* a transition sensor. Figure 2.6 shows the $\beta\gamma$ complex associating with the S4 transmembrane segment of one of the ion channel domains. Only when the G-protein is bound to the ion channel protein can GTP bind to the α subunit of the G-protein, activating the catalytic properties of the α subunit. The activated α subunit acts as a state sensor, since it is produced at a rate proportional to the occupancy of the open state.

Under application of a transmembrane voltage, the S4 segment is visualized as moving down across the membrane, carrying with it the gating charge. This conformational change closes the pore. If this event also causes the G-protein's $\beta\gamma$ components to simultaneously dissociate from the channel protein, then the rate at which this occurs is proportional to the rate of ion channel state transitions. If the $\beta\gamma$ has catalytic properties when it is independent of the α subunit, then this component acts as a transition sensor.

Note that the products of second-messenger cascades depend on the law of mass action, so that the rate of production or concentration of the effector, or end product of the cascade, can be proportional to $m_\infty(V)$ or $\left[\frac{d}{dV}m_\infty(V)\right]^\gamma$ raised to some power.

For instance, a G-protein mediated second messenger cascade can take

$$\left[\frac{d}{dV}m_\infty(V)\right]^{\frac{1}{n}} \Longrightarrow \frac{d}{dV}m_\infty(V), \qquad \text{where } n \text{ is integer.}$$

In particular, we are interested in the case $n = 2$, since this corresponds to the equilibrium model of a voltage-dependent ion channel with two states. Suppose the activated form of the G-protein, which is produced at a rate proportional to [O*], or the ion channel transition frequency, gives rise to two products, A and B. Suppose that the effector E, which could be a protein kinase, transcriptional regulatory protein, depends on A and B as follows:

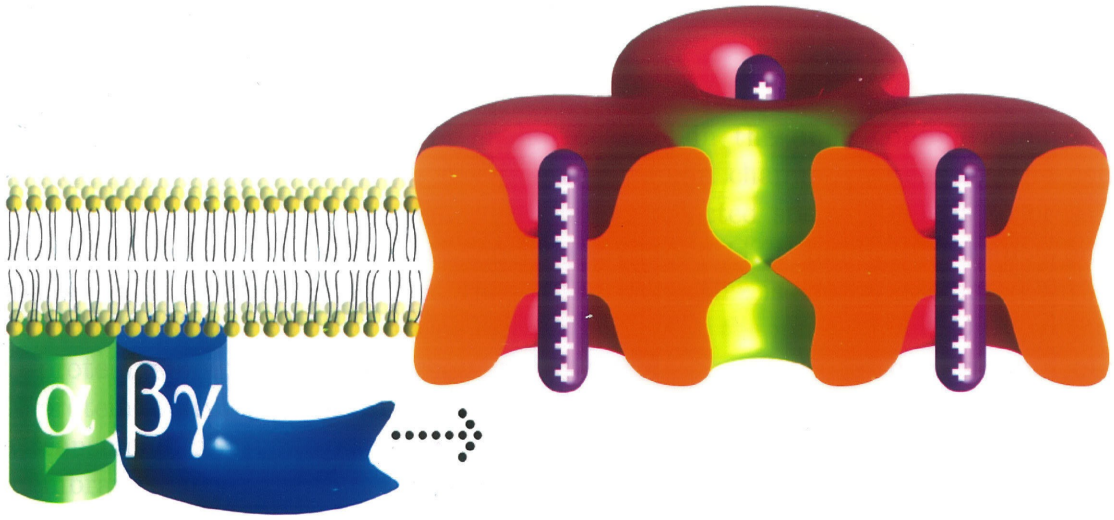
$$A + B + E \longrightarrow E^*.$$

In this case, the rate of E^* production is proportional to the product $[A][B]$ of the concentrations of $[A]$ and $[B]$. Assuming that $[A]$ and $[B]$ quickly reach equilibrium values, the rate of effector activation will be

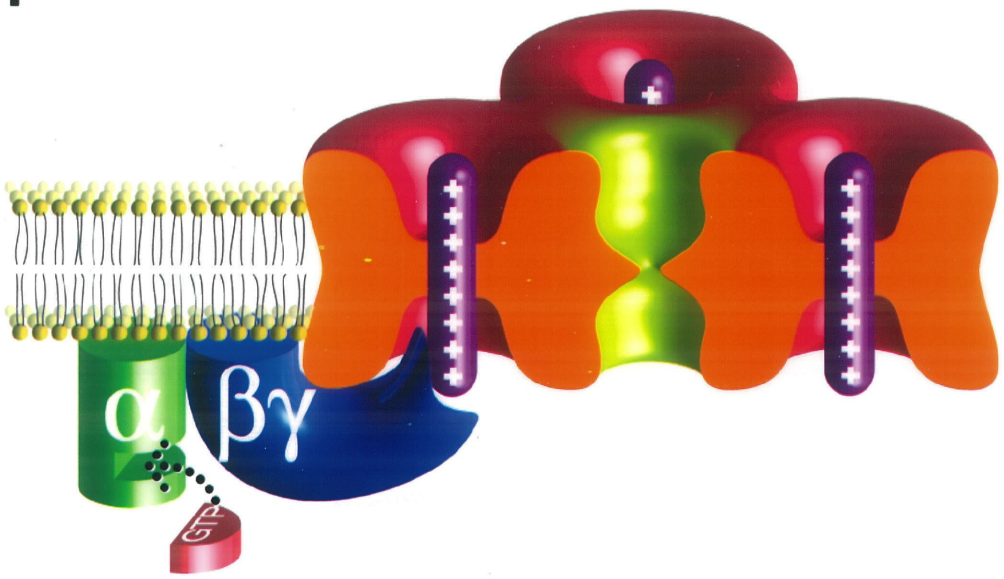
$$\begin{aligned} E^*(t) &\sim [A][B] \sim f_O^2 \\ &= \frac{d}{dV}m_\infty(V). \end{aligned}$$

(Here we assume that $E^*(t)$ is not inactivated, recycled, or degraded quickly and that E is present in abundant quantity.)

If each voltage-gated channel is associated with its own G-protein-mediated mech-



G-protein



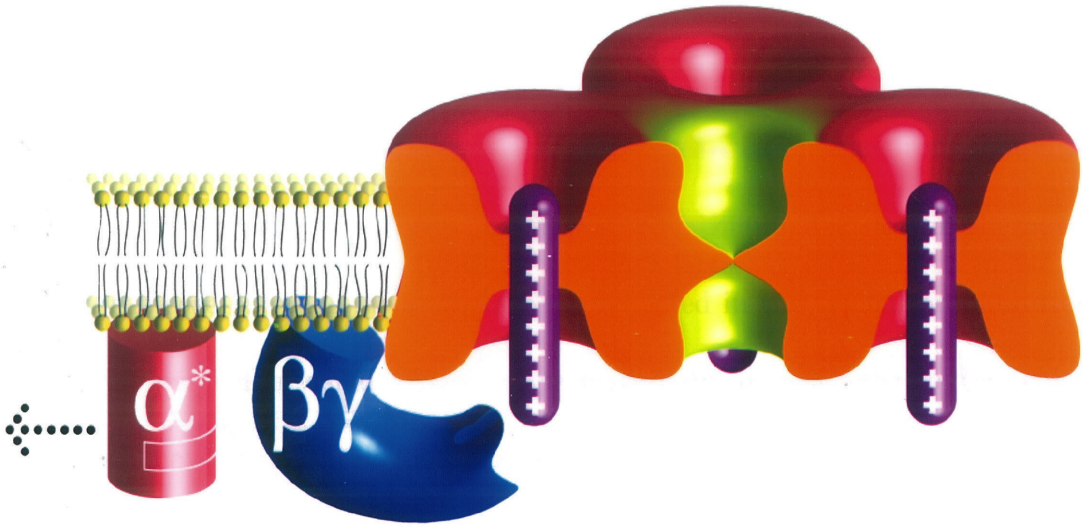


Figure 2.6: Schematic sketch of an ion channel in the open configuration. Four homologous domains form the α-subunit of an ion channel protein, at the center of which is the pore or P-region which allows selected species of ions to cross the membrane. The $\beta\gamma$ complex of the heterotrimeric GTP-binding protein typically binds to the ion channel protein. Here, the $\beta\gamma$ complex is shown associated with the S4 transmembrane segment of one of the domains of the ion channel. Only when the G-protein is bound to the ion channel protein can GTP bind to the α subunit of the G-protein, activating the catalytic properties of the α subunit. Under application of a transmembrane voltage, the S4 segment is visualized as moving down across the membrane, carrying with it the gating charge. This conformational change closes the pore. If the G-protein's $\beta\gamma$ complex simultaneously dissociates, then the G-protein can act as a both a state and a transition sensor through its two different components: the α subunit and the $\beta\gamma$ subunit complex (see text).

anism for self-regulation, then the functions $m_\infty(V)$ and $\left[\frac{d}{dV}m_\infty(V)\right]^\gamma$ for different voltage-gated ion channel types form a set of computational primitives. Using these primitives, we will show that biophysically plausible 'learning rules' exist by which a neuron could maximize the information about stimuli in its firing rate. The next section lays the theoretical foundation for information maximization. The reader who prefers a somewhat more heuristic approach can skip to section 2.5.

2.3 Information Maximization

Information from the senses must be compressed into the limited range of firing rates in spiking nerve cells. An efficient representation of information in neurons should take advantage of the regularity and scale invariances of stimulus features in the natural world. In the case of vision, this regularity is reflected in the typical probabilities of encountering particular visual contrasts, spatial orientations, or wavelengths (colors), all of which have been catalogued (Atick, 1992; Ruderman, 1995). Given these

probabilities, an optimized neural code would eliminate redundant features whilst devoting increased representation to commonly encountered features.

Using concepts from statistical mechanics, we can quantitatively define what it means for neural firing rate code to be 'optimal': Optimal noiseless transmission of information in a single nerve cell occurs when the response is matched to the statistics of naturally occurring stimuli, such that all firing rates are used with equal probability. If noise is present, the probability of firing rates should be weighted inversely proportionally to the standard deviation of the output noise.

Hodgkin-Huxley models of real neurons can exhibit complex behaviors on several timescales, such as firing patterns consisting of "bursts"—sequences of multiple spikes interspersed with periods of silence. We will, however, focus on models of regularly spiking cells, which adapt to a sustained stimulus by spiking periodically. To quantify how much information the time-averaged firing rate f of a regularly spiking neuron carries about a stimulus variable x, we use the definition of the mutual information $\mathcal{I}(f;x)$ between the stimulus x and the firing rate f in terms of entropy functions of the (continuous) probability distributions of f and x:

$$\mathcal{I}(f;x) = S(f) - \langle S(f|x)\rangle_x, \tag{2.8}$$

such that the mutual information is the difference between two entropies: the first is the unconditional firing rate entropy $S(f)$ of the probability distribution $p(f)$; the second is the average conditional firing rate entropy $\langle S(f|x)\rangle$.

We define these entropy functions and motivate their relationship to the mutual

information below:

If $p(f)$ is the probability, given the set of all stimuli, of a firing rate f, then the entropy $S(f)$ of the firing rates is defined as

$$S(f) = -\int \ln [p(f)]\, p(f)\, df, \tag{2.9}$$

where the integral is taken over all firing rates. In rough terms, the entropy measures how broad a single-peaked probability distribution of firing rates is. A broad distribution of firing rates is, of course, more suitable for representing many different stimulus amplitudes.

On the other hand, to optimize information transfer through the neuron, a stimulus x should reliably produce a firing rate f with little variation in f; therefore, the conditional probability distribution $p(f|x)$ given a stimulus x should be highly peaked. For a particular stimulus x, the distribution of firing rates in response to that stimulus will lead to a conditional entropy $S(f|x)$:

$$S(f|x) = -\int \ln [p(f|x)]\, p(f|x)\, df$$

Following Stein (1967), we can write down a simple approximate expression for the conditional entropy. As a consequence of the Central Limit Theorem, the distribution of firing rates for a particular stimulus x approaches a Gaussian distribution over long

times. Thus $S(f|x)$ will tend to

$$S(f|x) \sim \frac{1}{2}\log\left[2\pi\, e\, \sigma_f^2(x)\right] \tag{2.10}$$
$$= S(f_{\text{Gaussian}}|x),$$

where $\sigma_f^2(x)$ is the variance in the firing rate in response to a stimulus x as estimated over a set time interval T. This expression is an upper bound to the true conditional entropy because of the inequality $S(f|x) \leq S(f_{\text{Gaussian}}|x)$, true for any continuous probability distribution with the same fixed variance $\sigma_f^2(x)$ as the Gaussian. The deviation from equality introduced by considering even heavily skewed distributions of firing rates is typically minor (Stemmler, 1996).

Of course, the conditional entropy for a *particular* stimulus is less relevant than the *average* conditional entropy. The average conditional entropy is a measure of overall reliability: how well, on average, can the stimulus x be determined simply by observing the response f? As will be explored in the chapter on stochastic resonance, the dynamic range of responses in neurons with thresholds can be expanded by injecting uncorrelated fluctuations or noise in the input current to the neuron. But the drawback of adding input noise is that it also increases the noise in the output. The mutual information $\mathcal{I}(f;x)$ is a quantitative measure of the trade-off between these two effects.

We will assume that the probability distribution of inputs $p(x)$ is fixed and continuous, and that f and x are related by a differentiable and invertible function. This invertible function maps the probability distribution $p(x)$ into the probability

distribution $p(f)$, such that the infinitesimal probability measures are equal:

$$p(f)\,df = p(x)\,dx. \tag{2.11}$$

The assumptions that $p(x)$ exists and that $f(x)$ is invertible are sufficient to guarantee the existence of eq. 2.8. Even as $f(x)$ changes (by adapting some of the internal parameters of the neuron, for instance), we will assume that the one-to-one function $f(x)$ always exists.

Noise, as measured by the variance $\sigma_f^2(x)$ of the response or firing rate of the neuron, will always be treated as being additive. In other words, the deterministic function $f(x)$ exists; corruption by noise is modeled by adding a term Δf to the value of $f(x)$, where Δf can be a function of f, and hence a function of x. Multiplicative noise, a much more difficult subject, will not be considered here.

Substituting eqs. 2.9 and 2.10 into eq. 2.8 yields the following lower bound on the mutual information:

$$\mathcal{I}_{\mathrm{LB}}(f;x) = -\int \ln\Big(p[f(x)]\,\sigma_f(x)\Big)\,p(x)\,dx - \ln(\sqrt{2\pi e}). \tag{2.12}$$

Using eq. 2.11, we can also go one step further and rewrite the mutual information

in terms of $p(x) = p(f)df/dx$:

$$\mathcal{I}(f;x) = S(f) - \langle S(f|x)\rangle_x$$
$$\mathcal{I}_{\rm LB}(f;x) = \int_{-\infty}^{\infty} p(x)\ \ln\left[\frac{1}{\sigma_f(x)}\left(\frac{df(x)}{dx}\right)\right] dx - \int p(x)\ln p(x)\, dx, \tag{2.13}$$

where we have dropped the constant term $-\ln(\sqrt{2\pi e})$. Note that this expression is explicitly invariant with respect to linear rescaling of $f \to \alpha f$. If $f(x)$ is any nonlinear function, then the mutual information is always less than the original entropy of the inputs: $\mathcal{I}(f;x) < S(x)$.

Now that the mutual information $\mathcal{I}(f;x)$ has been defined, we examine what the consequences are of optimizing the information transfer through a neuron. First, we determine which probability distribution of output firing rates maximizes the mutual information; second, we will find the function $f(x)$ that achieves this probability distribution of $p(f)$.

To find the optimal probability distribution of output firing rates, we resort to the calculus of variations. Note that any probability distribution $p(f)$ is subject to the constraint

$$\int p(f)\, df = 1\,.$$

Introducing a Lagrangian multiplier λ to enforce this constraint, we augment the mutual information $\mathcal{I}_{\rm LB}(f;x))$ in eq. 2.12 by a Lagrangian term and take the variational

derivative

$$\frac{\delta \mathcal{I}_{\mathrm{LB}}(f;x))}{\delta p(f)} = \frac{\delta}{\delta p(f)}\left(-\int p(f)\Big\{\lambda + \ln\left[\sigma_f\, p(f)\right]\Big\}\, df\right)$$
$$= -(\lambda + 1 + \ln[\sigma_f\, p(f)]). \tag{2.14}$$

To extremize the functional, set the variational to zero, resulting in

$$p(f) = \frac{\text{constant}}{\sigma_f}. \tag{2.15}$$

Since the second variational derivative is negative, this solution for $p(f)$ maximizes $\mathcal{I}_{\mathrm{LB}}(f;x)$ The optimal probability distribution of firing rates is inversely proportional at each point to the standard deviation of the firing rate. If low firing rates are more reliable than high firing rates, in that low firing rates are subject to less noise or variation, then the optimal strategy is to bias the output firing rate probability distribution away from high firing rates. If, instead, the variation in the measured firing rate is independent of the stimulus x, then the optimal firing rate probability distribution is flat over the entire range of possible rates: each firing rate is used with equal probability; among engineers, this is known as response histogram equalization.

Stein (1967) considered two specific cases of spiking neurons:

(1) Completely regular neurons that spike perfectly periodically. If the time window T over which spikes are counted is not synchronized with the spike train, then the measurement of the firing rate will be subject to some error, or variability, which is reflected in a constant variance σ_f^2:

$$\sigma_f = \frac{1}{2\pi}$$

If we drop terms that do not depend on the function $f(x)$, the mutual information from eq. 2.13 scales as

$$\mathcal{I}_{\mathrm{LB}}(f;x) \sim \int \ln\left(\frac{df}{dx}\right) p(x)\, dx.$$

This leads to an optimal probability distribution

$$p_{\mathrm{optimal}}(f) = \text{ constant} \qquad f \in [f_{min}, f_{max}]$$

that is flat. Equivalently, additive and independent noise can be superimposed on the fundamental periodic firing behavior of the neuron at the output stage. If the noise is additive, it is not a function of the stimulus x, and thus all firing rates are equally "noisy," implying that $\sigma_f = \text{constant}$. The prediction of a flat output probability distribution has been tested experimentally in the blowfly by Laughlin (Laughlin, 1981), who matched the statistics of naturally occurring visual contrasts to the response amplitudes of the fly's (non-spiking) large monopolar cell (LMC). The response amplitude in the fly LMC is the amount of neurotransmitter released in response to a stimulus—this is the functional equivalent of the firing rate in a spiking neuron. Laughlin's implicit assumption

was that all response amplitudes were equally reliable.

(2) Randomly spiking neurons, in which the spike timing is described by a Poisson process with a mean firing rate. The variance of the firing rate behaves as

$$\sigma_f{}^2 = f$$

As a consequence, the mutual information scales as

$$\mathcal{I}_{\mathrm{LB}}(f;x) \sim \int \ln\left(\frac{d\sqrt{f}}{dx}\right) p(x)\, dx,$$

where, once again, we have dropped the term that does not depend on $f(x)$.

$$p_{\mathrm{optimal}}(f) = \frac{\mathrm{constant}}{\sqrt{f}} \qquad f \in [f_{min}, f_{max}]$$

Low firing rate responses are more typical than high firing rates. Since much of the neurophysiological evidence supports the Poisson hypothesis of neuronal spike timing, we mention one of the implications of an optimal rate code if the spike timing is Poisson: most often, a typical stimulus (particularly one from the natural environment or one drawn by random chance) will drive a particular cortical nerve cell only weakly. In general, only a few neurons in a population of cells will respond vigorously to any particular stimulus, such as the orientation of a bar in the visual field. However, this consequence of maximizing the information in single cells is *not* evidence for the sparse coding

hypothesis (Olshausen and Field, 1996), which predicts that the information in a population code is typically contained in the firing of a restricted and small subset of cells of the entire population. Sparse coding and information maximization in single neurons are, of course, not inconsistent with each other, since they make a similar prediction about the distribution of firing rates.

If the output probability distribution of firing rates is uniform over a fixed interval, what is the corresponding functional relationship between the stimulus x and the firing rate? To answer this, we show that for any fixed, differentiable probability distribution function $p(x)$, the entropy of the monotonically increasing and bounded function $f(x)$ will be maximal when $f(x) \propto \int_{-\infty}^{x} p(x')\,dx'$. Rewriting

$$S[f(x)] = -\int p[f(x)]\,\ln\Big\{p[f(x)]\Big\}\,df(x)$$

in terms of $p(x)$, we get

$$= -\int p(x)\,\ln[p(x)]\,dx + \int p(x)\ln\left|\frac{df(x)}{dx}\right|dx.$$

Note that

$$\int_{-\infty}^{\infty}\frac{d}{dx}f(x)\,dx = f_{max} - f_{min},$$

since $f(x)$ increases monotonically and is bounded above and below. This will be our constraint. Taking the variational derivative with respect to the derivative of $f(x)$,

we solve for

$$\frac{\delta}{\delta f'(x)}\left\{-\int\Big[p(x)\ \ln[p(x)]-p(x)\ \ln[f'(x)]+\lambda f'(x)\Big]\,dx\right\}=0.$$

The variational solution is simply

$$\frac{p(x)}{f'(x)}=\lambda$$

which leads to

$$f(x)=(f_{max}-f_{min})\int_{-\infty}^{x}p(x')\,dx'+f_{min}.$$

In other words, the firing rate that maximizes the entropy matches the cumulative distribution function of the inputs (i.e., the definite integral of the probability distribution). If the noise is constant and independent of the stimulus, then maximizing the entropy of the firing rates is equivalent maximizing the mutual information between the firing rate and the stimulus ensemble.

More generally, the firing rate response that optimizes the mutual information is:

$$f(x)=c_1\int_{-\infty}^{x}\sigma_f(x')p(x')\,dx'+c_0. \tag{2.16}$$

Eq. 2.3 is simply the special case of eq. 2.16 when $\sigma_f(x)=$ constant.

The mutual information will be the objective or "energy" function that a learning rule will seek to maximize. However, the mutual information is by no means the only

function that we could use for a learning rule. For instance, the function

$$R(f;x) = -\int \Big(p[f(x)]\,\sigma_f(x)\Big)^n\, p[f(x)]\,df,$$

where $n > 1$, possesses the same variational optimum as the mutual information $\mathcal{I}(f;x) \sim -\int \ln\Big(p[f(x)]\,\sigma_f(x)\Big)\, p[f(x)]\,df$, namely

$$p(f) = \frac{\text{constant}}{\sigma_f}. \tag{2.15}$$

Of course,

$$\ln y = \lim_{n\to 0} \frac{y^n - 1}{n},$$

so the fact that $R(f;x)$ and $\mathcal{I}(f;x)$ have the same optimal probability distribution comes as no surprise. The logarithm function, however, has an attractive additive property that polynomials in $p[f(x)]\,\sigma_f(x)$ do not share. This additive property is useful in optimization, since gradient-based learning rules will result in products of terms derived from the chain rule of calculus, products which turn into sums by taking the logarithm. For example, the entropy function of any composition of functions $f(h(g(V)))$ can be written as follows, as long as each function f, h, g is invertible as a function of its argument:

$$\begin{aligned} S\Big[\, f\{h[g(V)]\}\,\Big] = &-\Big\langle \ln[p(V)] \Big\rangle_{p(V)} + \Big\langle \ln\left|\frac{dg(V)}{dV}\right| \Big\rangle_{p(V)} \\ &+ \Big\langle \ln\left|\frac{dh[g(V)]}{dg(V)}\right| \Big\rangle_{p(V)} + \Big\langle \ln\left|\frac{df\{h[g(V)]\}}{dh[g(V)]}\right| \Big\rangle_{p(V)}, \end{aligned}$$

where $\langle \cdots \rangle$ indicates an integral over the probability distribution $p(V)$.

We are now armed with all the tools we need to investigate how a cell could "learn" to maximize the information in its firing rate about a stimulus distribution. Starting with the determination of a lower bound on the mutual information $\mathcal{I}_{\mathrm{LB}}(f;x)$ between the stimulus x and the firing rate f, we have established what the optimal firing rate distribution and the optimal mapping from stimulus amplitude to firing rate should be. In the next section, we will derive stochastic learning algorithms that change the parameters of the Hodgkin-Huxley model to achieve this desired input-output mapping. If a biological neuron could implement these algorithms (and we will argue that it can), it will be able to maximize the mutual information $\mathcal{I}(f;x)$ in response to arbitrary statistics of the input and to an (almost) arbitrary set of constraints.

2.4 The Adaptation or Learning Rule

Many learning or adaptation algorithms are based on "objective" functions that quantify how close a system is to an optimum; the minimum (or maximum) of said objective function defines the optimum. In our case, the objective function is the mutual information $\mathcal{I}_{\mathrm{LB}}(f;x)$ between the set of stimulus amplitudes x and the set of firing rates $f(x)$. Let us assume that the peak conductances and activation functions of eq. 2.1 can be adjusted through a learning mechanism, so that the function $f(x)$ is also a function of the Hodgkin-Huxley parameters.

The simplest parameter optimization scheme consists of computing the objective function and its gradient, or direction of steepest descent, with respect to the pa-

rameters, and then changing the values of the parameters in the direction of steepest descent or ascent, depending on whether the objective function is to be minimized or maximized. But this simple approach will not work in a biological system: first, a neuron will, in all likelihood, have no representation of the quantity $\mathcal{I}_{\mathrm{LB}}(f;x)$, i.e., neurons do not explicitly compute the amount of information present in the firing rate; secondly, the neuronal mechanisms for adaptation in cells have no direct "access" to the stimulus amplitudes; instead, these mechanisms will depend on some potentially unknown functions of the stimulus amplitude, such as the voltage across the cell membrane or the Ca^{2+} concentration inside the cell.

But to maximize the information transfer, does a neuron need to "know" the arrival rates of photons impinging on the retina or the frequencies of sound waves hitting the ear's tympanic membrane? Since ion channels only sense a voltage and not the stimulus directly, the answer to this question fortunately is no: maximizing the information between the firing rate f and any intermediate variable y, such as the average voltage $\langle V \rangle$, is equivalent to maximizing the information about the stimuli, as long as we can guarantee that the transformation from stimuli to firing rates is always one-to-one. If the only source of noise is additive noise in the output firing rate, this fact is an immediate consequence of a fundamental property of mutual information (Pinsker, 1964): the information between the firing rate and any invertible *function* of the stimulus is equal to the information between the firing rate and the stimulus itself. In a more detailed model, a learning mechanism that increases the channel density of active conductances would, of course, also increase the current

noise. In this case, the information between the firing rate and the average voltage, for instance, would be an upper bound on the information on the stimuli.

Thus a closed-form, analytic expression for the relationship $f(x)$ between firing rate and stimulus amplitude is not required. The function that relates to the stimulus x to the intermediate variable V can be completely unknown and never play a role in the learning algorithm, as long as we can assume that the mapping from x to V is one-to-one. So the answer to the question of whether a neuron needs to have 'direct knowledge' of the stimuli to maximize the mutual information about these same stimuli is, rather surprisingly, *no*. Note that, while the distribution of stimuli x is fixed, that of the intermediate variable V is not; as a consequence, a learning algorithm that relies solely on the voltage V will actually itself change the environment that the neuron "experiences" during learning.

Since a neuron must be able to adapt to a changing environment and shifting intra- and extracellular conditions (LeMasson *et al.*, 1993), the cell should possess learning and relearning mechanisms that can alter the synaptic and voltage-dependent conductances to reflect the environment and computational needs of the whole system. Such learning mechanisms we will assume operate on a continual basis, starting with the development of an animal and continuing throughout the adult life, thus providing a basis for nervous system plasticity.

Naturally, we will suppose that a learning mechanism exists that maximizes the information the neuron's firing rate contains about the stimuli the neuron receives. The mutual information depends on the entire stimulus set, whereas, at any one

time, only a single example of a pairing $(f; V)$ is available to a learning mechanism. As a consequence, the mutual information $\mathcal{I}_{\mathrm{LB}}[f; V]$ itself is usually unknown, and so the learning rule must construct an estimate of it. The standard approach to learning with an objective function that consists of an integral is known as stochastic approximation (Tsypkin, 1971). This approach, in fact, invokes a stochastic sampling of inputs, much like the Monte Carlo technique of integrating functions in high-dimensional spaces. For recent applications of this algorithm to entropy or mutual information maximization, see Linsker (1992), Schuster (1992), or Bell and Sejnowski (1995).

Suppose that q is a parameter for one of the conductances that the cell could adjust. For instance, q could be the number of voltage-dependent high-threshold calcium channels in the cell membrane of the neuron. The cell can change this number by inserting new channels into the membrane or recycling already present ones. Just by how much the cell should change the number of channels to maximize the mutual information is given by the stochastic approximation algorithm:

Each time a stimulus x occurs, the parameter q of the conductance is slightly modified in the direction that most increases a suitable estimate of the mutual information $\mathcal{I}_{\mathrm{LB}}[f; V]$:

$$\Delta q|_x = \eta \frac{\partial}{\partial q} \frac{\delta \mathcal{I}_{\mathrm{LB}}[f; V]}{\delta p[V(x)]}, \tag{2.17}$$

where η is a small, positive, real number known as the learning rate. In other words, the change in the parameter q is proportional to the gradient with respect to q of

the variational derivative of the mutual information. The variational derivative in eq. 2.17 is an instantaneous, but "noisy" estimate of the mutual information, since the expected change Δq averaged over all possible realizations of x is equal to the true gradient:

$$\langle \Delta q|_x \rangle_{p(x)} = \eta \frac{\partial}{\partial q} \Big\langle \ln\{\sigma_f \, p[f(V)]\} \Big\rangle_{p(V)} . \tag{2.18}$$

Equation 2.18 is the functional integral of the right-hand side in eq. 2.17, integrated over the probability distribution $p(V)$.

Recall that the variable V is some unknown, but numerically computable function of the stimulus x, which has a fixed probability distribution $p(x)$. After each successive update of the parameter q, the function relating V and x changes slightly. However, if the learning rate η is very small, $\mathcal{I}_{\rm LB}[f;V]$ remains fairly constant over the short term. If the stimuli are sampled randomly from the distribution $p(x)$, then averaging Δq_x over several steps of the learning mechanism is nearly equivalent to averaging over the stimulus ensemble.

To improve the estimate of the gradient, we can introduce a memory term that involves the previous history of the parameter q. For discrete update schemes, in which the parameter q is updated at unit time intervals, the change in q at time t depends on the change in q at time $t-1$ as follows:

$$\Delta q(t) = \alpha \, \Delta q(t-1) + \eta \frac{\partial}{\partial q} \frac{\delta \mathcal{I}_{\rm LB}[f;V]}{\delta p[V(x)]} ,$$

where $0 \leq \alpha < 1$. Alternatively, we can write down the differential update rule for continuous time learning:

$$\Delta\dot{q}(t) = -\frac{\Delta q(t)}{\tau_{\text{learning}}} + \eta \frac{\partial}{\partial q} \frac{\delta \mathcal{I}_{\text{LB}}[f;V]}{\delta p[V(x)]},$$

where τ_{learning} sets the time scale of the "memory" for past inputs.

Most learning rules in neural networks are, in fact, stochastic algorithms. In a stochastic algorithm, the evolution of parameters, such as the weights in an artificial neural network, can be described by a probability density function for the weight having the value w at time t.

If only one parameter is varied, the Fokker-Planck equation associated with eq. 2.17 (Radons, 1993) predicts that the probability distribution of the parameter q will evolve as:

$$\frac{\partial P(q,t)}{\partial t} = -\frac{\partial}{\partial q}\Big[\langle \Delta q \rangle_{p(x)} P(q,t)\Big] + \frac{\partial^2}{\partial q^2}\Big[\frac{1}{2}\langle (\Delta q)^2 \rangle_{p(x)} P(q,t)\Big]. \tag{2.19}$$

In general, even when $\langle \Delta q \rangle_{p(x)}$ is zero, such that a deterministic algorithm will have converged, $\langle (\Delta q)^2 \rangle_{p(x)}$ will be nonzero. Since this latter term is proportional to η^2, stochastic algorithms usually use an annealing schedule: the learning rate η is slowly decremented in a power-law or exponential fashion, until the parameters being learned are "frozen" into their optimal settings.

If the noise in the signal transduction from a stimulus x to a firing rate f is

additive, then eq. 2.13 predicts that the learning rule should be:

$$\Delta q|_x = \eta \frac{\partial}{\partial q} \ln\left(\frac{df}{dV}\right). \tag{2.20}$$

Note how the logarithm in the definition of the mutual information has simplified the functional derivative.

Now suppose that $f[I(V)]$ depends on a set of parameters q_i. The appropriate learning rule then consists of simultaneously perturbing all parameters as determined by the estimate of the gradient of $S\{f[I(V)]\}$ with respect to q_i. By applying the chain rule of calculus to eq. 2.20. we get

$$\frac{\partial}{\partial q_i}\frac{\delta S\{f[I(V)]\}}{\delta p(V)} = \left[\frac{df[I(V)]}{dI(V)}\right]^{-1} \frac{\partial df[I(V)]}{\partial q_i dI(V)} + \left[\frac{dI(V)}{dV}\right]^{-1} \frac{\partial dI(V)}{\partial q_i dV}.$$

We defer deriving explicit representations of this learning rule until the next sections; the explicit form will differ for different neuronal models.

Maximizing the mutual information $\mathcal{I}_{\mathrm{LB}}$ using eq. 2.20 is similar to a procedure known as simulated annealing, a technique often used for global optimization of high-dimensional problems. In simulated annealing, a 'free energy' is minimized by allowing the system to perform a biased random walk in parameter space. One should visualize the free energy as a landscape in parameter space with hills and valleys (fig. 2.7: as the system performs the random walk, it prefers states of lower energy (valleys) to those of higher energy (hills). A temperature, corresponding to the learn-

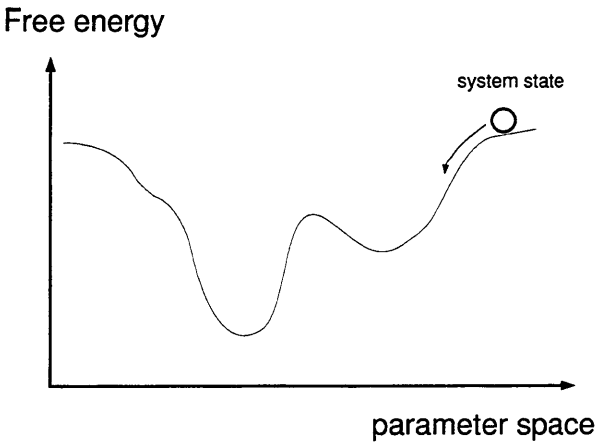


Figure 2.7: Simulated annealing is a stochastic method that seeks out the global minimum of the free energy by varying the parameters of a model system. The free energy landscape is fixed.

ing rate η in eq. 2.20, controls how likely the system is to move into a state of higher energy. If the temperature were zero, the system would never be allowed to escape from local minima in the free energy landscape. As time progresses, the temperature is cooled, freezing the system into a particular low-energy state; with luck, this state corresponds to the global minimum in the free energy.

We first point out the similarities of simulated annealing to information maximization, before discussing the one essential difference. Like simulated annealing, information maximization using eq. 2.20 also corresponds to a random walk in parameter space. In fact, the random walk is described by the Fokker-Planck equation 2.19; the stochastic nature of information maximization stems from the random nature of arriving stimuli x. The simulated annealing temperature corresponds to the learning rate η in eq. 2.20: both the temperature and the learning rate are decremented with time. Moreover, maximizing the mutual information $\mathcal{I}(f;x)$ can, in some instances, be interpreted quite literally as minimizing a free energy. Suppose

that the function relating the firing rate to the current is

$$f(I) = f_{max}\,\mathrm{erf}\left(\sqrt{\frac{\beta}{2}}I\right).$$

In analogy to statistical mechanics, define the free energy F of the current I as follows:

$$F = E(I) - \frac{1}{\beta}S(I), \tag{2.21}$$

where $\beta = 1/T$, the inverse of the temperature, and

$$\begin{aligned} E(I) = \quad \langle I^2/2 \rangle \quad &= \frac{1}{2}\int_0^\infty I^2\, p(I)\, dI \\ S(I) = \quad \langle \ln[p(I)] \rangle \quad &= -\int_0^\infty \ln[p(I)]\, p(I)\, dI \end{aligned}$$

are the energy and entropy of the somatic current distribution $p(I)$, respectively. The energy $E(I) = \langle I^2/2 \rangle$ corresponds to the power dissipated by the physical system.

For this particular $f(I)$ function, maximizing the entropy of the firing rate distribution is identical to minimizing the free energy of the currents. The optimal distribution of somatic currents I is

$$p(I) = \sqrt{\frac{2\beta}{\pi}}\exp\left(-\frac{\beta}{2}I^2\right),$$

which corresponds to the optimal probability distribution of firing rates $p(f) =$ constant if $f \in (0, f_{max})$.

For arbitrary fixed current-discharge relationships $f(I)$, we can define a generalized energy

$$\beta E(I) = -\ln\left|\frac{df(I)}{dI}\right|,$$

such that the optimal probability distribution $p(I)$ is

$$\begin{aligned} p(I) &= \lambda \frac{df}{dI} \\ &= \lambda \exp[-\beta E(I)], \end{aligned} \tag{2.22}$$

where λ is the normalization constant. As long as the generalized energy $E(I)$ can be bounded from below (i.e., as long as infinite negative energies are not possible), we can always interpret the maximization of the mutual information as a free energy minimization problem.

Recall that the goal of information maximization is not to increase the "randomness" or entropy of the firing rates at all costs. In the definition of mutual information, the unconditional entropy is balanced by a conditional entropy that represents the noise in the system. Noise represents but one type of constraint on the unconditional entropy; additional constraints arise through the nonlinearity of the $f(I)$ curve. These constraints enter as additional terms in the learning rule. In the free energy, entropy and energy are balanced relative to each other through the inverse temperature β, which acts as a Lagrangian parameter. (Note that the temperature β is not related to the temperature used in simulated annealing.) The same balancing act must occur

in information maximization.

Unlike simulated annealing or most artificial neural network learning algorithms, however, the free energy landscape in parameter space is *not* fixed for the information maximization procedure outlined here. Eqn. 2.20 does not correspond to true stochastic gradient ascent on the mutual information $\mathcal{I}_{\mathrm{LB}}(f;x)$. Such a procedure would require the computation of all derivatives, including dV/dx. Instead, each probability distribution of voltage $p(V)$ corresponds to a *different* free energy landscape. Since the algorithm changes the probability distribution $p(V)$ continually, it not only performs a drunken walk, but also deforms the energy landscape as it does so; one should, therefore, imagine the free energy landscape as made of jello, free to wobble. The minimum of $\mathcal{I}_{\mathrm{LB}}(f;x)$ must, however, always coincide with the minimum of $\mathcal{I}_{\mathrm{LB}}(f;V)$. It is for this reason that the learning algorithm will work.

2.5 The Set of Basis Functions

An alphabet zoo of different calcium (Ca^{2+}) conductances, denoted 'L', 'N', 'P', 'R', and 'T', is present in pyramidal neurons of the central nervous system (Helmchen *et al.*, 1996; Avery and Johnston, 1996). These conductances activate rapidly as a function of voltage (with a mean slope of about 10 mV^{-1}), and have been traditionally divided into two classes: low-voltage activated (LVA) channels such as the 'T' type, and high-voltage activated (HVA) channels such as the 'L'(L = long-lasting) type. In reality, a spectrum of Ca^{2+} ion channel types exists, which differ from each other in the midpoint of voltage activation or their pharmacological properties. For instance,

the activation function for the 'N' type channel lies midway between that of the 'T' and 'L' types.

A similar myriad number of potassium K^+ conductances exists, including the delayed rectifier current, various types of A-currents, delay currents, muscarine-sensitive currents, and Q-currents. In addition, there are fast, calcium-gated potassium conductances, whose activation functions rapidly shift to lower potentials in the presence of increasing internal Ca^{2+} currents.

In the spike initiation zone in the axon hillock region of a neuron, a set of powerful sodium and potassium conductances gives rise to the nerve impulse—the action potential. A set of weaker voltage-dependent conductances in the dendrites of pyramidal neurons in neocortex and hippocampus modulates both the subthreshold and suprathreshold response to stimuli (for a recent review, see Yuste and Tank, 1996). The learning mechanism we will develop will adjust this latter set of conductances.

We will assume a Hodgkin-Huxley formalism for all conductances, eschewing more detailed models, such as the Goldman-Hodgkin-Katz equations, that account for current rectification. For each electrotonic compartment of a neuron, the Hodgkin-Huxley model corresponds to a coupled set of nonlinear differential equations for the membrane potential V and the gating variables m (for activation) and h (for inactivation) for each conductance. In each electrotonic compartment of the neuron, the

voltage V evolves as

$$C\frac{dV}{dt} = \sum_j g_j^{\text{passive}}\,(E_j - V) + \sum_i g_i^{\text{active}}\, m_i^{p_i}(V)\, h_i^{q_i}(V)\,(E_i - V), \tag{2.23}$$

where C is the membrane capacitance, g_j^{passive} are the voltage-independent synaptic and leak conductances, g_i^{active} are the peak values of the voltage-dependent conductances, p_i and q_i are integers, and E_i are fixed, ion-specific reversal potentials. The index $i = 1, 2, \cdots, N$ runs over all N independent conductances within the compartment that have different voltage dependencies.

If we only consider conductances that activate, but do not inactivate ($q = 0$), then such Ca^{2+}, K^+, Na^+ or Cl^- conductances are analogous to the "neurons" of artificial neural networks, which are characterized by sigmoidal relationships between input and output. (The Boltzmann function is typically called the logistic function in the neural network literature.)

For Hodgkin-Huxley conductances with a single gating particle and only two states of the underlying ion channel, open and closed, the steady-state value of the conductance g_i corresponding to channel type i is given by the sigmoidal Boltzmann function:

$$g_i = \bar{g}_i\, m_{\infty,i}(V) = \frac{\bar{g}_i}{1 + \exp[-s_i(V - V_i)]},$$

where $\bar{g}_i$ is the peak conductance, s_i is the slope at maximal inflection of the activation function, and V_i is the midpoint voltage. The peak conductance for any one type is

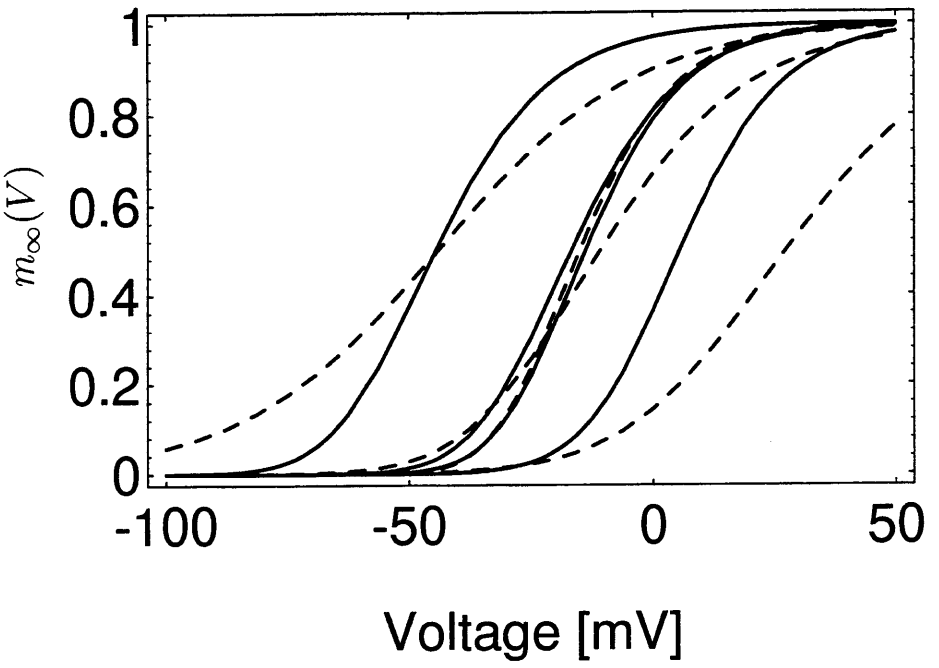


Figure 2.8: An activation curve in the Hodgkin-Huxley model corresponds to a time-dependent basis function in an artificial neural network. Illustrated as solid lines are the steady-state activation functions for calcium Ca^{2+} 'L', 'N', and 'T' conductances and the Hodgkin-Huxley Na^+ conductance taken from Lytton and Sejnowski, 1991. The dotted lines illustrate various potassium conductances, such as 'M', 'A', 'K_D','K_C' at 10 mM [Ca^{2+}]. Parameters for K^+ conductances taken from *ibid.* and Johnston and Wu, 1995.

the product of the density of channels in the membrane and the peak single channel conductance.

A learning mechanism that adjusts the steady-state activation functions of conductances will change any one of three different types of parameters, singly or in

combination:

peak conductance/channel $*$ channel density	$\bar{g}_i$
midpoint (half-activation) voltage	V_i
conductance slope	s_i

If the conductance model is more complicated than the one given by the Boltzmann function, there will be more than three parameters that could be adjusted during "learning." In addition, learning mechanisms could, in principle, change the time constants of activation and inactivation. However, we will defer consideration of learning rules that change the time constants until section 2.13.

Conceptually, we may think of the voltage activation curves of different types of Ca^{2+} and K^+ channels as providing a set of sigmoidal basis functions. During learning, we will change the parameters of these "sigmoidal basis functions," so that the firing rate response matches the target function that maximizes information transmission. But the analogy to feedforward perceptrons only goes so far: voltage and conductances interact in a nonlinear manner, complicating the analysis. The voltage sets the state of the conductance, but the voltage-dependent conductance in turn resets the voltage, establishing a nonlinear feedback loop.

A standard result from the theory of universal approximation states that one can approximate any sufficiently smooth ($\mathcal{L}_2$) function to arbitrary accuracy by the superposition of a sufficient number of basis functions (Ripley, 1996), as illustrated

in fig. 2.9. This result lies at the heart of most of learning theory for artificial neural networks.

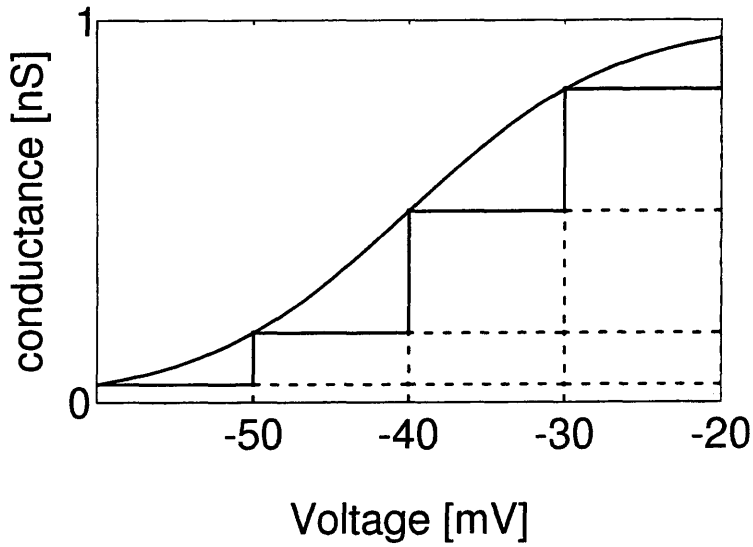


Figure 2.9: If we take the limit of infinite slopes in the steady-state activation curves of conductances, then the smooth curves of fig. 2.8 become step functions. In this schematic, the smooth, voltage-dependent activation function of the overall conductance is approximated by three step functions as a function of the dendritic voltage V. If the true conductance is a sum of N conductances, each based on a channel type with a steplike activation curve, then a learning rule should change the density of these channel types to achieve the best approximation to the smooth curve as shown.

If the target conductance is a known function of voltage, then the following supervised learning rule suggests itself for adapting the peak conductances associated with each activation curve for the voltage-dependent conductances:

$$\Delta \bar{g}_i = \eta \cdot (\text{Error}) \cdot \left[\frac{d}{dV} m_{\infty,i}(V) \right]^{\gamma}, \tag{2.24}$$

where η is a learning rate, the error is a measure of how close the model's performance

is to the target performance for that input, and γ is some positive, real number. (The error can be any metric or pseudo-metric, such as the Kullback-Leibler distance used by Bayesian statisticians.) This error is provided by a so-called 'teacher'; hence the learning rule is called a supervised learning rule. If $\gamma = 1$, this learning rule reduces to gradient descent on the derivative of $1/2\,(\text{Error})^2$ with respect to the intermediate variable V. In situations involving nonlinear feedback, such a learning rule is preferable to straightforward gradient descent on the error: in effect, the bell-shaped curve of $\frac{d}{dV}m_\infty(V)$ selects a subset of weights for updating each time an input is drawn from the input space, instead of updating all weights. Nonlinear feedback can make a system sensitive to small changes in parameters; straightforward gradient descent thus has a greater potential for instability.

Using the schematic of fig. 2.9, we can decoct the essence of the supervised learning rule 2.24. For simplicity, suppose the voltage is quantized in units of ΔV. In fig. 2.9, the actual smooth voltage-dependent activation curves have been replaced by a discrete number of step functions spaced ΔV apart. Consider, for instance, an input that leads to a voltage $k\Delta V$, where k is an integer. In response to such an input, the learning rule of Eq. 2.24 adjusts *only* the peak conductance associated with the activation curve centered at $k\,\Delta V$, leaving all other conductances unaffected. Even in the presence of the nonlinear feedback loop between voltages and voltage-dependent conductances, the effect of this isolated adjustment is fairly specific.

Learning rules to change a set of basis functions are typically derived from an objective function that measures how close a system's performance is to ideal or

perfect performance. These objective functions will naturally lead to learning rules that involve the activation functions $m_{\infty,i}(V)$ and their derivatives $\frac{d}{dV}m_{\infty,i}(V)$. It is hence not surprising that the the unsupervised learning rule that maximizes the information transfer of the neuron bears a striking similarity to the supervised learning rule for mapping a range of inputs onto an arbitrary set of outputs:

$$\Delta\bar{g}_i = \eta\left[\frac{d}{dV}m_{\infty,i}(V)\right]. \tag{2.25}$$

This learning rule will be derived in section 2.7 from first principles. The only difference to the previous learning rule is that no error measure multiplies the term on the right-hand side of the equation: the teacher that provided feedback on how close the system was to the 'right' answer has quite literally disappeared. Why such a simple maneuver should yield the appropriate unsupervised learning rule is the nontrivial subject of the next sections.

Note that the learning rule of eq. 2.25 bears no relationship to correlation-based or Hebbian learning, which is the approach that has dominated thinking about learning in biological systems for the past fifty years. In the classic postulate of Hebb (Hebb, 1949), learning is a consequence of correlated activity between pre- and postsynaptic neurons, generally thought to be subserved by long-term potentiation (LTP) and long-term depression (LTD). If two neurons i and j are connected by a synapse W_{ij}, then the strength of that synapse is changed in (anti-)Hebbian learning according to

$$\Delta W_{ij} = \pm\eta\Big(f_i\, f_j - \theta\Big), \tag{2.26}$$

where f_i, f_j represent the activity of the two neurons, respectively, θ is a threshold, and the sign determines whether learning is anti-Hebbian or Hebbian. To be precise, we have omitted from both eqs. 2.25 and 2.26 a term to control the peak growth of the parameter being changed.

The difference in the two learning rules arises because the equivalent of a synapse between neurons does not, in general, exist at the subcellular level: different voltage-dependent conductance types are typically not coupled to each other except through the transmembrane voltage. (An exception to this rule is the calcium-dependent potassium conductance.) Since the states of conductances are only indirectly linked to the input-output relationship of the neuron, the learning rule 2.25 makes no direct use of correlations.

2.6 Feedforward Learning

To explore what the learning rule of eq. 2.20 on page 47 actually means, we begin with a zeroth order model: Suppose that a fixed function, independent of any of the subcellular parameters of the neuron, relates each possible synaptic input to a steady state voltage, which, in turn, is related by some variable function to the response, or firing rate, of the neuron. We thus imagine that each input stimulus "clamps" the voltage to a particular known value, but that the response of the neuron can be changed as a function of that voltage.

The assertion that a steady state voltage exists is already an assumption, which will be treated in detail in section 2.8. For the full dynamical system of a spiking

Hodgkin-Huxley model, we will generalize the notion of steady state to include periodic functions of voltage—the quasi-steady state—before taking the last step and examining fully time-varying, non-periodic functions of voltage.

Clamping the voltage leads to a steady state current I across the neuronal membrane composed of three components, namely synaptic, passive (leak), and active currents:

$$\begin{aligned} I(V) &= I_{\text{synaptic}} + I_{\text{leak}} + I_{\text{active}} \\ &= g_{\text{syn}}(E_{\text{syn}} - V) + g_{\text{leak}}(E_{\text{leak}} - V) + \sum_i \bar{g}_i \, m_{\infty,i}^p(V) h_{\infty,i}^q(V)(E_i - V), \end{aligned}$$

where we have simply given names to some of the components of eq. 2.1. (The synaptic and leak components are simply terms with $p_i = q_i =$ in eq. 2.1.) To simplify matters even further, we introduce the shorthand notation $\phi_{\infty,i}(V)$ to denote the product of the steady-state conductance fraction and the driving potential:

$$I(V) = I_{\text{synaptic}} + I_{\text{leak}} + \sum_i \bar{g}_i \phi_{\infty,i}(V).$$

Together with the peak conductance $\bar{g}_i$, $\phi_{\infty,i}(V)$ yields the so-called "window"-current for the i-th voltage-dependent conductance.

Since the relationship between synaptic input and steady-state voltage is fixed, the probability distribution of voltage inputs is also fixed and unchanging. Changing the active conductance parameters in the equation above will affect the relationship between the steady-state current I and the steady-state voltage V, but, by assump-

tion, not change the probability distribution of voltages. If we let the neuron's firing rate response to the steady-state current be some saturating, nonlinear function $f(I)$, then learning can occur by changing $I(V)$. Note that we will assume that the function $f(I)$ to be given, *deus ex machina*, until we derive the $f(I)$ relationship directly from the full, spiking Hodgkin-Huxley equations in section 2.10.

In summary, a stimulus x is converted to a firing rate f in three stages:

$$x \rightarrow V \rightarrow I \rightarrow f,$$

i.e., the stimulus maps onto a voltage, the voltage onto a current, and finally the current onto a firing rate. In the zeroth order model, only the mapping from $V \rightarrow I$ is variable.

Learning rules for artificial neural networks are typically based on a fixed probability distribution of voltage inputs, so a close analogy to neural network learning exists for this zeroth order model. Even though this approach might be expected to make the problem more tractable, we will show that the resulting adaptation rules for the conductances turn out to be horrendously complicated. Only by using the feedback inherent in the full nonlinear dynamical system will the adaptation rules simplify considerably.

To obtain an explicit representation for the learning rules, we start with the innocuous-looking equation from section 2.4:

$$\Delta q|_x = \eta \frac{\partial}{\partial q} \ln \left(\frac{df}{dV} \right). \tag{2.20}$$

For the sake of simplicity, we will typically make the assumption that the noise in the model is additive and independent of the firing rate, so maximizing the mutual information between stimuli and firing rates is equivalent to maximizing the entropy of the firing rate distribution. This assumption is inherent in the choice of eq. 2.20 as the learning rule.

The first complication arises because of the logarithm in eq. 2.20. Computing the rules for adapting the peak conductances, midpoint voltages, and voltage slopes of various voltage-dependent conductance types requires taking the partial derivative of $\ln(dI/dV)$, the logarithm of the derivative of the steady state current with respect to the steady state voltage. This leads to a term

$$\left(\frac{dI(V)}{dV}\right)^{-1} = \left\{\sum_j \bar{g}_j \frac{d}{dV}\left[\phi_{\infty,j}(V)\right]\right\}^{-1} \tag{2.27}$$

in the learning rule. All update rules, whether for the peak conductance, the midpoint voltage, or the conductance slope of a particular ion channel type, are thus coupled to all other ion channel types through the sum $\sum_j \bar{g}_j \frac{d}{dV}\left[\phi_{\infty,j}(V)\right]$. This coupling reflects the fact that the mutual information is a *global* property of the stimulus set.

However, this coupling between the learning rules for different ion channel types is inherently biophysically implausible—how would a neuron compute the sum over derivatives of different activation functions required in the learning rule? A learning rule for changing the parameters of conductance should be completely *local*: it should depend only on the voltage and the properties of the ion channel itself. In addition, $\frac{dI(V)}{dV}$ can tend to zero, rendering the learning rule unstable.

Similar problems plague attempts to maximize the information in networks of nonlinear neurons by changing the synaptic strengths of connections. Straightforward stochastic gradient ascent on the mutual information or entropy of firing rates is another form of feedforward learning, and leads to nonlocal learning rules. This can easily be seen by generalizing from the simplest case of one presynaptic neuron and one postsynaptic neuron; note that only in this simplest case is the resulting learning rule still local, since only one parameter is present.

In applying the information maximization approach to changing the synaptic strengths in a simple network, we treat the active, voltage-dependent conductances inside single cells as fixed, incorporating their effect into the nonlinear neuronal response function $f(I)$. Suppose that the current I driving the postsynaptic neuron is linearly proportional to the firing rate of the presynaptic neuron:

$$f_{\text{post}}(I) = f(\alpha f_{\text{pre}}), \tag{2.28}$$

where α is the measure of synaptic strength, and f_{pre} and f_{post} are the firing rates of the presynaptic neuron and the postsynaptic neuron, respectively. If the presynaptic firing rate f_{pre} is a random variable with a fixed distribution, then the stochastic approximation learning rule for α is

$$\begin{aligned} \Delta\alpha &= \frac{\partial}{\partial\alpha_i} \ln\Big[\alpha f'(\alpha f_{\text{pre}})\Big] \\ &= \frac{1}{\alpha} + \frac{f''(\alpha f_{\text{pre}})}{f'(\alpha f_{\text{pre}})}, \end{aligned} \tag{2.29}$$

where the prime denotes a derivative with respect to the argument. The extension to multiple input neurons and multiple output neurons is immediate (see, for example Schuster, 1992; Bell and Sejnowski, 1995) and involves the computation of the inverse synaptic weight matrix that describes the couplings between neurons. This inverse matrix replaces α^{-1} in the equation above and requires an inherently *global* computation. The next section will explore a more powerful and clever approach to information maximization that will avoid such global computations.

This is not to say that all global computations are biologically implausible. For instance, a spiking neuron's intracellular calcium concentration is thought to track the average firing rate; LeMasson *et al.* (1993) propose conductance adaptation rules designed to maintain a constant firing rate by keeping the calcium concentration steady. Here, the calcium concentration plays the role of a global variable. However, information maximization is a comparatively much harder task. Even though the calcium ion is the journeyman of neurobiology, involved in many neuronal processes, from the release of neurotransmitter at presynaptic terminals to the activation of protein kinases in the phosphorylation cascades, it is but one molecule. As such, it can represent one variable, e.g., the average firing rate, but not all the different moments of the firing rate distribution required for information maximization. We do not intend to imply that calcium does *not* play a major role in changing the information transfer properties of a neuron, only that adaptation mechanisms should have a component directly tied to or initiated by the ion channels that underlie the individual conductances (see section 2.2). Appendix A.1 shows how calcium-based

information maximization can, in fact, be an instantiation of feedforward learning, with all its attendant complications.

The second complication in deriving explicit learning rules from eq. 2.20 arises from the nonlinearity $f(I)$ relating the current to the firing rate. This nonlinearity leads to a multiplicative term $h(I) = \left[\frac{df}{dI}\right]^{-1} \frac{d^2 f}{dI^2}$, which can take on a bewildering variety of forms, reflecting the sensitivity of the $f''(I)$ term to changes in the saturating nonlinearity. Technically speaking, $h(I)$ is the radial curvature of the $f(I)$ function. The function $h(I)$ is plotted on a logarithmic scale versus current in fig. 2.11 for different $f(I)$ functions. We write out $h(I)$ for several choices of $f(I)$ in table 2.1.

Table 2.1

$f(I) = f_{\max} \operatorname{erf}\left(\sqrt{\frac{\beta}{2}} I\right)$	$h(I) = -\beta I$	Error Function
$f(I) = f_{\max} \tanh\left(\frac{\beta}{2} I\right)$	$h(I) = -\frac{\beta}{f_{\max}} f\left(I\right)$	Tanh
$f(I) = \frac{a}{1 + \exp\left[-\beta(I - I_0)\right]} - c$	$h(I) = \frac{\beta}{a}\left\{a - 2\left(f(I) + c\right)\right\}$	Boltzmann
$f(I) = a I^{\beta}$	$h(I) = -\frac{1 - \beta}{I}$	Power-Law
$f(I) = a \exp(\beta I)$	$h(I) = \beta$	Exponential
$f(I) = f_{\max} \frac{I}{I + k_d}$	$h(I) = -\frac{2}{I + k_d}$	Michaelis-Menten
$f(I) = a I$	$h(I) = 0$	Linear

For the first three choices of $f(I)$, maximizing the entropy on f can be interpreted as minimizing a free energy of the currents I, because we can, in principle, write the

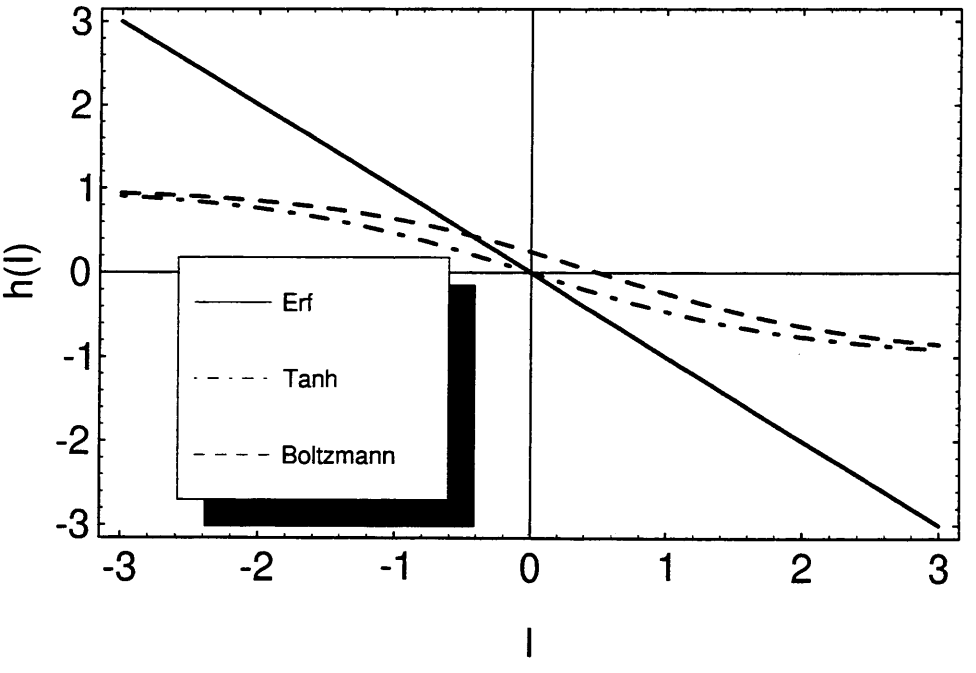


Figure 2.10:

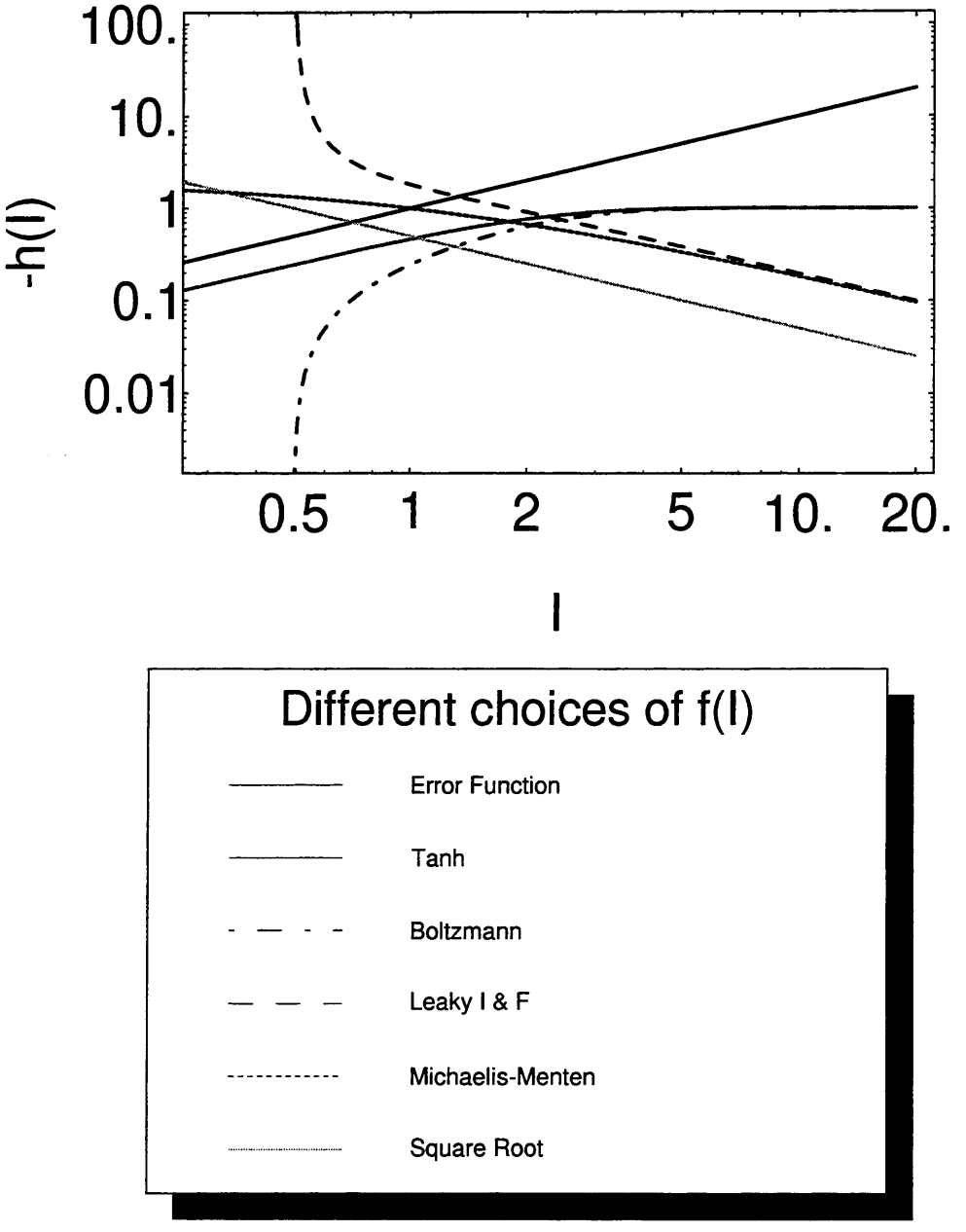

100.
10.
1
0.1
0.01
-h(I)
0.5
1
2
5
10.
20.
I
Different choices of f(I)
Error Function
Tanh
Boltzmann
Leaky I & F
Michaelis-Menten
Square Root

Figure 2.11:

function $h(I)$ as the functional derivative of a scalar quantity that is bounded from below. In fig. 2.10, we see that the $h(I)$ term in the learning rule for these firing rate nonlinearities penalizes high currents:

In principle, hard thresholds in the $f(I)$ function present a problem, since the multiplicative factor $h(I)$ diverges near the threshold. Any realistic model of a neuron will not contain such hard thresholds, since the presence of noise will "smear out" any discontinuity in the derivative of the $f(I)$ relationship. As a general rule, the firing rate for both the Hodgkin-Huxley and leaky integrate-and-fire model behaves as

$$f(I) = \frac{\theta - I}{\sigma} \exp\left[-\frac{(\theta - I)^2}{\sigma^2}\right],$$

if I is well below the threshold θ and noise is present (see Chapter 3). In this regime, a constant noise model for the firing rate is absolutely invalid. Since the timing of spikes is nearly Poisson for spiking models with subthreshold currents, a more appropriate noise model is $\sigma_f^2 \propto f$. In this case, maximizing the mutual information (or constrained entropy) causes the multiplicative factor $h(I)$ to tend to

$$h_{\text{Poisson}}(I) = \frac{f''(I)}{f'(I)} - \frac{1}{2}\frac{f'(I)}{f(I)}$$

in the subthreshold regime.

Hence the asymptotic limit of $h(I)$ is

$$h(I) \sim \frac{\theta - I}{\sigma^2} \qquad \text{as } I \to -\infty.$$

In general, if a neuron is to adapt its firing rate probability range to be uniform over all possible firing rates, then the twice differentiable, bounded function $f(I)$ must be quickly saturating. By quickly saturating, we mean simply

$$\lim_{I->\pm\infty} \frac{f'(I)}{f''(I)} \neq 0$$

Neither leaky integrate-and-fire models with a fixed refractory period nor Hodgkin-Huxley models will have quickly saturating current-discharge relationships, such as would be given by a tanh function, for instance. This poses somewhat of a quandary: while the nonlinearity near the threshold current for firing is well understood, what the appropriate nonlinearity for firing rate saturation should be is less clear. This issue will be explored in detail in section 2.8.

While writing down the learning rules in their fullest form is not particularly edifying, we can gain some appreciation for the complexity of the feedforward learning by considering the case where conductances only activate, but do not inactivate. Calculating the learning rules from eq. 2.20 on page 47 explicitly for the changes in peak conductance $\bar{g}_i$, midpoint voltage V_i, and slope s_i of the various conductance types yields:

$$\Delta\bar{g}_i = \eta \left(\frac{dI(V)}{dV}^{-1} \left\{ \left[\frac{d}{dV} m_{\infty,i}(V) \right] (E_i - V) - m_{\infty,i}(V) \right\} + h\big(I(V)\big)\, m_{\infty,i}(V)\, (E_i - V) \right) \tag{2.30a}$$

$$\Delta V_i = \eta \left(\frac{dI(V)}{dV}^{-1} \bar{g}_i \left\{ -\left[\frac{d^2}{dV^2} m_{\infty,i}(V) \right] (E_i - V) + \left[\frac{d}{dV} m_{\infty,i}(V) \right] \right\} - h\big(I(V)\big)\, \bar{g}_i \left[\frac{d}{dV} m_{\infty,i}(V) \right] (E_i - V) \right) \tag{2.30b}$$

$$\Delta s_i = \eta \left(\frac{dI(V)}{dV}^{-1} \bar{g}_i \left\{ \frac{(V - V_i)}{s_i} \left[\frac{d^2}{dV^2} m_{\infty,i}(V) \right] (E_i - V) + \frac{\big(E_i + V_i - 2V\big)}{s_i} \left[\frac{d}{dV} m_{\infty,i}(V) \right] \right\} - h\big(I(V)\big)\, \bar{g}_i \frac{(V_i - V)}{s_i} \left[\frac{d}{dV} m_{\infty,i}(V) \right] (E_i - V) \right), \tag{2.30c}$$

where η is the learning rate, and $h(I) = \left[\frac{df}{dI} \right]^{-1} \frac{d^2 f}{dI^2}$ is the function that reflects the nonlinearity of $f(I)$ in the learning rule, and we have used the Boltzmann function for the activation curves: $m_{\infty,i}(V) = 1/\{1 + \exp[-s_i(V - V_i)]\}$, where s_i is the slope of the activation function and V_i is the midpoint voltage of half-activation.

The first term in each of the learning rules maximizes the entropy of the currents. Note that these adaptation rules *per se* do not rule out negative conductances, which would make the Hodgkin-Huxley system biophysically unrealistic and numerically unstable. Therefore, the adaptation rule for the peak conductance is made conditional: if the peak conductance $\bar{g}_i$ were to become negative, then the peak conductance is simply set to zero instead. Increasing the entropy of the currents alone without any

safeguards can lead to rampant and unchecked growth of peak conductances and slopes. The multiplicative factor $h(I)$ in the second term of the learning rules keeps this behavior in check.

2.7 Computing with the Steady State–The Role of Feedback

In a Hodgkin-Huxley model of a neuron, voltages and conductances are nonlinearly coupled: the conductances affect the voltage, which, in turn, sets the voltage-dependent conductances. At the level of the single cell, the true input is not a voltage, but a synaptic conductance. While the function relating synaptic conductance to voltage is highly nonlinear and unknown, the system must satisfy the strict physical constraint of charge conservation: when the neuron is firing periodically, the average current injected by the synaptic and voltage-dependent conductances must equal the average current discharged by the neuron; when the neuron is in a steady state, the same constraint holds true. This constraint will be the key to obtaining a local learning rule that completely uncouples conductances from each other, as we shall see below.

We illustrate the steady state case with a simplified model of a single neuron containing one electrotonic compartment, displayed in fig. 2.12. Current is injected through a synaptic conductance g_{syn}, whose value is drawn randomly from a fixed probability distribution. A set of Ca^{2+} and K^{+} voltage-dependent conductances acts

to amplify or deamplify the synaptic current—these conductances are subject to the learning mechanism. The current discharged across the leak conductance g_{soma} is converted into a response by a fixed function $f(I)$.

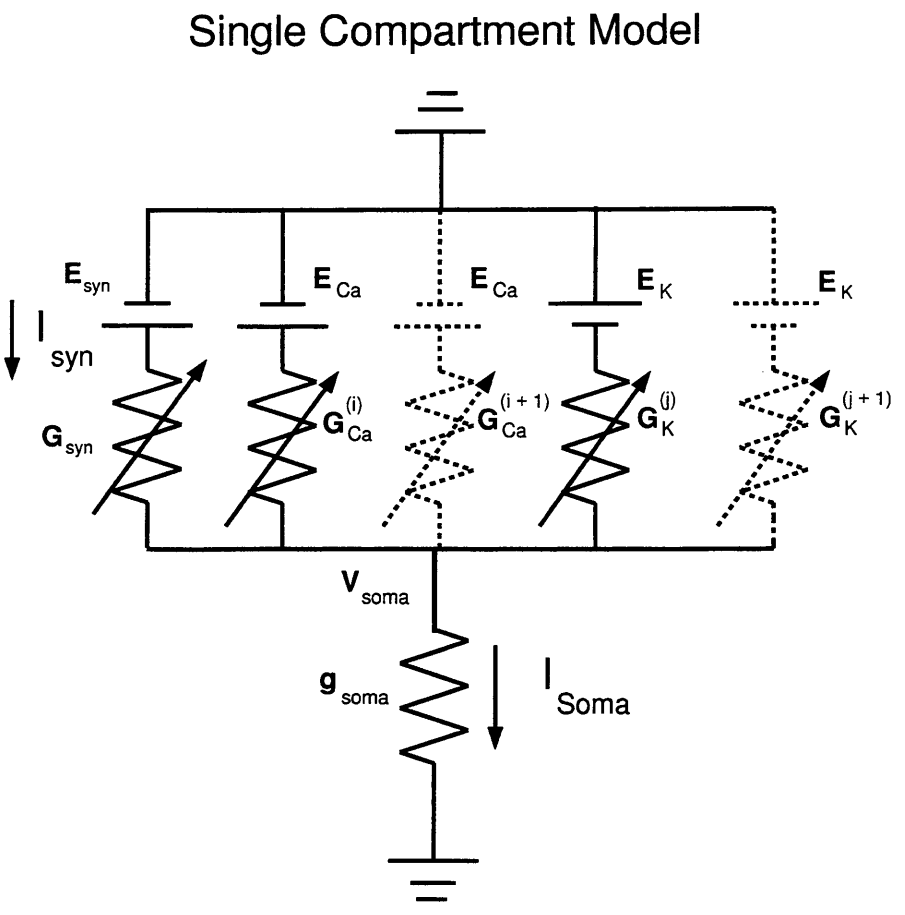


Figure 2.12: Single compartment model with multiple conductances. The dashed lines are meant to suggest a variable number of different channel types.

Whether adaptation of the voltage-dependent conductances will be successful depends on whether a stable steady state voltage exists for each input to the neuron at each stage in learning. In the case of periodically spiking neurons, is the simple periodic state stable for all possible inputs? These are nontrivial issues, particularly for spiking neurons. In the next section, the existence of the true steady state be examined, which will motivate the use of an auxiliary function that is used to simplify

the learning rules for changing conductance parameters.

We will, however, postpone some of the details of the full derivation of the adaptation rules, instead emphasizing how charge conservation leads to learning rules that are *local*, so that the mechanism for changing one conductance does not depend on the values of any other conductances. For the reader less interested in formal derivations, a more intuitive explanation of the learning rule is given below eq. 2.31.

As in the feedforward learning case in the previous section, the learning rule requires the computation of $(dI(V)/dV)^{-1}$, which arises from eq. 2.20 on page 47. However, now the probability distribution of the voltages is no longer fixed, but that of the synaptic conductances is. Recall that in the noiseless case, the mutual information between response and steady state voltage $I(f; V^*)$ is identical to the mutual information between response and input $I(f; g_{\text{syn}})$. Since the steady state voltage V^* and the synaptic conductance g_{syn} are related by an invertible transformation, we can write g_{syn} as a function of V^*. By virtue of charge conservation, we can write the steady-state current across the leak conductance either as

$$I_{\text{soma}}(V^*) = g_{\text{soma}} V^*,$$

reflecting the current discharged, or as the the current injected

$$I_{\text{soma}}(V^*) = g_{\text{syn}}(V^*)(E_{\text{syn}} - V^*) + \sum_{i=1}^{N} \bar{g}_i \, m_{\infty,i}^p(V^*) \, h_{\infty,i}^q(V^*)(E_i - V^*).$$

Note that $g_{\text{syn}}(V^*)$ does not explicitly depend on the parameters of the voltage-

dependent conductances, and thus plays no role in the partial derivatives in the learning rule 2.20 on page 47. We conclude, therefore, that

$$\left(\frac{dI(V)}{dV}\right)^{-1} = \left\{\sum_j \bar{g}_j \frac{d}{dV}\left[m_{\infty,j}^p(V) h_{\infty,j}^q(V)(E_j - V)\right]\right\}^{-1} \tag{2.27}$$

in feedforward learning is replaced by

$$\left(\frac{dI(V)}{dV}\right)^{-1} = \frac{1}{g_{\text{soma}}} \tag{2.30}$$

thanks to the nonlinear coupling between voltage-dependent conductances and voltage. Since the term $(dI(V)/dV)$ need no longer be written as a sum over terms containing different conductance types, a local learning rule results.

If we contract all constant factors in eq. 2.20, we obtain the beautifully simple formula for changing the peak conductances of the model:

$$\boxed{\Delta\bar{g}_i = \eta\left[\frac{d}{dV}\phi_{\infty,i}(V) + c(V)\,\phi_{\infty,i}(V)\right],} \tag{2.31}$$

where η is a positive, real number called the learning rate, $\phi_{\infty,i}(V) = m_{\infty,i}^p(V) h_{\infty,i}^q(V)(E_i - V)$ is the product of the steady-state conductance fraction and the driving potential, and $c(V)$ is a function that depends on the nonlinearity in the current-discharge (f-I) relationship, the constraints on the firing rate distribution, and the noise model. In the simplest case, $c(V)$ is a function that is zero for most average voltages, equal to a positive constant if the firing rate is below a minimum target, and equal to a negative

constant if the firing rate is above a maximum target voltage. Thus,

$$c(V) = \begin{cases} \gamma & \text{if } V < V_l, \\ 0 & \text{if } V_l \leq V \leq V_h, \\ -\gamma & \text{if } V > V_h. \end{cases} \tag{2.32}$$

when $f(V)$ can be approximated by a linear function over the range $[f(V_l), f(V_h)]$.

The dominant term in the learning rule is $\frac{d}{dV}\phi_{\infty,i}(V)$, which acts to increase the entropy of the outputs. If the ion channels activate, but do not inactivate, then this term is simply bell-shaped function of voltage $\frac{d}{dV}m_{\infty,i}(V)$ that reflects the transitions of ion channels between open and closed states at the microscopic scale (see section 2.2). By assumption, the stimulus x and the voltage V are related to the firing rate response f by monotonically increasing functions. The first term in the learning rule changes primarily those conductances that increase the slope of the firing rate response to x. A higher slope means that more of the neuron's limited range of firing rates is devoted to representing the stimulus x and its immediate neighborhood. Since the learning rule is democratic yet competitive, only the most frequent inputs "win" and thereby gain the largest representation in the output firing rate.

The second term $c(V)\,\phi_{\infty,i}(V)$ limits the steady-state voltages into which the neuron is allowed to settle. Without this term, the conductances would grow without bound. Examples of the uniform (flat) probability distribution of firing rates resulting from applying this learning rule in numerical simulations are shown in fig. 2.16 on page 88 and fig. 2.27 on page 123.

Using the Boltzmann equation for the (in)activation functions, we can also write down update rules for the other conductance parameters, namely the midpoint voltages and slopes of the (in)activation functions. The equation for the midpoint voltages V_i is

$$\Delta V_i = -\eta\, \bar{g}_i \left\{ \frac{d^2}{dV^2}\left[m_{\infty,i}^{p_i}(V)(E_i - V)\right] + \left[c(V)(E_i - V) - 1\right] \frac{d}{dV} m_{\infty,i}^{p_i}(V) \right\} h_{\infty,i}^{q_i}(V).$$

For the slopes s_i, we have

$$\begin{aligned}\Delta s_i =& \eta \frac{\bar{g}_i}{s_i} \left\{ (V - V_i) \frac{d^2}{dV^2}\left[m_{\infty,i}^{p_i}(V)(E_i - V)\right] \right. \\ & \left. + \left[c(V)(V - V_i)(E_i - V) + (E_i - V_i)\right] \frac{d}{dV} m_{\infty,i}^{p_i}(V) \right\} h_{\infty,i}^{q_i}(V).\end{aligned}$$

Even though more complicated than eq. 2.31, the equations for changing the midpoint voltages and slopes are still local: the change in the conductance parameter depends only on the state of that conductance and the voltage across the membrane. To set these last two equations into the context of eq. 2.31, we rewrite the dominant term in the update rules for ΔV_i and Δs_i as a function of $\Delta \bar{g}_i$. For the purpose of making the resulting equations compact, we consider only the case when $p_i = 0$ or $q_i = 0$:

$$\begin{aligned}\Delta V_i &= -\bar{g}_i \left(\frac{d}{dV} \Delta \bar{g}_i \right) \\ \Delta s_i &= \frac{\bar{g}_i}{s_i} \left(\Delta \bar{g}_i + (V - V_i) \frac{d}{dV} \Delta \bar{g}_i \right).\end{aligned}$$

Even though the learning rate η is a common factor multiplying all the learning

rules for all conductance types, η by itself does not represent a global coupling of different adaptation mechanisms to each other. If we rescale all parameters q_i by a random, nonzero vector α, such that $\bar{q}_i = \alpha_i q_i$, then all the learning rules for the new parameters $\bar{q}_i$ are rescaled accordingly. Equivalently, we can rescale the learning rate for each parameter arbitrarily, without respect to the other parameters. In fact, there exists a natural scale for changing the peak values of the voltage-dependent Ca^{2+} and K^+ conductances, which is set by the one conductance that does *not* change in the model of fig. 2.12: the fixed conductance g_{soma}. On this scale, the peak conductances evolve as

$$\tau_i \frac{d\bar{g}_i}{dt} = \eta(t)\left[g_{\text{soma}} \frac{d}{dV}\phi_{\infty,i}(V) + g_{\text{soma}}^2 c(V)\,\phi_{\infty,i}(V)\right], \tag{2.33}$$

where we have taken the liberty to convert the discrete update rule in eq. 2.31 into a differential equation and introduced a time scale τ_i. When we consider time-varying signals in section 2.13, we will show that this time scale τ_i is related to the time constants in the equations $dm_i/dt = (m_{\infty,i}(V) - m_i)/\tau_i(V)$ governing the activation of voltage-dependent conductances. The learning rate $\eta(t)$ is dimensionless and the function $c(V)$ has the dimensions of an inverse current (see sections 2.6 and 2.8). Given these facts, a quick check reveals that both sides of equation 2.33 are dimensionally consistent.

2.8 Stability of the Steady State*

The steady state voltage V^* is given by setting CdV/dt to zero in the Hodgkin-Huxley equations. To ease the notational burden, define $\mathbf{G}$ as the vector of conductances in eq. 2.23, which includes the synaptic conductance and the voltage-dependent conductances. Let $\mathbf{E}$ be the associated vector of reversal potentials. We thus rewrite the Hodgkin-Huxley set of differential equations as

$$C\frac{dV}{dt} = \mathbf{G}\cdot\mathbf{E} - V\,\mathbf{G}\cdot\mathbf{1}$$
$$\tau_i(V)\frac{dm_i}{dt} = m_{\infty,i}(V) - m_i.$$

The steady state voltage V^* is the solution to

$$V^* = \frac{\mathbf{G}(V^*)\cdot\mathbf{E}}{\mathbf{G}(V^*)\cdot\mathbf{1}}, \tag{2.34}$$

in which all activation variables m_i have been replaced by their steady-state values $m_{\infty,i}(V^*)$. Since g_{soma} is one of the components of $\mathbf{G}(V^*)$, we have of course $\mathbf{G}(V^*)\cdot\mathbf{1} \geq g_{\text{soma}} > 0$, so eq. 2.34 is well-defined. For any fixed choice of peak conductances, $||\mathbf{G}(V^*)||$ is bounded; hence at least one fixed point (steady state voltage) of the differential equation must exist. The steady state voltage V^* will be bounded by the lowest and highest ionic reversal potentials, which are typically the ones for potassium E_{K} and calcium E_{Ca}.

If the fixed point is stable, the Hodgkin-Huxley system will return to the fixed

point under a perturbation $V^* \to V^* + \delta V$ and $m_i \to m_{\infty,i}(V^*) + \delta m_i$. To determine the conditions that need to be satisfied for the fixed point to be stable, we linearize the Hodgkin-Huxley equations around the fixed point:

$$C\frac{d(\delta V)}{dt} = -\mathbf{G}(V^*)\cdot\mathbf{1}\,\delta V + \sum \bar{g}_i(E_i - V^*)\,\delta m_i$$
$$\tau_i(V^*)\frac{d(\delta m_i)}{dt} = m'_{\infty,i}(V^*)\,\delta V - \delta m_i,$$

where we have used the fact that $m_{\infty,i}(V^*) = m_i$ at the fixed point to eliminate the derivative of $\tau_i(V^*)$, and we have ignored the inactivation variable h_i, since including it will not change the stability criterion in eq. 2.35 below. This set of equations can be written as a matrix differential equation with entries in the first row and column and along the diagonal, zeroes everywhere else.

By applying the definition of the determinant in terms of the Levi-Civita tensor and noting that only the diagonal term and combinations of $\bar{g}_k(E_K - V^*)$ with $m'_{\infty,k}(V^*)$ remain, it can be shown that the determinant of the stability matrix is

$$\frac{(-1)^N}{C\prod_{k=1}^{N-1}\tau_k(V^*)}\left[\sum G(V^*) - \sum_{k=1}^{N-1}\bar{g}_k\,(E_k - V^*)\,m'_{\infty,k}(V^*)\right] = \frac{(-1)^{N-1}}{C\prod_{k=1}^{N-1}\tau_k(V^*)}\,\frac{dI}{dV}(V^*),$$

where N is the rank of the matrix, i.e., the number of variables in the Hodgkin-Huxley differential equations, and the *total* derivative dI/dV is

$$\frac{dI}{dV}(V^*) = \mathbf{G}'(V^*)\cdot\left(\mathbf{E} - V^*\cdot\mathbf{1}\right) - \mathbf{G}(V^*)\cdot\mathbf{1}. \tag{2.35}$$

Since the determinant is equal to the product of eigenvalues, and the eigenvalues must all be negative for the system to be stable, a necessary condition for stability is that $S(V^*) < 0$, or in other words,

$$\sum_{k=1}^{N-1} \bar{g}_k \, m_{\infty,k}(V^*) + g_{\text{soma}} + g_{\text{syn}} > \sum_{k=1}^{N-1} \bar{g}_k \, m'_{\infty,k}(V^*) . (E_k - V^*).$$

To establish that the mapping $g_{\text{syn}} \to V^*$ is absolutely monotonic and increasing, we compute dV^*/dg_{syn} explicitly. Writing the steady state as the solution to an equation with two variables V^* and g_{syn},

$$F(V^*, g_{\text{syn}}) = \mathbf{G}(V^*) \cdot \mathbf{E} - V^* \, \mathbf{G}(V^*) \cdot \mathbf{1} = 0,$$

we have

$$\begin{aligned} \frac{dV^*}{dg_{\text{syn}}} &= -\frac{\partial F / \partial g_{\text{syn}}}{\partial F / \partial V^*} \\ &= \frac{E_{\text{syn}} - V^*}{\mathbf{G}(V^*) \cdot \mathbf{1} - \mathbf{G}'(V^*) \cdot \left(\mathbf{E} - V^* \cdot \mathbf{1} \right)} \\ &= \frac{E_{\text{syn}} - V^*}{-\dfrac{dI}{dV}(V^*)}. \end{aligned}$$

Thus as long as $E_{\text{syn}} > V^*$ and the fixed point is stable, i.e., $-\frac{dI}{dV}(V^*) > 0$, then dV^*/dg_{syn} will be positive. By the implicit function theorem, this implies that, within the local neighborhood of g_{syn}, the mapping $g_{\text{syn}} \to V^* \to I_{\text{soma}} \to f$ is a composition of absolutely monotonic, increasing functions. A mapping of g_{syn} onto a compact set

of unstable fixed points is inconsistent with an increasing function.

Note that the mapping of $g_{\text{syn}} \rightarrow V^*$ may only be piecewise continuous and differentiable, because it can be punctuated by discontinuous jumps due to the presence of unstable fixed points. Unstable fixed points can cause some ambiguity, since the steady state reached will depend on the initial conditions. In devising a learning rule, we obviate such ambiguity by choosing a constant initial voltage V. If no unstable fixed points exist, the mapping $g_{\text{syn}} \rightarrow V^*$ will be continuous and differentiable.

Before learning begins, we will assume initial conditions such that $V^*(g_{\text{syn}})$ is an increasing function of g_{syn}. Since the optimal response of the neuron is proportional to the cumulative distribution function of the inputs, the target $V^*(g_{\text{syn}})$ is also an increasing function. To maintain a monotonic mapping from input to response output, the learning rule must enforce the constraint $E_{\text{syn}} > V^*$. This constraint sets a limit to the maximal current across the leak conductance g_{soma}, even when the synaptic conductance g_{syn} becomes very large:

$$I_{max} = \lim_{g_{\text{syn}} \rightarrow \infty} \frac{g_{\text{syn}}\, g_{\text{soma}} E_{\text{syn}}}{g_{\text{syn}} + g_{\text{soma}}} = g_{\text{soma}} E_{\text{syn}}.$$

While the current delivered can be increased by inserting channels with reversal potentials $E_{\text{rev}} > E_{\text{syn}}$, this would lead to the undesirable situation in which the synaptic current decreases as the voltage increases.

If the current is limited, the neuron is not allowed to reach the absolute maximum f_{max} of its response. We thus seek a systematic method to map the inputs onto a range $[f_l, f_h] \in [0, f_{max}]$ with uniform probability, where $f_l \geq 0$ and $f_h \leq f_{\text{max}}$. The

most natural approach to this problem is to define an auxiliary function $a(f)$:

$$a(f) = \begin{cases} 0 & f < f_h, \\ f & f_l \leq f \leq f_h, \\ 0 & f > f_h. \end{cases} \tag{2.36}$$

When noise is added to the output, but not the input, maximizing the mutual information is equivalent to entropy maximization. Ideally, we would like to use the entropy function

$$S[a(f)] = -\int_0^\infty p[a(f)] \ \ln\{p[a(f)]\}\, dh$$

in a stochastic approximation learning rule. However, $a(f)$ is discontinuous and has derivative zero for $f > f_h$ and $f < f_l$. Hence $a(f)$ is not invertible and, therefore, utterly useless. Instead, define

$$a(f) = \begin{cases} f_l \exp\left[+\frac{(f-f_l)}{\lambda_l}\right] & f < f_l \\ f & f_l \leq f \leq f_h, \\ f_h \exp\left[-\frac{(f-f_h)}{\lambda_h}\right] & f > f_h. \end{cases} \tag{2.37}$$

This is a fair approximation to the previous definition of $a(f)$, particular if λ_l and λ_h

are large. The extremal solution to $S[a(f)]$ with $a(f)$ given by eq. 2.37 is

$$p(f) = \text{constant} \times \frac{da}{df}$$

$$= \begin{cases} c_1 \exp\left[+\frac{(f-f_l)}{\lambda_l}\right] & f < f_l, \\ c_2 & f_l \leq f \leq f_h, \\ c_3 \exp\left[-\frac{(f-f_h)}{\lambda_h}\right] & f > f_h. \end{cases}$$

With the constraint that $p(f)$ be continuous at $f = f_l$ and $f = f_h$ and that $\int_0^\infty p(f)\,df = 1$, these constants are $c_1 = c_2 = c_3 = \left\{(f_h - f_l) + \lambda_h + \lambda_l[1 - \exp(-f_l/\lambda_l)]\right\}^{-1}$.

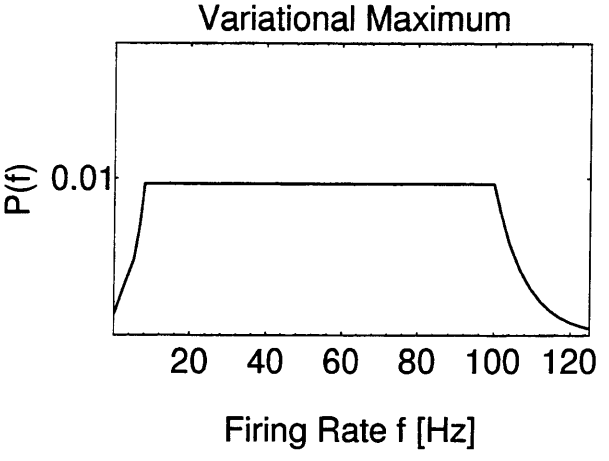


Figure 2.13: The auxiliary function $a(f)$ from eq. 2.37 allows one to define a target probability distribution for the firing rates even in the absence of saturating nonlinearities in the output firing rate. This optimal probability distribution is uniform over a range of firing rates between $[f_l, f_h]$, and exponentially decaying outside this region.

More rigorously, $S[a(f)] = S^-[a(f)] + S[a(f)] + S^+[a(f)]$ where $S^-[a(f)]$ is defined

on the support $f < f_l$, $S[a(f)]$ on the support $f_l \leq f \leq f_h$, and $S^+[a(f)]$ on the support $f > f_h$. The goal of stochastic approximation will be to simultaneously maximize $S^-[a(f)]$, $S[a(f)]$ and $S^-[a(f)]$.

The auxiliary function $a(f)$ leads to an additional term in the learning rule if $f \notin [f_l, f_h]$. For $q_i = \bar{g}_i, V_i, s_i$, we have

$$\Delta q_i = (\Delta q_i)_{\text{orig}} + \eta \left\{ -\frac{1}{\lambda_h} \frac{df}{dI} \frac{\partial I}{\partial q_i} \right\}$$

whenever $f > f_h$, or

$$\Delta q_i = (\Delta q_i)_{\text{orig}} + \eta \left\{ +\frac{1}{\lambda_l} \frac{df}{dI} \frac{\partial I}{\partial q_i} \right\}$$

whenever $f < f_l$.

Since the function $f(I)$ is fixed and current conservation relates the voltage V to the current I, we can replace the condition $f < f_l$ by $V < V_l$ and $f > f_h$ by $V > V_h$ above. When $f(I) \propto I$, the factor $\pm\lambda^{-1} df/dI$ multiplying the second term in the learning rule of eq. 2.31 is absolutely required to ensure convergence of learning. In this case, it replaces the factor $h(I) = [df/dI]^{-1} \, d^2f/dI^2$ arising from the nonlinearity of the f-I curve, which becomes zero if the f-I curve is linear.

2.9 Simulations Gallery

In simulations of the simple model shown in fig. 2.12 on page 72, the steady state is reached by solving a time-discretized iterative map of the Hodgkin-Huxley dynamics, as illustrated in fig. 2.14. The conductance is computed for a particular voltage,

then either the voltage or conductance is clamped, yielding the next voltage in an iterative map. For instance, when voltages are successively clamped in this iterative map procedure, the voltage at iteration or time $t+1$ is

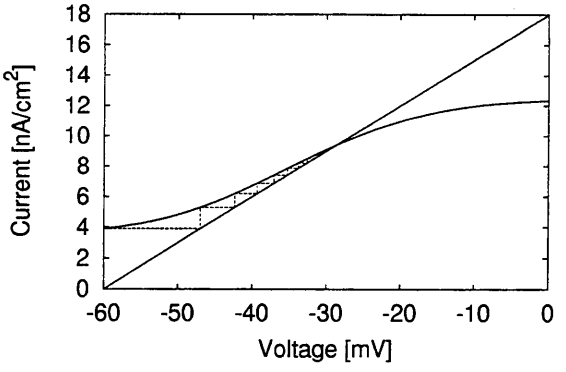
$$V(t+1) = \frac{1}{g_{\text{soma}}} \sum_i \bar{g}_i m_{\infty,i}^p [V(t)] \, h_{\infty,i}^q [V(t)] \, [E_i - V(t)] \, .$$

The limit of this map, $\lim_{t\to\infty} V(t)$, is the steady state voltage.

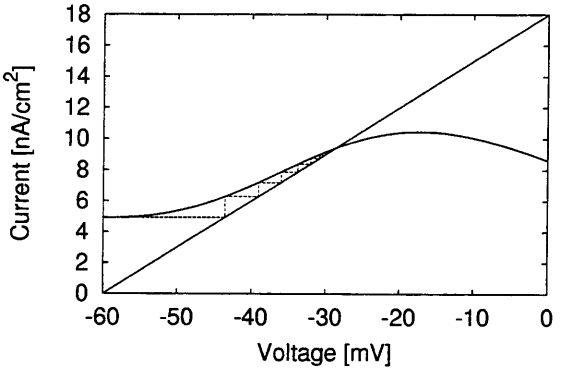
For simplicity and without loss of generality, numerical simulations were done with simple activation functions for each conductance and no inactivation functions. The number of conductance types was varied between six and twenty for both Ca^{2+} and K^+ conductances. At the beginning of learning, the midpoints of the Ca^{2+} activation functions were equally spaced between the resting potential and some maximum voltage. K^+ conductances also started out equally spaced, but the range of midpoint voltages was offset relative to the Ca^{2+} conductances to below the resting potential. All peak conductances are initially set to zero.

Since the simple model does not explicitly model the spiking mechanism, current across the leak resistor g_{soma} is transduced into a "firing rate" or response by a fixed function, such as the function $f(I) = 418/(1 + \exp[1.22(I - 0.38)]) - 161.78$ shown in fig. 2.15. In section 2.10, the challenges to this simple model are addressed by considering a true spiking model.

We take the input variable to be the input current to a presynaptic neuron with the same $f(I)$ function (illustrated in fig. 2.15). The firing rate of the presynaptic neuron is taken to be linearly proportional to the synaptic conductance driving the



(a) conductance clamp



(b) voltage clamp

Figure 2.14: Two equivalent numerical methods can be used to compute the steady state. In both cases, the value of all conductances is computed for an initial membrane potential, usually the resting potential. On the left in panel (a), the conductances are clamped, giving rise to a current through the somatic conductance g_{soma} to ground; this current leads to a new membrane potential, as indicated by the straight line. This process is iterated until the value of the current no longer changes; the dynamical system's evolution traces out the zig-zag line. On the right in panel (b), the voltage instead of the conductance is clamped, and the resulting current is recorded; as before, the current determines the next voltage.

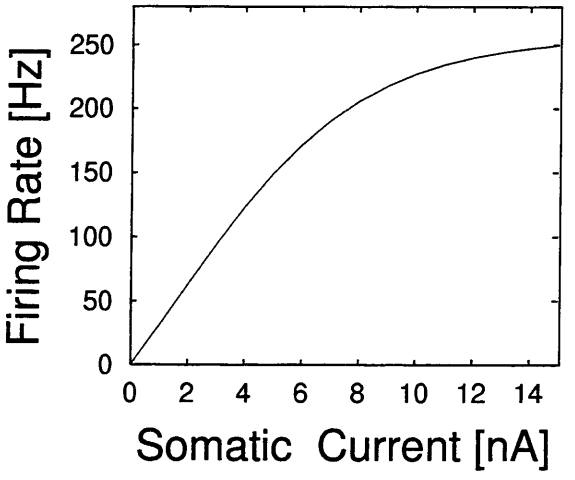
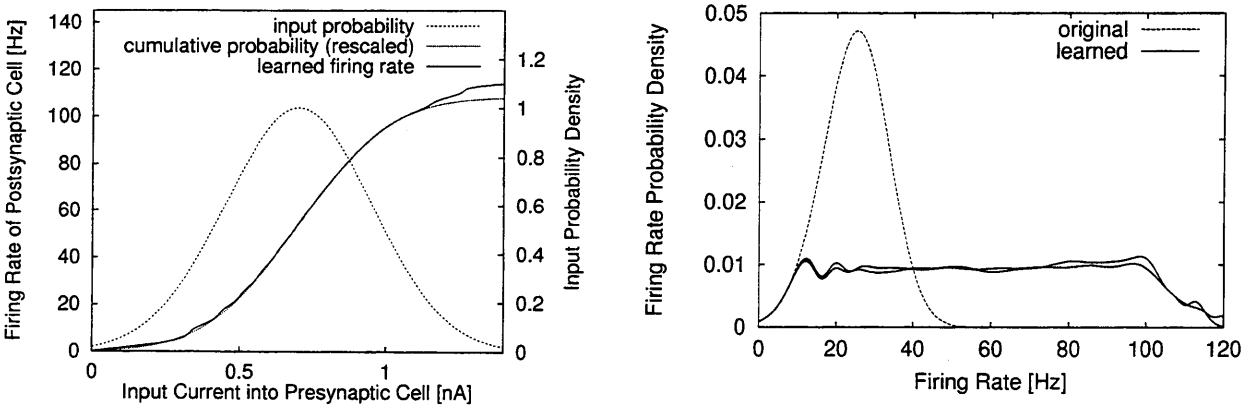


Figure 2.15: Transduction of current across the somatic leak resistor g_{soma} into a firing rate is given by a fixed function in the simple model. An analytical approximation to the "f-I" relationship that can be directly derived from the Hodgkin-Huxley equations is one of the subjects of section 2.10.



(a) (b)

Figure 2.16: 500,000 iterations of the update rules were used to adjust the peak conductances, midpoint voltages, and slopes of 20 Ca^{2+} conductances. The midpoint voltages of the calcium conductances were originally spaced 2 mV apart; the maximal slope at the inflection point of each conductance activation curve was initially 2.72 mV. The input-output mapping the system learns approximates the integral $f(x) = C\int_{-\infty}^{x} p(z)\,dz$, which is none other than the cumulative distribution function of the inputs. The learning rate decayed exponentially: $\eta = \eta_0 \exp(-t/\tau_{\text{learning}})$, with $\eta_0 = 10^{-5}$ and $\tau_{\text{learning}} = 50,000$ iterations. After 500,000 iterations, the system settles into a state in which all firing rates are used with nearly equal probability, except at the edges of the firing rate range. To show the variation in the learning algorithm from run to run, the results of two identical runs are superimposed in panel (b).

postsynaptic neuron to fire.

Illustrated in fig. 2.16 is a Gaussian distribution of input currents to the presynaptic cell with a mean of 0.7 nA and a standard deviation of 0.25 nA. Whether the input variable is the synaptic conductance, the firing rate of a presynaptic cell, the current injected into a presynaptic cell, or the arrival rate of photons at the retinal photoreceptors is irrelevant as long as the noiseless transformation of the input to output is one-to-one and any noise present in the system is additive. Note that noise was not added to the input in these simulations. Below, we will consider the effect of external feedback loops adding to the synaptic input of the postsynaptic neuron;

this consideration motivates the the choice of input we use here.

Fig. 2.16 shows the result of training the neuron model of fig. 2.12 on the Gaussian distribution of input currents described above, using the learning rule

$$\Delta\bar{g}_i = \eta\Big[\frac{d}{dV}\phi_i(V) + c(V)\,\phi_i(V)\Big] \tag{2.31}$$

for the peak conductances and the corresponding learning rules for ΔV_i and Δs_i for the midpoint voltages and slopes, respectively. The postsynaptic cell "learns" the structure of the input, matching its firing rate to the cumulative distribution function of the input currents. In so doing, the neuron adapts itself so that all firing rates within the neuron's dynamic range are used roughly equally often.

Limiting the maximal dynamic range of firing rates or alternatively limiting the average firing rate imposes a constraint on the entropy of the firing rate distribution. The function $c(V)$ incorporates this constraint. When the postulated somatic f-I relationship is linear or nearly linear, then the function $c(V)$ is extremely simple, as shown in eq. 2.32 on page 75. In the linear case, we can essentially ignore the conversion from current into firing rate completely, and simply maximize the entropy of the currents. Changing the f-I relationship adds the term $h(I) = [df/dI]^{-1}d^2f/dI^2$ to the function $c(V)$. The optimal distribution of currents or voltages after learning remains proportional to df/dI, but df/dI is no longer constant. The non-uniform voltage probability distributions that result from applying the learning rule eq. 2.31 to nonlinear f-I relationships are shown in fig. 2.17.

By choosing the input variable to be the current to a presynaptic neuron, we can

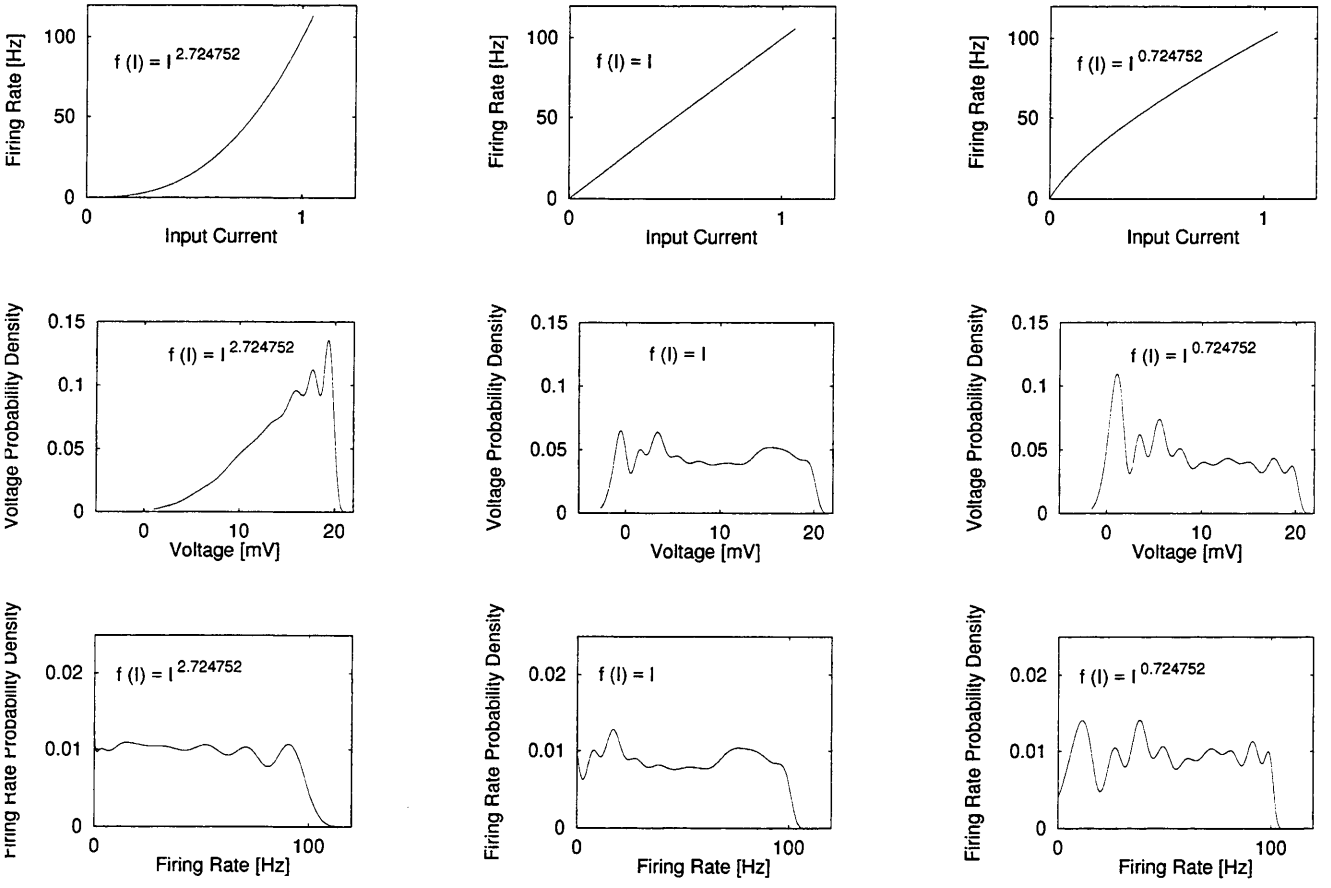


Figure 2.17: Different current-discharge (f-I) relationships lead to different target voltage probability distributions, as given by eq. 2.22 on page 50. The target firing rate probability distribution remains uniform in all cases. The same input current distribution was used as in fig. 2.16(a). Graphs show the resulting probability distributions after 200,000 iterations of the learning rule, with an initial learning rate of $\eta = 5 \times 10^{-6}$. (For $f(I) = I^{0.724752}$ $\eta = 10^{-8}$).

explore the effect of synaptic feedback in the model. Most of the synapses made onto a neocortical neuron originate from other neurons within cortex (McGuire *et al.*, 1984; Kisvarday *et al.*, 1986; White, 1989). Anatomical studies reveal that not more than 10% or the synapses onto excitatory stellate cells in visual cortex arise directly from the lateral geniculate nucleus (Ahmed *et al.*, 1994), which is the intermediate stage between the retina and the cortex in the visual processing stream. These anatomical studies led to the 'canonical microcircuit' hypothesis of Douglas and Martin (1991),

which implies that excitatory feedback dominates cortical processing. A number of studies (Douglas *et al.*, 1995; Somers *et al.*, 1995) have explored the possible functional roles for such massive, excitatory recurrent connections. Of course, it should be noted that the mere presence of predominantly excitatory synapses in the anatomy does not necessarily dictate that the functional effect is excitatory; for an alternative view, see Stemmler *et al.* (1995).

The simplest abstraction of the many network feedback loops present in mammalian cortex is an autapse, i.e., a neuron making a synapse onto itself. With such

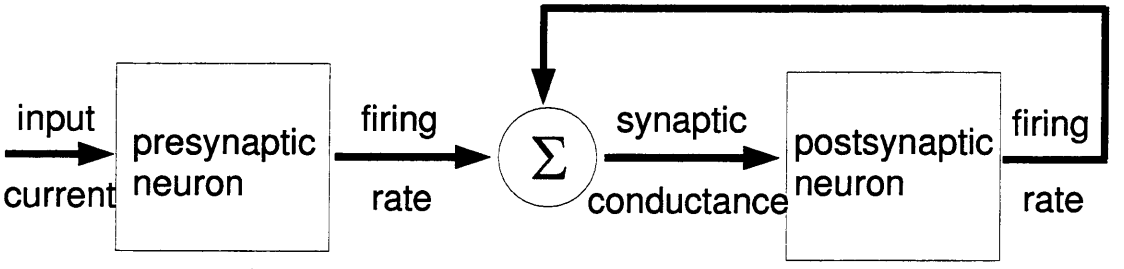


Figure 2.18: Model of synaptic feedback.

feedback, the synaptic conductance of the postsynaptic cell is

$$g_{\text{syn}} = \alpha f_{\text{pre}} + \beta f_{\text{post}},$$

where f_{pre} and f_{post} are the firing rate of the presynaptic and postsynaptic cells, respectively, and α and β are synaptic weights.

From a theoretical point of view, synaptic feedback is a trivial extension to the previous model: even though the probability distribution of synaptic conductances is no longer fixed, we know that no amount of processing can change the maximum mutual information between the original input and the firing rate of the postsynaptic

neuron—this is known as the 'data processing inequality' (Cover and Thomas, 1991).

If the sign of the synaptic feedback is constant as a function of firing rate, then feedback will preserve the one-to-one nature of the mapping from input to firing rate. In such a case, we can apply the learning rule 2.31 without any additional modification. Fig. 2.19 shows the response of the model with six Ca^{2+} and six K^{+} conductances trained on a Gaussian input current distribution to a presynaptic cell of mean 0.75 nA/cm^2 and standard deviation 0.25 nA/cm^2. The firing rates of the presynaptic and the postsynaptic neurons add together in equal proportion to drive the postsynaptic cell; the synaptic weight converting the summed firing rate into a synaptic conductance has the numerical value of 8×10^{-4} $nS/(cm^2Hz)$.

To measure the progress of the learning algorithm to maximize the mutual information or entropy, we plot the entropy of the firing rate distribution as a function of time in fig. 2.20. Typically, two phases in the learning process emerge: a fast phase, characterized by rapid increase in the entropy of the firing rate distribution as measured in bits, followed by a slow phase, during which the firing rate distribution slowly creeps towards the target distribution. Since the actual goal of learning is to maximize the mutual information between firing rates and stimuli subject to certain constraints, the entropy of the firing rates, which is only one component of the mutual information, need not always increase. In addition, the algorithm is based on a stochastic estimate of the mutual information gradient with respect to an intermediate variable; hence, departures from true gradient ascent on the mutual information are to be expected.

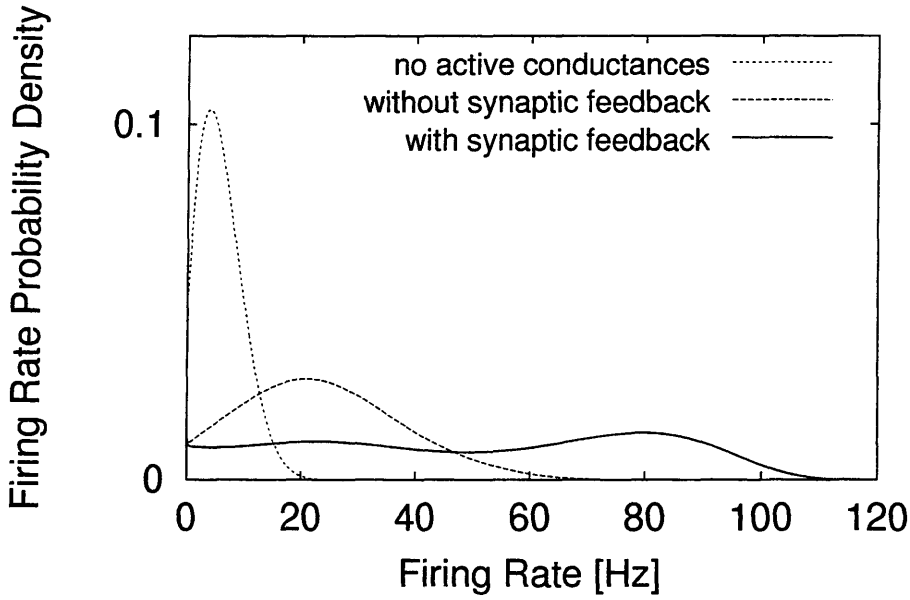
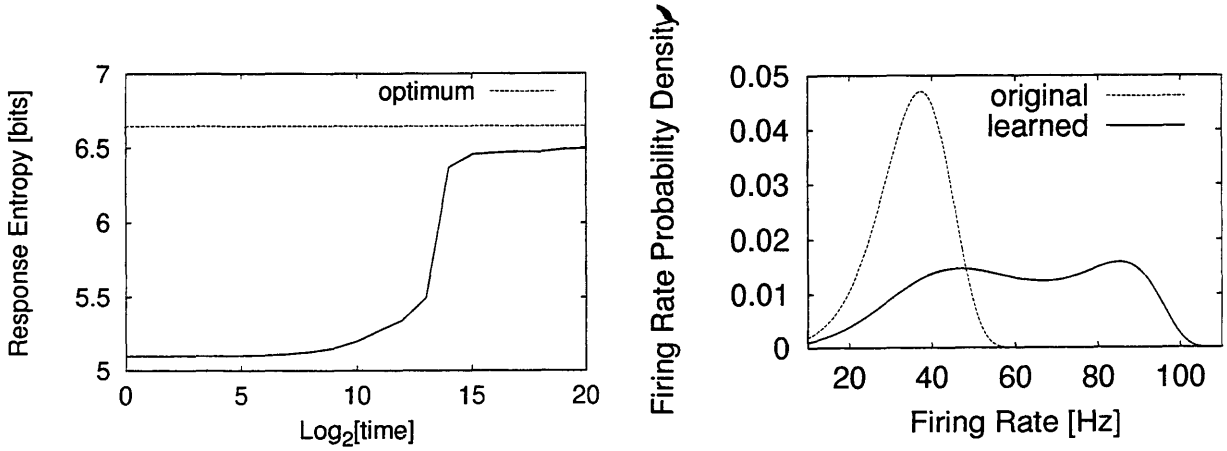


Figure 2.19: The model of fig. 2.18 was trained for 2×10^7 iterations on a Gaussian distribution of input currents to a presynaptic neuron (see text). The initial learning rate was $\eta_0 = 2.5 \times 10^{-6}$, and the firing rate function $f(I)$ was chosen to be concave: $f(I) = I^{2.72}$. Without any active conductances, firing rates cluster near zero, even in the presence of synaptic feedback. Since the model was trained in the presence of synaptic feedback, the model achieves a nearly uniform probability distribution only while the synaptic feedback remains on; removing the feedback without changing the parameter settings for the active conductances results in a peaked distribution of firing rates.

To get a better understanding of the dynamics of the learning rule, we perform the following *Gedankenexperiment*: let the activation curve slope of all conductances remain fixed at a constant value and examine a "point" conductance centered at $V_{\frac{1}{2}}$. Here, $V_{\frac{1}{2}}$ can take on any arbitrary value. We ask how this fictitious conductance would evolve under the learning rule 2.31 in the presence of a fixed number of real, simultaneously evolving voltage-dependent conductances. One can imagine that the changing probability distribution of voltages exerts a dissipative, average "force" on the point conductance, causing the peak conductance at $V_{\frac{1}{2}}$ to grow or decay as a



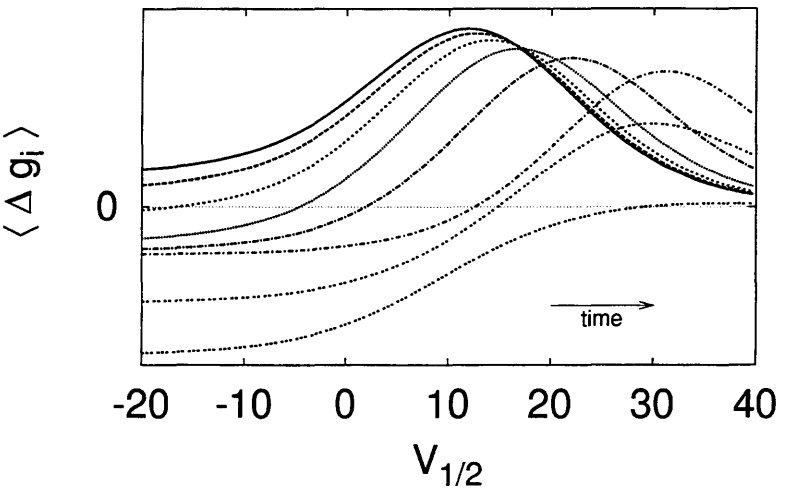
(a) (b)

Figure 2.20: Progress of the information maximization algorithm as a function of time for one run of the learning algorithm with six Ca^{2+} and six K^+ conductances. For simplicity, only the peak conductances were allowed to change. The learning rate was kept constant and small at $\eta = 2 \times 10^{-8}$. The learning algorithm was iterated 524,288 times; time is plotted on a logarithmic scale. The fact that the algorithm exhibits a fast and a slow learning phase leads to the sigmoidal shape of the learning curve in the semilogarithmic graph. On the right, the final shape of the output probability distribution of firing rates is plotted.

function of time. This average "force" corresponds to the true gradient of the free energy function of the firing rates with respect to the voltage across the cell membrane, as given by eq. 2.18. (Recall that the learning algorithm, by design, seeks out the minimum of the free energy function, albeit in a highly stochastic fashion.)

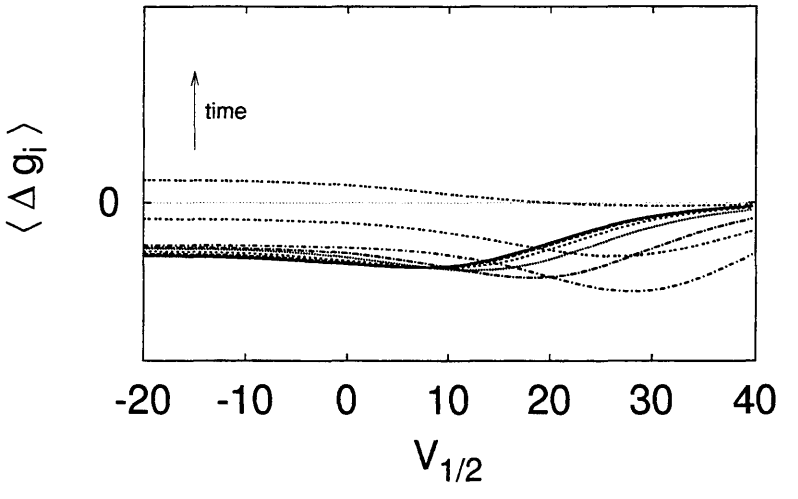
$$\Delta \bar{g}_i(t) = \left\langle \left[\frac{d}{dV} \phi_i(V) + c(V)\, \phi_i(V) \right] \right\rangle_{p[V(t)]} \tag{2.38}$$

Fig. 2.21 displays the gradient of the free energy with respect to voltage in eq. 2.38 as a function of the midpoint voltage of the point conductance. During learning, the Ca^{2+} peak conductances centered near the resting potential are increased first, am-



(a) Evolution of Ca^{2+} peak conductances

plifying the overall response of the neuron. As learning progresses, the Ca^{2+} conductances that activate at high voltages are strengthened, whereas those conductances that were strengthened earlier are made weaker again. The gradient for the Ca^{2+} conductances thus sweeps out a traveling wave moving to the right in time. Not represented in fig. 2.21 is the half-wave rectification needed in the learning rule to keep the conductances positive (since a negative conductance would be unphysical). Since the peak conductances all start out at zero, the K^+ conductances, therefore, only start changing later in the learning process. As amplification by Ca^{2+} conductances becomes dominant, K^+ conductances begin to counteract the amplification, allowing for a fine-tuning of the neuron's response to input. Convergence of the stochastic learning algorithm in the mean implies that the gradient in eq. 2.38 vanishes for all nonzero conductances; for all zero-valued conductances, the gradient should be either



(b) Evolution of K^+ peak conductances

Figure 2.21: "Force" exerted on a point peak conductance by the learning rule 2.31 as a function of time and the midpoint voltage $V_{\frac{1}{2}}$ of the point conductance. The real conductances evolving are six Ca^{2+} and six K^+ conductances, exactly as in fig. 2.20 all other parameters remain fixed. The learning rate in this case is kept constant at $\eta = 2 \times 10^{-8}$. Since the evolution of the voltage probability distribution depends on the exact sequence of randomly drawn inputs, the figure shows the result of eq. 2.38 averaged over 20 runs, starting with identical initial conditions. The curves correspond to a geometric progression in time, each successive curve representing twice as many iterations of the learning rule as the one before. The initial curve (solid line) represents $t = 4$, or the fourth iteration of the learning rule. The arrows indicate the general trend in time: the Ca^{2+} curves, for instance, form snapshots of a "wave" traveling to the right along the voltage axis.

zero, or negative, because of the half-wave rectification in the adaptation mechanism for peak conductances. The last curve in fig. 2.21a, for instance, shows a persistent negative gradient for conductance midpoint voltages near or below the resting potential (at 0 mV); however, no real nonzero conductances remain within that voltage range in the learning process.

2.10 The Hodgkin-Huxley Spiking Model

Are learning rules for information maximization compatible with realistic models of spiking neurons, such as the Hodgkin-Huxley model? One theoretical approach is to let the cell change the Ca^{2+} and K^{+} conductance parameters to modulate the current delivered to a set of fixed spiking conductances, as given, for instance, by the Hodgkin-Huxley equations. The other approach is to make all conductance parameters subject to adaptation, without reserving a predetermined set of conductances dedicated to the spike mechanism: can a non-spiking cell learn to spike? The answer to this last question is yes, but we shall defer consideration of the abiogenesis of spiking to the next section. For the moment, we simply wish to point out that the two approaches are fundamentally different, each fraught with its own set of difficulties.

We consider the following scenario: Human beings and other primates make rapid eye movements known as saccades about three times a second. A saccade typically lasts from 30 to 70 milliseconds (Becker, 1989), which leaves an average of 230-300 milliseconds during which the gaze is fixed. Therefore, in the time between saccades, the mean luminance across the visual receptive field of a neuron in visual cortex

remains constant. Since the visual response (at least in the magnocellular stream) is suppressed during the saccade, viewing a static natural scene is tantamount to visual neurons receiving a train of quasi-static stimuli, each stimulus lasting several hundred milliseconds, followed by another stimulus, then by another.

Within the time scale of the stimulus, we will presume that, after an initial transient response to the stimulus onset, the cell will adapt to the stimulus and settle into a periodically spiking state. Such a cell would belong to the class of "regularly" spiking cells, whose response can be characterized by the rate of action potentials. In mathematical terms, what does it mean for a neuron to be in the firing rate regime? In essence, the duration τ_{duration} of the stimulus must be much greater than the period T between spikes, i.e., each stimulus must elicit many spikes:

$$\tau_{\text{duration}} \gg T.$$

After an initial transient in response to each stimulus, we assume that the voltage waveform $V(t)$ settles into a simple periodic limit cycle as dictated by the somatic spiking conductances. This steady-state period of $V(t)$, or firing frequency, is then the relevant function of the stimulus. As mentioned in the overview, the firing rate of a neuron is but one representation of information, as opposed to spike time encoding, for instance. If the condition $\tau_{\text{duration}} \gg T$ is not met, then the question of whether information is encoded in the timing of individual spikes is more appropriate. We will not, however, treat this question here.

The method of *averaging* assumes that the firing rate f can be written as a func-

tion of the mean current $\langle I \rangle$ reaching the somatic compartment averaged over the period T between successive spikes. Averaging allows one to make direct use of the strict physical constraint of charge conservation: if the dynamical system is in a periodic state, the average current injected by the synaptic and voltage-dependent conductances over one period of the firing cycle must be equal to the average current discharged by the neuron. In averaging, we assume that the firing rate $f(I)$ only depends on the mean current averaged over the firing period:

$$\langle I \rangle = \frac{1}{T} \int_0^T I(t)\, dt, \qquad \text{where } I(t) \text{ is periodic with period } T.$$

We thus write f as $f(\langle I \rangle)$.

We assume that the time-dependence of the stimuli can be described as a series of square pulses of width Δt in time:

$$g_{\text{syn}}(t) = \sum_n g_{\text{syn}}(n) \Theta\big[t - n\Delta t\big] \Theta\big[(n+1)\Delta t - t)\big],$$

where each $g_{\text{syn}}(n)$ is an input conductance drawn randomly from a fixed probability distribution. A particular input conductance $g_{\text{syn}} \in (a, b)$ gets mapped onto a unique function $V(t)$ of unique period T that describes the membrane voltage. The averaged current $\langle I \rangle$ is a functional of $V(t)$ and T, whereas the firing rate f is a bijective function from the variable $\langle I \rangle$ on the real line to the positive real line. Note that throughout we require f to be a *bounded* function. We thus have a composition of

mappings that transforms a particular g_{syn} into a firing rate:

$$(a,b) \subset (R^1_+ \cup \{0\}) \rightarrow T \times \mathcal{L}^2[0,T] \rightarrow R^1 \rightarrow (R^1_+ \cup \{0\}),$$

where R^1_+ is the space of positive real numbers, and $\mathcal{L}^2[0,T]$ the space of square integrable real functions of period T.

Note that $\langle I[V(t)] \rangle$ is a *functional*, i.e., a mapping from the space of functions (that describe $V(t)$) onto the real line. In the previous sections, we dealt only with functions relating voltage to current, without considering any implicit time-dependence. The term in the learning rule (eq. 2.31) that read

$$\Delta q|_x = \eta \frac{\partial}{\partial q} \ln \left(\frac{dI}{dV} \right)$$

now reads

$$\frac{dq}{dt} = \eta \frac{\partial}{\partial q} \ln \left(\frac{\delta \langle I \rangle}{\delta V(t)} \right), \tag{2.39}$$

where $\dfrac{\delta \langle I \rangle}{\delta V(t)}$ is a functional derivative.

If the cell adjusts the parameters of its conductances to modulate the current delivered to the spike initiation zone, two questions must first be addressed before delving into the issue of learning:

1) What is the functional form of the firing rate $f(I)$ as a function of the current reaching the nerve cell's spike initiation zone?

2) The learning rule exploits the linearity of a fictitious conductance g_{soma} in fig. 2.12. If we replace this conductance by a set of spiking conductances for sodium and potassium, will the effective conductance still be linear?

The answers to these questions turn out to be quite simple, even if highly non-trivial. In section 2.11, we show, by using the theory of dynamical systems, that the adapted firing rate for most generic Hodgkin-Huxley spiking models will be linear. Also in section 2.11, we will argue that the relationship between the mean current $\langle I \rangle$ and the average voltage $\langle V \rangle$ is generally quadratic. By extending the neuron to two compartments separated by a coupling conductance G, one compartment representing the dendrites, the other the soma, an approximately linear relationship between $\langle I \rangle$ and $\langle V \rangle$ is recovered. For our purposes, the two-compartment model, rather than the single compartment model, constitutes the *minimal* model for studying the learning of modulatory voltage-dependent conductance parameters.

When a linear relationship between $\langle I \rangle$ and $\langle V \rangle$ results because of charge conservation, the functional derivative $\dfrac{\delta \langle I \rangle}{\delta V(t)}$ becomes a constant.

To compute the partial derivatives in eq. 2.39, we need to examine the *average* contribution of the time-dependent active conductances for Ca^{2+} and K^+ to the mean current $\langle I[V(t)] \rangle$. This depends on the average activation

$$\frac{1}{\tau_{\text{duration}}} \int_0^{\tau_{\text{duration}}} m[V(t)]\, dt \approx \frac{1}{T} \int_t^{t+T} m[V(t)]\, dt = \langle m[V(t)] \rangle,$$

where the RHS is the integral over one period of the quasi-steady-state of the system. Any transient behavior of the conductance, including adaptation over short time

scales, is averaged out by taking $\tau_{\text{duration}} \gg T$. In the previous formulation of the learning rule, the steady-state $m_{\infty,i}(V)$ activation functions formed a basis set. Here, the averaged activation functions $\langle m_i(V) \rangle$ serve this role.

The functional derivatives simplify significantly if the time constants $\tau_i(V)$ are, in reality, voltage-independent. In this case, integrating both sides of the differential equation for $m(t)$ results in

$$\frac{1}{T}\int_0^T \tau \frac{dm}{dt} = \frac{1}{T}\int_0^T m_\infty[V(t)]\,dt - \frac{1}{T}\int_0^T m(t)\,dt.$$

Since the LHS is zero, it follows that

$$\langle m[V(t)]] \rangle = \frac{1}{T}\int_0^T m_\infty[V(t)]\,dt$$

The average $\langle m[V(t)]] \rangle$ is thus independent of the value of τ, the conductance time constant. We are free to choose any reasonable τ.

In passing, we point out that the set of "basis functions" $\langle m_i \rangle$ is not equivalent to the basis set used in the previous model, since

$$\langle m_\infty[V(t)]] \rangle \neq m_\infty[\langle V(t) \rangle].$$

Instead,

$$\langle m_\infty[V(t)]\rangle < m_\infty[\langle V(t)\rangle] \qquad \text{in regions where } m_\infty[\langle V(t)\rangle] \text{ is convex}$$

$$\langle m_\infty[V(t)]\rangle > m_\infty[\langle V(t)\rangle] \qquad \text{in regions where } m_\infty[\langle V(t)\rangle] \text{ is concave,}$$

leading to the overall behavior that $\langle m_\infty[V(t)]\rangle$ is a shallower function of $\langle V(t)\rangle$ than $m_\infty[\langle V(t)\rangle]$. As a general rule of thumb, steeper activation functions are better at shaping the neuron's firing rate response to match the optimal response curve. (This statement can be made more quantitative—see, for example Barron (1993).)

We assume throughout that $m_\infty(V)$ is described by a Boltzmann function of the type $m_\infty(V) = 1/\{1+\exp[-s(V-V_{\frac{1}{2}})\}$, where V is the transmembrane voltage, s is the inverse slope, and $V_{\frac{1}{2}}$ is the midpoint voltage of half-activation. If $\tau(V) =$ constant, then applying the definition of the functional derivative to $\langle m_i[V(t)]\rangle$ is straightforward:

$$\frac{\delta}{\delta V(t)}\langle m_i(V)\rangle \doteq \lim_{\varepsilon\to 0}\frac{1}{T}\left\{\frac{\int_0^T m_{\infty,i}[V(s)+\varepsilon\delta(s-t)]\,ds - \int_0^T m_{\infty,i}[V(s)]\,ds}{\varepsilon}\right\}$$

$$= \begin{cases} \dfrac{s_i}{T}\left(m_{\infty,i}[V(t)] - m_{\infty,i}[V(t)]^2\right) & \text{if } 0 < t < T \\ 0 & \text{if } t < 0 \text{ or } t > T \\ \dfrac{s_i}{2T}\left(m_{\infty,i}[V(0)] - m_{\infty,i}[V(0)]^2\right) & \text{if } t = 0 \\ \dfrac{s_i}{2T}\left(m_{\infty,i}[V(T)] - m_{\infty,i}[V(T)]^2\right) & \text{if } t = T \end{cases}$$

where the last two cases follow by integrating with the Dirichlet kernel $\delta_n(t) =$

$1/2\pi\ \sin[(n+1/2)t]/\sin[1/2\ t]$ or the kernel $\delta_n(t) = n/\pi\ (1+n^2t^2)^{-1}$, and letting $n \to \infty$. For $\tau(V) \neq$ constant, we still need to take the functional derivative, even though it will no longer take such a simple form as above.

If we now integrate the functional derivative over the duration of the stimulus, assuming $\tau_{\text{duration}} \gg T$,

$$\frac{1}{\tau_{\text{duration}}}\int_0^{\tau_{\text{duration}}} \frac{\delta}{\delta V(t)}\langle m_i[V(t)]\rangle\, dt = s_i\left(\langle m_{\infty,i}(V)\rangle - \langle m_{\infty,i}(V)^2\rangle\right)$$
$$= \left(\left\langle \frac{d}{dV} m_{\infty,i}(V)\right\rangle\right).$$

In the previous time-independent description, we implicitly assumed that each randomly selected stimulus is presented for the same length of time. We can relax this assumption to some degree, so that the weight a stimulus given is equal to the product of duration and frequency with which the stimulus is presented. The condition, of course, is that the timescale of the longest-duration stimulus be much smaller than the timescale of learning.

Integrating over a standard stimulus duration, we recover a learning rule similar to eq. 2.31, at least if we restrict ourselves to conductances that simply activate:

$$\Delta\bar{g}_i = \eta(t)\left[\left\langle \frac{d}{dV} m_i(E_i - V)\right\rangle + c(\langle V\rangle)\left\langle m_i(E_i - V)\right\rangle\right] \tag{2.40}$$

Fig. 2.27 on page 123 shows the result of applying this learning rule to a probability distribution of conductance inputs in a Hodgkin-Huxley model with two compartments.

2.11 Effective Conductance and Current-Discharge Relationships for Hodgkin-Huxley Spiking Models with Adaptation*

The goal of this section is to relate three quantities to each other by appealing to the theory of dynamical systems: the firing rate, the mean voltage, and the average current in a spiking Hodgkin-Huxley model. These relationships will be needed to estimate $d\langle I\rangle/d\langle V\rangle$ required in the stochastic approximation learning rule. Initially, we consider a neuron that consists of a single electrotonic compartment.

The typical behavior of a Hodgkin-Huxley relaxation oscillator consists of charge building up followed by a sudden and stereotypical voltage discharge that is the action potential. To capture the basic behavior of the voltage as a function of time, we can choose a simplified model in which the voltage ramps linearly in time up to a threshold; when the threshold voltage is reached, the voltage is constrained to undergo a stereotypical spike, which resets the voltage to zero. This simplified model's *sole* purpose, as illustrated in fig. 2.22, is to make imminently clear the qualitative relationship between the firing rate and the mean voltage averaged over the firing period:

In most spiking models, the duration of the spike, t_{spike}, remains fairly constant as the firing frequency $f = 1/T$ changes. Thus $\langle V_{\text{spike}}\rangle$, the average voltage *during*

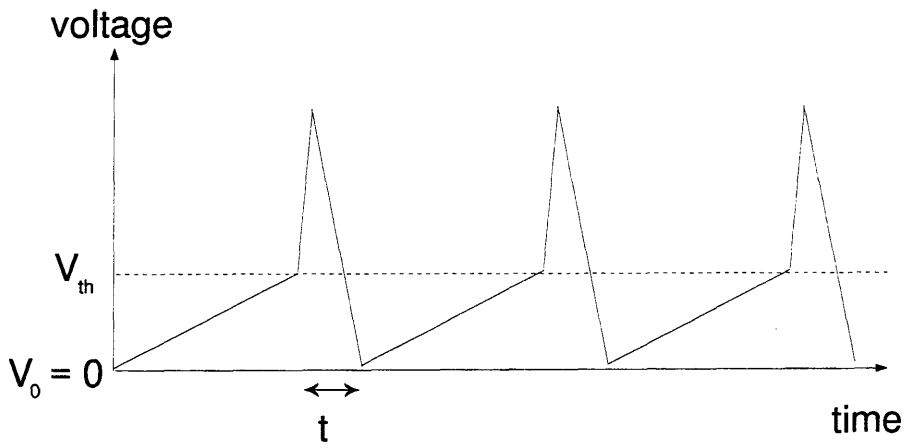


Figure 2.22: Schematic simplification of a spiking cell's voltage trace

the time t_{spike}, is constant. Qualitatively, then,

$$\begin{aligned}\langle V \rangle &= \frac{1}{T}\int_0^T V(t)\,dt \\ &= \frac{1}{T}\Big[(T - t_{\text{spike}})\,1/2\,V_{th} + t_{\text{spike}}\langle V_{\text{spike}}\rangle\Big] \\ &= 1/2\,V_{th} + f\,t_{\text{spike}}\Big(\langle V_{\text{spike}}\rangle - 1/2\,V_{th}\Big),\end{aligned}$$

Therefore, $\langle V \rangle$ is *linear* in f.

We can thus write, to a good approximation, a relationship valid for most models of spiking cells:

$$\langle V \rangle = af + V_0 \tag{2.41}$$

To relate the spike-averaged voltage to the spike-averaged current, we cannot use an integrate-and-fire model: in such a model, the mean voltage *decreases* as the mean current increases, which is quite the opposite behavior to that of a Hodgkin-Huxley model. In fact, one of the requirements for learning to be feasible is that the current

increases with increasing voltage, such that periodic fixed points of the dynamics exist. Therefore, an integrate-and-fire model is not a valid simplification of the dynamical system of Hodgkin-Huxley equations for our purposes.

To determine the current-discharge relationship that describes the transduction from the current reaching the spike initiation zone into a firing rate f, we borrow a deep result from the theory of dynamical systems (Guckenheimer and Holmes, 1983; Strogatz, 1991):

For non-degenerate saddle-node bifurcations of dynamical systems, the set of differential equations for the Hodgkin-Huxley model can be reduced to a one-dimensional equation near the transition from a steady-state voltage to periodic firing. This transition to sustained oscillations occurs when $dI/dV|_{V^*}$ of eq. 2.35 on page 79 becomes zero as the injected current increases to $I_{\text{threshold}}$. For currents near the threshold, the dynamics on the 'center manifold' are given by

$$\dot{x} = -(I - I_{\text{threshold}}) - x^2$$

This equation is also known as the canonical form of the saddle-node bifurcation. For currents above threshold, the period of the voltage oscillation scales as (Strogatz, 1991):

$$\begin{aligned} \frac{1}{f} &\sim \int_{-\infty}^{\infty} \frac{1}{(I - I_{\text{threshold}}) + x^2}\, dx \\ &\sim (I - I_{\text{threshold}})^{-\frac{1}{2}} \end{aligned}$$

where we write the period of the oscillation as the inverse of the firing frequency f.

Applying the method of averaging to the current, we can write:

$$f = b\sqrt{\langle I \rangle - I_{\text{threshold}}} \qquad \text{for } \langle I \rangle > I_{\text{threshold}} \tag{2.42}$$

In the language of physics, this is a second-order transition. Note that this is a

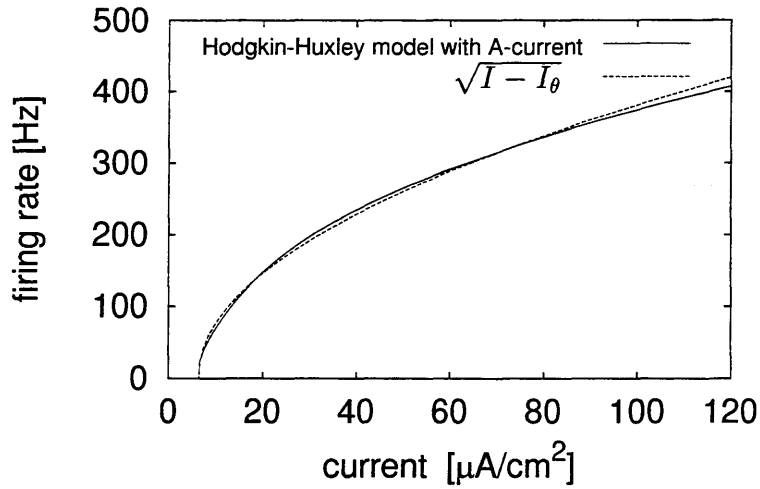


Figure 2.23: Firing frequency as a function of synaptic input conductance for the Hodgkin-Huxley model of Connor *et al.* (1977), shown with a square root nonlinear fit. A square-root $f(I)$ relationship is a consequence of a saddle-node bifurcation (see text).

universal result, and is not restricted to any particular Hodgkin-Huxley model with a particular set of conductances. The only requirement is that the bifurcation must be of the type that allows arbitrarily low firing rates. A plot of the reciprocal of the first ISI versus current injected into neurons in the central nervous system will typically have a square root functional shape. Adaptation currents will, of course, modify this fundamental form of the $f(I)$ curve, as we will show below. Combining eqs. 2.41 and

2.42, we have

$$c\sqrt{\langle I\rangle - I_{\text{threshold}}} = \langle V\rangle - V_0$$

or

$$\langle I\rangle = \frac{\Big(\langle V\rangle - V_0\Big)^2}{c^2} + I_{\text{threshold}} \qquad \langle V\rangle > V_0. \tag{2.43}$$

In other words, $\langle I\rangle$ is a quadratic polynomial in $\langle V\rangle$.

To show that $\langle I\rangle$ remains a quadratic polynomial even when adaptation is added, we need to first look at the effect of adaptation on the $f(I)$ curve. It is simple to show that adaptation linearizes the $f(I)$ curve by making the following *Ansatz*:

Assume that the adaptation current is purely subtractive, and that we can replace it by its mean value:

$$\begin{aligned} I_{\text{adapt}} &= g_{\text{K(Ca)}}\, f\, \tau_{\text{K(Ca)}}\Big(\langle V\rangle - E_{\text{K}}\Big) \\ &= k\, f + h\, f^2, \end{aligned}$$

where, using the relationship between f and $\langle V\rangle$ from eq. 2.41,

$$\begin{aligned} k &= g_{\text{K(Ca)}}\, \tau_{\text{K(Ca)}}\Big(V_0 - E_{\text{K}}\Big) \\ h &= a\, g_{\text{K(Ca)}}\, \tau_{\text{K(Ca)}}. \end{aligned}$$

In other words, the adaptation current is the product of the K(Ca) or AHP conductance opened during or after each spike, the frequency of spiking, the time constant τ of this conductance, and the average driving potential, which is the difference between the reversal potential for potassium and the average voltage $\langle V \rangle - E_{\mathrm{K}}$. The typical time constant of adaptation in neocortical pyramidal cells is roughly $\tau = 50$ msec, as estimated by Ahmed *et al.* (1995). It is essential here that the time constant be greater than the average interspike interval $\tau \gg T_{\mathrm{ISI}}$.

Solve for the steady state of

$$f(\langle I \rangle - I_{\mathrm{adapt}}) = b\sqrt{\langle I \rangle - \left(I_{\mathrm{adapt}} + I_{\mathrm{threshold}}\right)} = b\sqrt{(\langle I \rangle - I_{\mathrm{threshold}}) - (k\,f + hf^2)}$$

$$f = -\frac{(b')^2 k}{2} + \sqrt{\frac{(b')^4 k^2}{4} + (b')^2 (I - I_{\mathrm{threshold}})}, \tag{2.44}$$

where

$$(b')^2 = \frac{b^2}{(1 + hb^2)}.$$

In the limit as $(\langle I \rangle - I_{\mathrm{threshold}}) \to 0$,

$$f \sim \frac{\langle I \rangle - I_{\mathrm{threshold}}}{k} \tag{2.45}$$

with local curvature around the threshold of

$$f'' = -\frac{2}{(b')^2 k^3} = -\frac{2}{k^3}\Big(\frac{1}{b^2} + h\Big)$$
$$= \mathcal{O}\Big(\big(g_{\mathrm{K(Ca)}}\,\tau_{\mathrm{K(Ca)}}\big)^{-2}\Big).$$

Thus the stronger the adaptation factor, the greater the linearization, since the curvature measures the deviation from linearity. As long as b is large, eq. 2.45 will be a valid approximation. Thus, even though b is a quantity which can be directly (and laboriously) calculated via analysis of the local bifurcation, the value of b does not play an important role in the description of the adapted firing rate.

In the opposite limit of high currents, as $(\langle I\rangle - I_{\mathrm{threshold}}) \to \infty$

$$f = -\frac{(b')^2 k}{2} + b'\sqrt{\langle I\rangle - I_{\mathrm{threshold}}} + \frac{(b')^4 k^2}{8b'\sqrt{\langle I\rangle - I_{\mathrm{threshold}}}}.$$

Combining equations 2.44 and 2.41, we have

$$\langle I\rangle = \frac{1}{(ab)^2}\Big(\langle V\rangle - V_0\Big)^2 + \frac{1}{a} g_{\mathrm{K(Ca)}}\tau_{\mathrm{K(Ca)}}\Big(\langle V\rangle - E_{\mathrm{K}}\Big)\Big(\langle V\rangle - V_0\Big) + I_{\mathrm{threshold}} \tag{2.46}$$

where $g_{\mathrm{K(Ca)}}$ is the adaptation conductance opened during each spike, and $\tau_{\mathrm{K(Ca)}}$ is the exponential decay time of the adaptation conductance. We thus still have the result that $\langle I\rangle$ is a quadratic polynomial in $\langle V\rangle$. Note that even though the current-discharge is *linear*, the neuron is by no means a linear device!

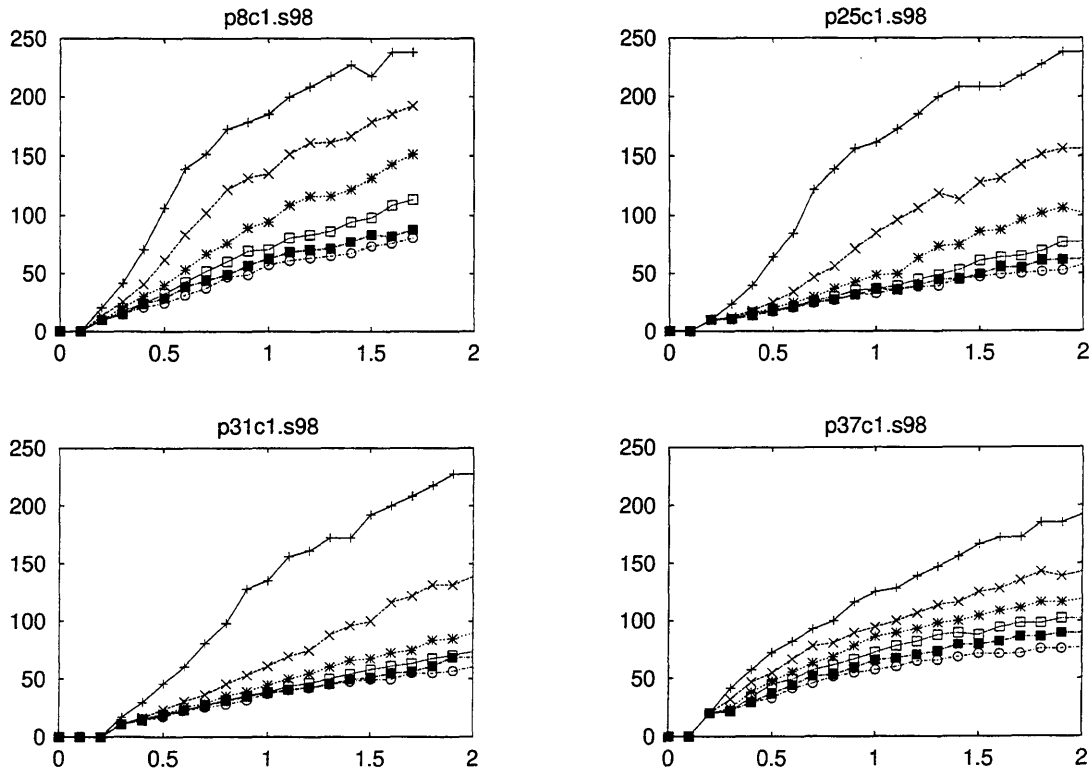


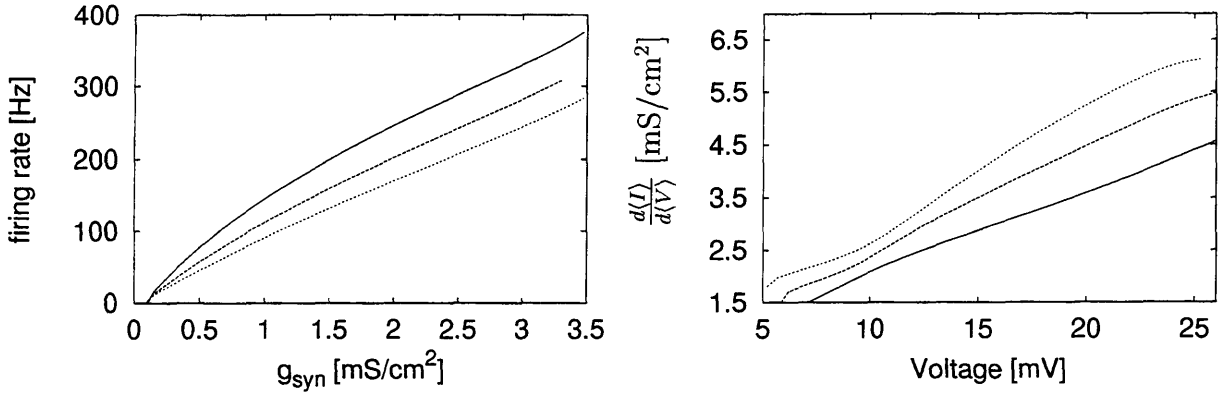
Figure 2.24: Current-discharge relationships of *in vitro* cells from cat area 17, plotting the inverse of successive interspike intervals (first through sixth ISI) at different values of current injection in nA. As adaptation effects accumulate, the firing rate changes from a "square-root"-like function to a linear function. Data courtesy of N. Berman, graph by G. Holt.

The derivative of the average current with respect to the average voltage is

$$\frac{d\langle I\rangle}{d\langle V\rangle} = \frac{2}{a^2}\Big(\frac{1}{b^2} + h\Big)\Big(\langle V\rangle - V_0\Big) + \frac{k}{a} \tag{2.47}$$

Increasing the adaptation current increases both the DC level and the slope of the effective conductance from the somatic compartment to ground.

By redefining $I_{\text{threshold}}$ and $\langle V_0\rangle$ as follows, we can rewrite eq. 2.44 in the form of



(a) (b)

Figure 2.25: A mean-field approach casts the adapted current-discharge (f-I) relationship into an analytical and tractable form. The same *linear* relationship holds if we plot the adapted firing rate versus the synaptic conductance g_{syn}, as shown in fig. (a). The constant conductance g_{soma} of fig. 2.12 is replaced by an (approximately) affine linear function of the mean voltage $\langle V\rangle$, as illustrated in fig. (b) and given by eq. 2.47.

eq. 2.42:

$$I'_{\text{threshold}} = I_{\text{threshold}} - \frac{(b')^2 k^2}{4}$$

$$V'_0 = V_0 - \frac{a(b')^2 k}{2}$$

Using this linear approximation to the derivative presumes that the biological learning mechanism has access to the values of the slope and offset of this function. The primary disadvantage, however, is that the dynamic range of possible mean voltages $\langle V\rangle$ is limited, as compared to the average slope of typical activation functions, which are in the range of 10 mV^{-1}. Another, relatively minor, disadvantage is that the linear approximation to $d\langle I\rangle/d\langle V\rangle$ breaks down near the threshold voltage.

Suppose now that we add a dendritic compartment separated by a resistance R

from the somatic compartment. Then the current injected into the somatic compartment will be:

$$\begin{aligned}\langle I\rangle &= \frac{\langle V_d - V_s\rangle}{R} \\ &= \frac{1}{R}\Big(\langle V_d\rangle - V_0' - c\sqrt{\langle I\rangle - I'_{\text{threshold}}}\Big) \\ &= \frac{\langle V_d\rangle - V_0'}{R} + \frac{c^2}{2R^2}\left\{1 - \sqrt{1 + \frac{4R^2}{c^2}\Big(\frac{\langle V_d\rangle - V_0'}{R} - I'_{\text{threshold}}\Big)}\right\},\end{aligned}$$

where $c = ab'$. In the limit $R \to 0$ and $c \gg R$,

$$\langle I\rangle \approx \frac{(\langle V_d\rangle - V_0')^2}{c^2} + I'_{\text{threshold}} + \mathcal{O}\Big(R, \frac{R}{c^2}\Big)$$

If $R \gg c$ and $\langle I\rangle \gg I'_{\text{threshold}}$,

$$\begin{aligned}\langle I\rangle \approx& \frac{\langle V_d\rangle - V_0'}{R} - \frac{c}{R}\sqrt{\frac{\langle V_d\rangle - V_0'}{R} - I'_{\text{threshold}}} \\ \frac{d\langle I\rangle}{d\langle V_d\rangle} \approx& \frac{1}{R} - \frac{c}{2R^{\frac{3}{2}}}(\langle V_d\rangle - V_0' - RI'_{\text{threshold}})^{-\frac{1}{2}}.\end{aligned} \tag{2.48}$$

By separating the modulatory Ca^{2+} and K^+ conductances from the spiking conductances, i.e., by placing the two sets of conductances into different electrotonic compartments, we recover an approximation to the simple model in fig. 2.12. In essence, we will write $\frac{d\langle I\rangle}{d\langle V_d\rangle} \approx$ constant.

2.12 Hodgkin-Huxley Spiking Model Simulations

To implement information maximization learning rules in spiking neurons, a simplified model of a neuron consisting of two electrotonic compartments, illustrated in fig. 2.26, was constructed. The soma (or cell body) contains the classic Hodgkin-Huxley sodium and delayed rectifier conductances with the addition of a transient potassium "A"-current and an adaptation current. The equations for the two compartments, the soma and the dendrite, are:

$$\begin{aligned} C\frac{dV_{\text{soma}}}{dt} =& G\left(V_{\text{dendrite}} - V_{\text{soma}}\right) \\ &+ \bar{g}_{\text{Na}}\, m^3 h\left(E_{\text{Na}} - V_{\text{soma}}\right) \\ &+ \bar{g}_{\text{K,DR}}\, n^4\left(E_{\text{K}} - V_{\text{soma}}\right) \\ &+ \bar{g}_{\text{K,A}}\, a^3 b\left(E_{\text{K}} - V_{\text{soma}}\right) \\ &+ \bar{g}_{\text{adapt}}\left(E_{\text{K}} - V_{\text{soma}}\right) \\ &+ g_{\text{memb}}\left(E_{\text{Cl}} - V_{\text{soma}}\right) \end{aligned} \tag{2.49}$$

$$\begin{aligned} C\frac{dV_{\text{dendrite}}}{dt} =& \bar{g}_{\text{syn, LGN}}\left(E_{\text{syn, LGN}} - V_{\text{dendrite}}\right) \\ &+ \sum_i \bar{g}_{\text{Ca},i}\, m_i(V_{\text{dendrite}})\left(E_{\text{Ca}} - V_{\text{dendrite}}\right) \\ &+ \sum_j \bar{g}_{\text{K},j}\, n_j(V_{\text{dendrite}})\left(E_{\text{K}} - V_{\text{dendrite}}\right) \\ &+ g_{\text{memb}}(E_{\text{Cl}} - V_{\text{dendrite}}) \end{aligned} \tag{2.50}$$

$$\tau_x(V)\frac{dx}{dt} = x_\infty(V) - x(V) \qquad x = a, b, h, m, n, m_i, n_j \tag{2.51}$$

All parameters for the somatic compartment, with the exception of the adaptation conductance, are given by the standard model of (Connor *et al.*, 1977). This choice of somatic spiking conductances allows spiking to occur at arbitrarily low firing rates, as is typically observed in cortical cells. As discussed in the previous section, the voltage dynamics undergo a saddle-node bifurcation from a steady state voltage to a periodic limit cycle as the synaptic conductance is increased. Saddle-node bifurcations are a universal class of dynamical bifurcations typified by the following current-discharge ("f(I)") relationship in the absence of adaptation: $f(\langle I \rangle) = b\sqrt{\langle I \rangle - I_\theta}$, where b is a constant, and I_θ is a threshold current.

Firing frequency adaptation is modeled as a steeply voltage-dependent process with a long time constant:

$$\frac{dg_{\text{adapt}}}{dt} = \frac{1}{\tau_{\text{adapt}}}\left(-g_{\text{adapt}} + \frac{g_{\text{adapt, max}}}{1 + \exp[(V_{\frac{1}{2}} - V)/k]}\right),$$

with $g_{\text{adapt, max}} = 50$ mS/cm^2, $\tau_{\text{adapt}} = 50$ msec, $V_{\frac{1}{2}} = -10$ mV, and $k = 0.5$ mV. The time constant of 50 milliseconds is based on the work of Ahmed *et al.* (1997) in cat visual cortex. Because of the extremely steep voltage dependence, the adaptation conductance g_{adapt} only increases during an actual action potential. This behavior models a calcium-dependent potassium conductance without explicitly representing the calcium concentration inside the somatic compartment; in a real cell, calcium ions enter the cell during action potential through high-voltage activated channels, thereby activating Ca(K) channels. Because of the long time constant in the decay of the adaptation conductance, adaptation will be roughly proportional to the firing

frequency of the Hodgkin-Huxley model: $I_{\text{adapt}} = b'f$, where b' is the constant of proportionality. As we showed in section 2.11, the effect of an adaptation current with a long time constant is to linearize the firing frequency as a function of input current: $f(\langle I \rangle) = (\langle I \rangle - I_\theta)/b'$. Note that b' here is unrelated to the constant b in the unadapted firing rate ($f(\langle I \rangle) = b\sqrt{\langle I \rangle - I_\theta}$).

A coupling conductance $G = 1.0$ mS/cm^2 connects the somatic and dendritic compartment. Each compartment also contains a fixed chloride conductance $g_{\text{memb}} = 0.3$ mS/cm^2, as in the original Hodgkin-Huxley model. The reversal potential for Ca^{2+} is 130 mV above resting potential, similar to that of Na^+. To drive current from the dendritic compartment into the somatic one, the synaptic reversal potential was set to 50 mV above the leak (chloride) reversal potential.

For simplicity, we model the adjustable calcium and potassium conductances in the dendritic compartment as having a single activation function (with exponent $p_i = 1$) and no inactivation function. In addition, the time constant for these conductances is taken to be voltage-independent and equal to 10 milliseconds. Further, we insist that the steady state of these conductances is consistent with physical equilibrium, such that the steady state activation function $m_{\infty,i}(V)$ for the i-th conductance is given by a Boltzmann function: $m_{\infty,i}(V) = 1/\{1 + \exp[-(V - V_i)/s_i]\}$, where s_i is the slope of the activation function and V_i is the midpoint voltage of half-activation. The initial conditions were such that the peak values of all modulatory conductances were set to zero. The midpoint voltages for the Ca^{2+} conductances were spaced evenly between 35 and 75 mV above resting potential; for the K^+ conductances, the midpoints were

spaced between 5 and 40 mV above the resting potential. The slope for all modulatory voltage-dependent conductances was initially set to 6.7 mV.

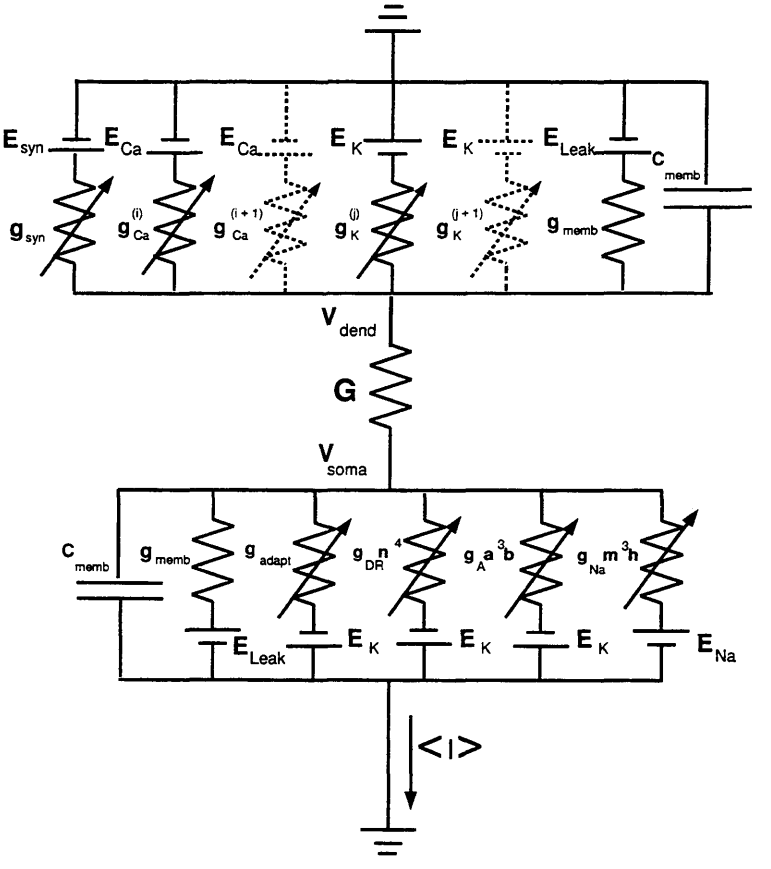


Figure 2.26: The Hodgkin-Huxley model neuron contains two compartments to represent the cell's soma and dendrites. For the sake of simplicity, both compartments are of equal size and are connected by a coupling conductance $G = 1.0$ mS/cm^2. To maximize the information transfer, the parameters for six calcium and six potassium voltage-dependent conductances in the dendritic compartment are varied, while the somatic conductances responsible for the cell's spiking behavior are held fixed.

The learning rules for changing the conductances in periodically spiking systems can be written down as a simple generalization of those in section 2.7 for steady state voltages:

$$\Delta\bar{g}_i = \eta(t)\left[\left\langle \frac{d}{dV}\Big\{m_i\,[E_i - V(t)]\Big\}\right\rangle + c(\langle V\rangle)\left\langle m_i\,[E_i - V(t)]\right\rangle\right], \tag{2.40}$$

where the angular brackets indicate an average of the quantities inside the brackets over one interspike interval of the periodically spiking Hodgkin-Huxley model. The term dm_i/dV is a shorthand notation for the functional derivative $\delta/\delta V(t)\langle m_i(t)\rangle$; if the time constant is voltage-independent, then $\langle dm_i/dV\rangle = \langle dm_{\infty,i}(V)/dV\rangle$, so the notational abuse of interchanging a functional derivative with an ordinary derivative is not too grave a crime.

Since the firing rate is nearly linear in the average current reaching the soma, we use the simple constraint function of eq. 2.32 to limit the dendritic voltages:

$$c(\langle V\rangle) = \begin{cases} \gamma & \text{if } \langle V\rangle < 25 \text{ mV above rest} \\ 0 & \text{if } 25 \text{ mV} \leq V \leq 60 \text{ mV} \\ -\gamma & \text{if } V > \text{mV} \end{cases}$$

with $\gamma = 2.5$.

Purely numerical considerations motivate the restriction of the spiking Hodgkin-Huxley model to simple activation functions and voltage-independent time constants. In the most general case, the numerical computation of the functional derivatives requires the solution of an auxiliary differential equation for each parameter being changed. To show that we do rush in where angels fear to tread, we will bite the bullet in the last section, not to mention mix metaphors, and actually simulate a coupled set of 109 differential equations to change the parameters of twelve conductances.

With the restrictions mentioned above, we need make only one approximation to numerically implement the adaptation mechanism using only the original system of

differential equations. In this approximation,

$$\langle m_i \, E_i - V(t) \rangle \approx \langle m_i \rangle \langle E_i - V(t) \rangle$$

for the purpose of taking the functional derivative with respect to $V(t)$. If the time constant τ of the conductance is long compared to the fast timescales in the Hodgkin-Huxley spiking mechanism, then m_i is fairly constant over the spiking period, and the approximation is justified. When $|E_i| \gg |V(t)|$, any approximation will do.

In fig. 2.27, the learning rule of eq. 2.40—generalized to also change the midpoint voltage and steepeness of the activation functions—is applied to a Gaussian distribution of synaptic conductance inputs to the model. A synaptic input is drawn at random from the probability distribution every 400 milliseconds, such that the stimulus train is a step-like sequence of pulses. The adapted firing rate is measured by taking the inverse of the last interspike interval within each 400 millisecond period. With this training regimen, the cell "learns" the statistical structure of the input, matching its adapted firing rate to the cumulative distribution function of the conductance inputs (fig. 2.27a). The distribution of firing rates shifts from an initially peaked distribution to a much flatter one, so that all firing rates are used nearly equally often (fig. 2.27b). To emphasize the fact that learning has occurred even in the absence of steady state voltages, we display the time course of the voltage in the dendrites as insets to fig. 2.27a—note the spikes.

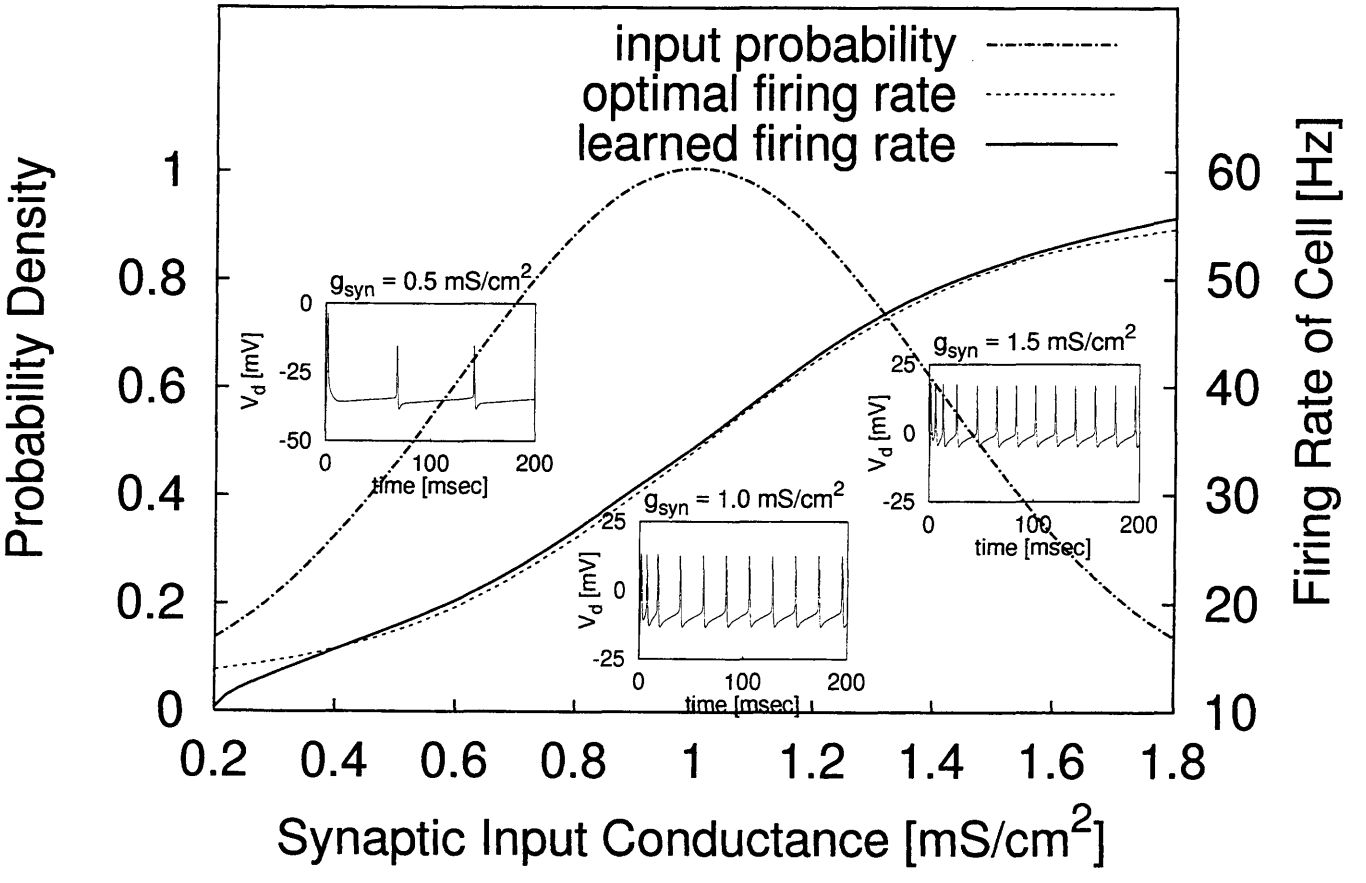
The adapted peak conductances, midpoint voltages, and activation slopes for the Ca^{2+} and K^+ conductances are listed in tables 2.2 and 2.3, respectively.

Table 2.2 Adapted parameters of the voltage-dependent Ca^{2+} conductances for fig. 2.27

peak conductance $\bar{g}_i/g_{\text{leak}}$	midpoint voltage $V_{1/2}$ [mV]	slope s [mV]
0.099	25.4	4.29
0.284	30.4	4.31
0.447	30.9	5.11
0.657	39.8	11.5
2.02	54.4	8.27
2.46	54.3	7.80

Table 2.3 Adapted parameters of the voltage-dependent K^+ conductances for fig. 2.27

peak conductance $\bar{g}_i/g_{\text{leak}}$	midpoint voltage $V_{1/2}$ [mV]	slope s [mV]
2.06	23.5	2.69
1.94	23.5	2.68
0.350	32.1	3.96
0.036	42.5	3.97
0	51.3	4.02
0	58.9	3.96

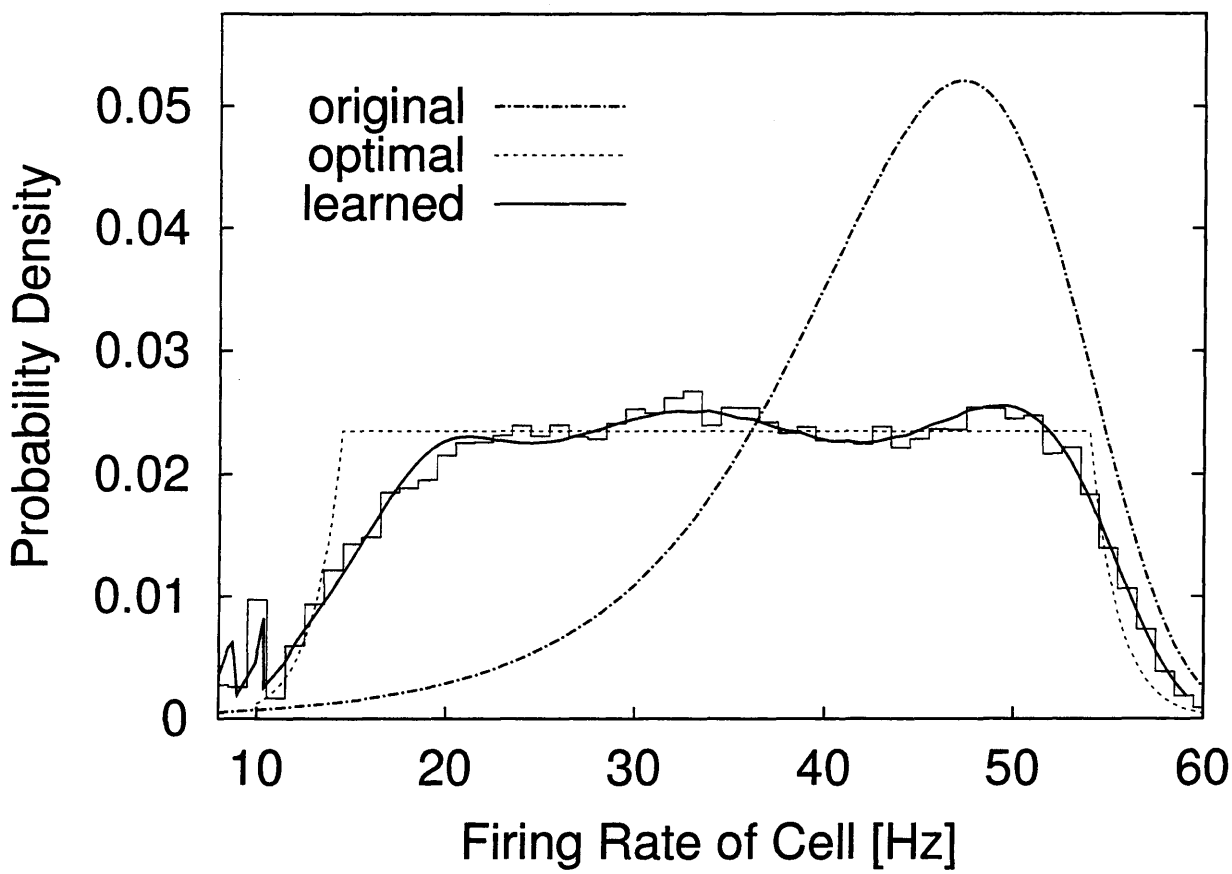


(a)

Quantifying the entropy of the distribution of firing rates before and after adaptation in fig. 2.27 yields an information gain of 0.47 bits. If each stimulus lasts 400 msec, then the bit rate increases by roughly 1.2 bits/second.

For different statistics of the stimulus distribution, the neuron will learn the appropriate response function to match. In fig. 2.28, the response becomes steeper as the stimulus amplitude distribution becomes narrower, or the entire response shifts as the underlying input distribution shifts.

Since the coupling conductance G between the dendritic and somatic compartment



(b)

Figure 2.27: The inputs to the model are synaptic conductances, drawn randomly from a Gaussian distribution of mean 1.0 mS/cm^2 and standard deviation of 0.4 mS/cm^2 with the restriction that the conductance be non-negative. 120,000 iterations of the learning rule of eq. 2.31 were used to adjust the peak conductances, midpoint voltages, and slopes of six Ca^{2+} and six K^+ dendritic conductances to maximize the information in the firing rate. As is standard in most stochastic learning schemes, the learning rate decayed with time: $\eta = \eta_0 \exp(-t/\tau_{\text{learning}})$, with $\eta_0 = 0.025$ and $\tau_{\text{learning}} = 40,000$ iterations. The optimal firing rate response curve (dotted line in fig. 2.27a on the previous page) is asymptotically proportional to the cumulative probability distribution of inputs: $f_{opt}(g_{\text{syn}}) \sim f_0 + C \int_0^{g_{\text{syn}}} p(g)\, dg$. Figure (b) shows the probability distribution of firing rates before and after learning. Learning shifts the distribution from a peaked distribution to a flat one, so that the neuron uses each firing rate within the range [15 , 55] Hz equally often in response to randomly selected synaptic inputs. Superimposed on the graph of the "learned" probability distribution of firing rates is a normalized frequency histogram, counting how often each firing rate occurred in 50,000 iterations of the adapted model.

plays an important role in the assumptions needed to derive the learning rule 2.40, we repeat the entire simulation for fig. 2.27 using two other values of the coupling conductance and compare the results in fig. 2.29. Each curve represents 120,000 iterations of the learning rule, with the same initial conditions for the modulatory voltage-dependent conductances in the dendritic compartment. The learning rule constrains the voltage in the dendritic compartment to remain within the same range in each case; the current delivered to the soma when the dendritic compartment is at a particular voltage, however, changes as a function of the coupling conductance. (The function $c(\langle V \rangle)$ in eq. 2.40 remains the same.) For this reason, the input conductances are mapped onto a different range of firing rates for different values of the coupling conductance.

Changing how tightly the stimulus amplitudes are clustered around the mean will increase or decrease the slope of the firing rate response to input, without necessarily changing the average firing rate. This gives rise to a specific experimental prediction that can be evaluated using standard *in vitro* current clamp methods. Suppose that a neuron is subjected to electrical stimulation during its developmental period, such that the stimulation consists of current injections of random amplitudes. Information maximization learning rules then predict that the neuron will change the slope (gain) of its response as the variance of the input changes.

In fig. 2.30, we model a similar experiment: after the neuron has adjusted to one probability distribution of stimuli, the same distribution is made narrower. The neuron responds to this "artificial" manipulation of the input statistics by increasing

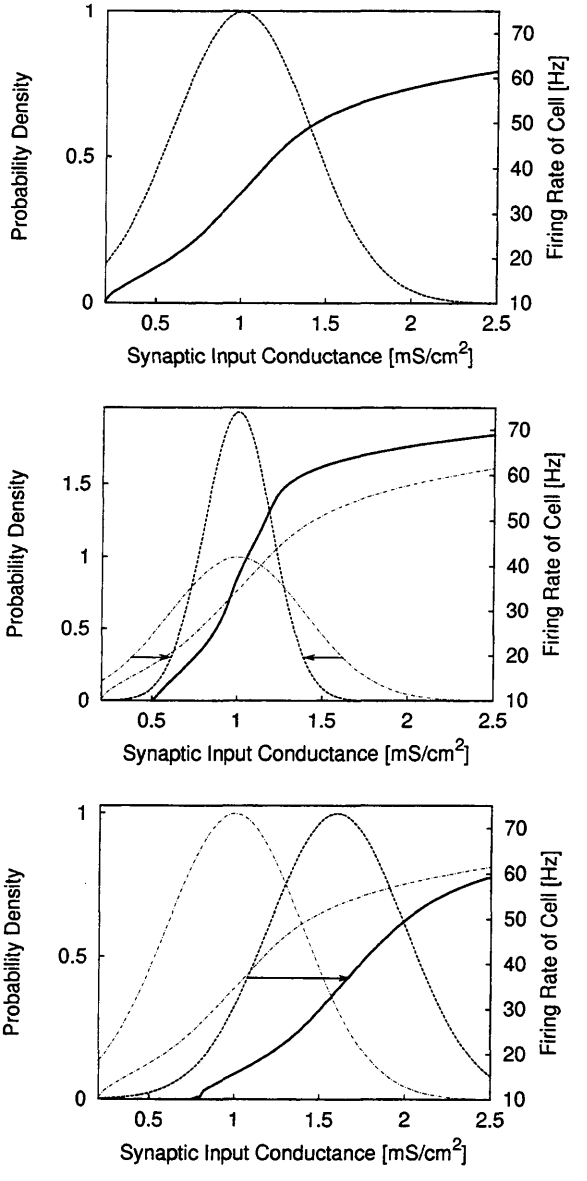


Figure 2.28: By adapting the voltage-dependent conductances in the dendritic compartment, the model neuron attempts to match its response function to the underlying cumulative distribution function of input conductances. The top panel replots the firing rate response as a function of synaptic conductance input from fig. 2.27. The middle panel shows the response of the neuron after 120,000 iterations of training on a narrower Gaussian input probability distribution with width $\sigma = 0.2$ nS/cm^2. Initial conditions at the start of training were kept the same as before. In the lower panel, the variance of the Gaussian changes, but the midpoint is shifted to the right by 0.6 nS/cm^2 to 1.6 nS/cm^2. For comparison purposes, we superimpose the input probability distribution and firing rate response from the top panel as light dashed-dotted curves. The mean and variance of the stimulus probability distributions (dotted lines) set the midpoints and slopes of the firing rate curves (solid lines). Same parameters and initial conditions as in fig. 2.27.

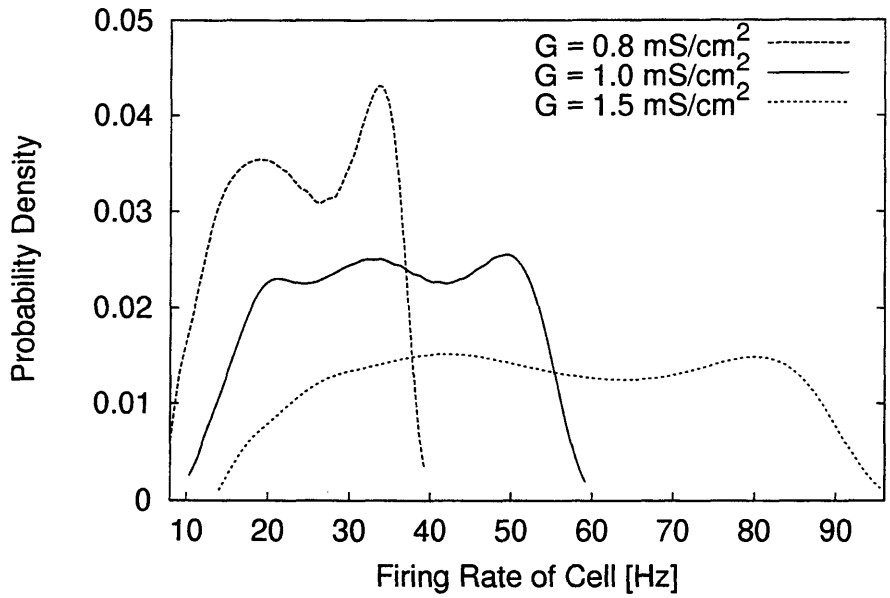


Figure 2.29: The learning rule 2.40 is strictly only valid when the coupling conductance G is not much larger than the mean conductance across the somatic membrane. By varying G but not the constraints imposed by the learning rule on the voltages in the dendritic compartment, the model learns to map the range of synaptic conductances onto different ranges of firing rates. All initial conditions and parameters, except for the coupling conductance, are identical to those used in fig. 2.27.

its firing rate response gain.

The learning rule of eqn. 2.31 depends on the membrane voltage and the states of the ion channels underlying the voltage-dependent conductances. If neurons adapt their voltage-dependent conductances in such a manner as to maximize the information transmitted, then *in vitro* current-clamp and voltage-clamp experiments should both reveal changes in these conductances. Clamping the voltage, however, will initiate the neuron's learning mechanism to change the conductances, but will not allow these conductances to achieve the intended effect of changing the neuron's response. Therefore, *in vitro* voltage-clamp experiments over long durations (hours) should induce neurons to behave in a variety of "pathological" ways in subsequent current-clamp experiments: previously regularly spiking neurons will produce tonic firing patterns, bursting behavior, plateau potentials, or become completely quiescent.

While the detailed substrate for learning information maximization at both the single cell and network level awaits experimental elucidation, the terms in the learning rule of eqn. 2.40 have simple biophysical correlates, as outlined in section 2.2: the derivative term, for instance, is reflected in the stochastic flicker of ion channels switching between open and closed states. In that section, we showed that the transitions between simple open and closed states at a steady state voltage will occur at a rate proportional to $\left[\frac{d}{dV}m_\infty(V)\right]^\gamma \sim \left[m_\infty(V) - m_\infty(V)^2\right]^\gamma$, where the exponent γ is $1/2$ or 1, depending on the kinetic model used for the ion channels. The average rate of ion channel transitions during one period of a regularly spiking Hodgkin-Huxley model will be (approximately) proportional to $\left[\langle m\rangle - \langle m\rangle^2\right]$ ($\gamma = 1$), as discussed in

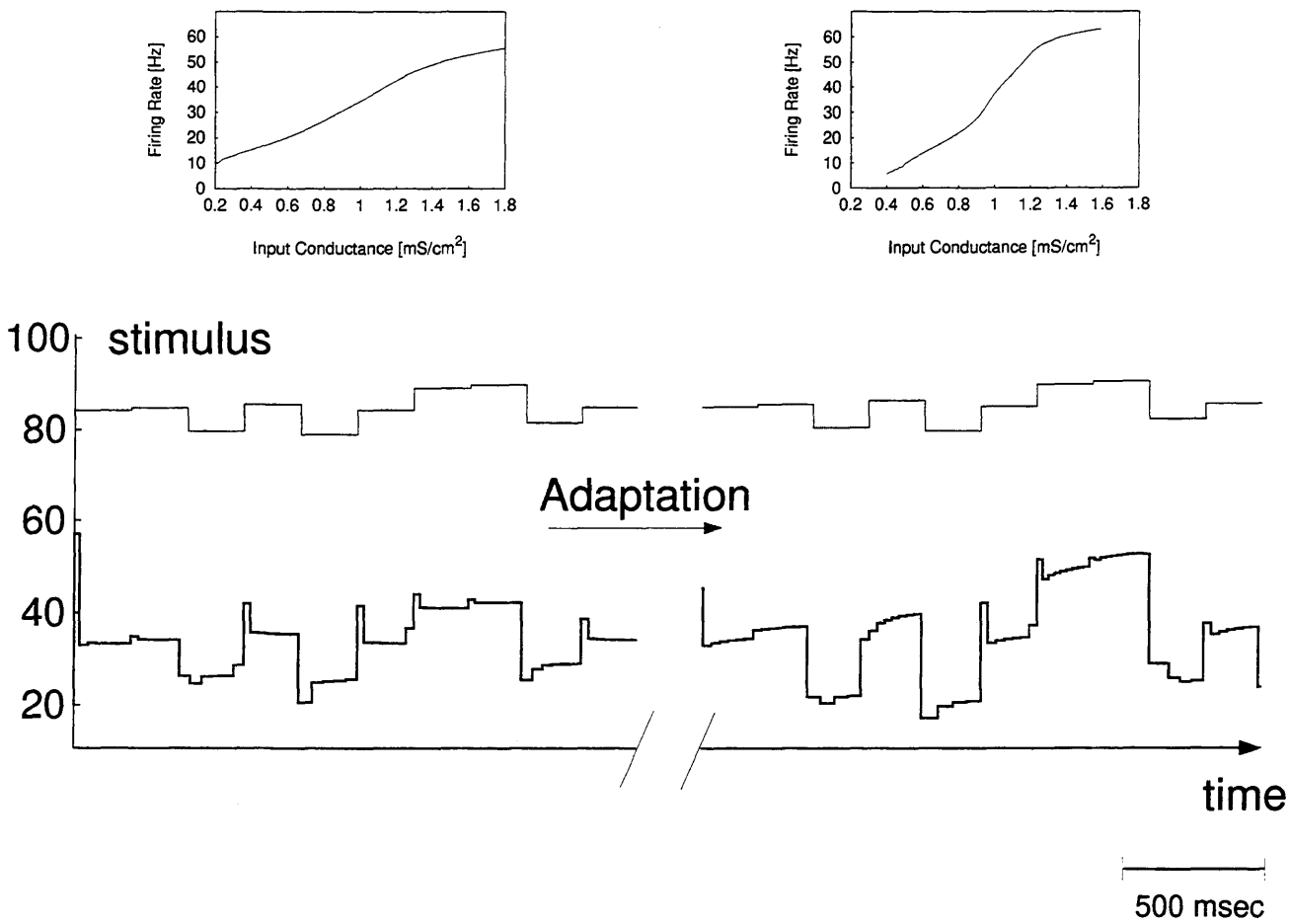


Figure 2.30: Response of the model after readaptation to a new stimulus distribution. Every 200 milliseconds, a synaptic conductance is chosen from a Gaussian distribution of mean 1.0 mS/cm^2 and standard deviation of 0.2 mS/cm^2, which is twice as narrow as the previous distribution of inputs in fig. 2.27. On the left, the voltage-dependent conductances were set by training on the original stimulus distribution; on the right, the values of the conductances were relearned by training for 200,000 iterations on the new, narrower stimulus distribution. The instantaneous firing rate is plotted as a function of time in response to the identical stimulus train before and after readaptation. Note that the firing rate response curve becomes steeper as a function of input conductance after readaptation.

appendix A.2. If we replace

$$\Big(\langle m_{\infty,i}(V)\rangle - \langle m_{\infty,i}(V)^2\rangle\Big)$$

in the learning rule of eq. 2.40 by

$$\Big(\langle m_i\rangle - \langle {m_i}^2\rangle\Big),$$

then we will overestimate the derivative, unless $V(t)$ is constant or τ is zero. This overestimate, however, is not serious, and has no discernible, systematic effect in the learning rule simulations.

To change the information transfer properties of the cell, a neuron could use state-dependent phosphorylation of ion channels or gene expression of particular ion channel subunits, possibly mediated by a G-protein initiated second messenger cascade, to modify the properties of voltage-dependent conductances. The tools required to adaptively compress information from the senses are thus available to single neurons at the subcellular level.

In the next section, we turn to another aspect of the stimulus statistics, namely the structure of correlations in the input, and ask how an optimally tuned neuron responds to correlated input.

2.13 Time-Varying Input Signals

A full description of the statistics of the stimuli reaching an individual neuron will not only specify stimulus amplitude distribution, but also describe the correlations between successive stimuli in time. Until now, we have assumed that each stimulus is independently drawn from the amplitude distribution and that the neuron is allowed to reach a steady state of firing. In this section, we examine what happens when we drop this assumption.

Following Atick and Redlich (1992) and Bell and Sejnowski (1995), we show that maximizing the entropy or mutual information of a time series $X(n\Delta t)$ implies that the mutual information between any two distinct samples of the time series $X(k\Delta t)$ and $X(l\Delta t)$ is *minimized.* Of course, this only holds when each element in the time series is bounded in magnitude from above and below, i.e., subject to the constraint that $0 \leq X(k\Delta t) < X_{\max}$ for any k.

Given a collection of random variables $\{X(t_1), X(t_2) \cdots, X(t_n)\}$, the entropy $S(\{X(t_1), X(t_2), \cdots, X(t_n)\})$ can be expanded using a chain rule (Cover and Thomas, 1991):

$$\begin{aligned} S(X(t_1), X(t_2), \cdots, X(t_n)) &= S(X(t_1)) + S(X(t_2)|X(t_1)) + S(X(t_3)|X(t_1), X(t_2)) + \cdots \\ &= \sum_{k=1}^{n} S(X(t_k)|X(t_{k-1)}, \cdots, X(t_1)) \\ &\leq \sum_{k=1}^{n} S(X(t_k)). \end{aligned}$$

In other words, the entropy of the set is less or equal to the sum of the individual entropies, since any conditional entropy $S(X|Y)$ is less than the original entropy $S(X)$. The same expansion and inequality holds true if we replace the entropy S by the mutual information with respect to a set of variables $\{V_1, \cdots, V_n\}$.

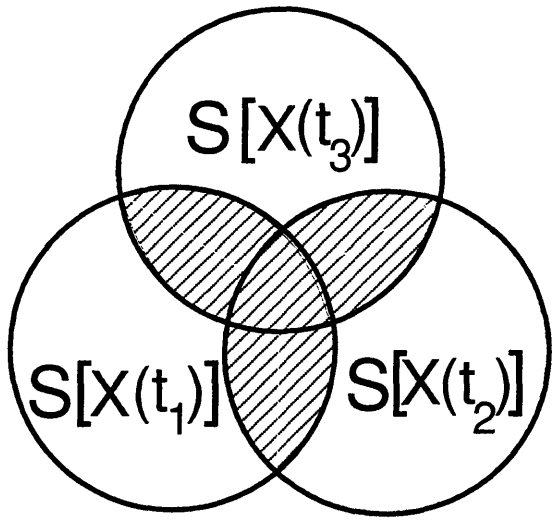


Figure 2.31: The entropy measures the volume or number of accessible states of a system. If the variable $X(t_1)$ depends on the variable $X(t_2)$, then the joint entropy of both is reduced; the shared entropy between two variables is indicated by the overlap regions above. A necessary condition for maximizing the overall entropy is that the variables $X(t_1), X(t_2), X(t_3)$ become statistically uncorrelated and independent of each other.

Bounding the values of $X(k\Delta t)$ bounds the entropy of $X(k\Delta t)$. Thus,

$$\sum_{k=1}^{n} S(X(t_k)) \leq nS_{\max}.$$

In the Venn diagram of fig. 2.31, $S[X(t_2)] - S[X(t_2)|X(t_1)]$ is represented by the overlap of two circular discs. Since the area (i.e., the entropy) of any disc is bounded, maximizing the entropy of the set implies eliminating these overlaps. But the overlap between any two such discs is exactly the mutual information between the two

elements of the time series:

$$I[X(t_2), X(t_1)] = S[X(t_2)] - S[X(t_2)|X(t_1)]$$

Hence, maximizing the entropy of a time series also decorrelates successive elements of the series, such that the value of the element $X(n\Delta t)$ cannot be predicted by the value of $X(\,(n-1)\Delta t)$ in the previous time interval. A time-varying signal that is strongly correlated in time changes slowly; if a device decorrelates the signal, these slow changes are subtracted from the output.

We need to recast the Hodgkin-Huxley equations into a new form to apply the ideas behind information maximization. Start with the full expression for the time-dependent active conductances for Ca^{2+} and K^+:

$$\frac{dm}{dt} = \frac{m_\infty(V) - m}{\tau(V)}$$

$$m(t) = \int_{-\infty}^{t} m_\infty[V(t')] \frac{\exp\left(-\dfrac{t-t'}{\tau[V(t')]} \right)}{\tau[V(t')]}\, dt'$$

or alternatively, with an initial condition for m at $t = 0$:

$$m(t) = \int_{0}^{t} m_\infty[V(t')] \frac{\exp\left\{ -\dfrac{t-t'}{\tau[V(t')]} \right\}}{\tau[V(t')]}\, dt' + m(0)\exp[-\frac{t}{\tau[V(t)]}]$$

We can rewrite the Hodgkin-Huxley equations,

$$C\frac{dV}{dt} = g_{\text{leak}}(E_{\text{leak}} - V) + g_{\text{syn}}(t)(E_{\text{syn}} - V) + \sum_i \bar{g}_i m_i(t)(E_i - V)$$
$$\tau_i(V)\frac{dm_i}{dt} = m_{\infty,i}(V) - m_i,$$

in terms of an integro-differential equation (with $E_{\text{leak}} = 0$):

$$\Big[g_{\text{leak}} + C\frac{d}{dt}\Big]V = g_{\text{syn}}(t)(E_{\text{syn}} - V) + \sum_i \left\{ \bar{g}_i(E_i - V)\int_0^\infty m_{\infty,i}\big[V(t-t')\big]\frac{\exp\Big\{-t'/\tau_i\big[V(t-t')\big]\Big\}}{\tau_i\big[V(t-t')\big]}dt' \right\} \tag{2.52}$$

If the time constants $\tau_i(V)$ are voltage-independent, then

$$\begin{aligned} I(t) &= \Big[g_{\text{leak}} + C\frac{d}{dt}\Big]V \\ &= g_{\text{syn}}(t)(E_{\text{syn}} - V) + \sum_i \left\{ \bar{g}_i\,(E_i - V)\int_0^\infty m_{\infty,i}\big[V(t-t')\big]\frac{\exp(-\frac{t'}{\tau_i})}{\tau_i}dt' \right\}. \end{aligned}$$

Discretize time into units Δt and examine the mapping

$$\{V(0), V(\Delta t), V(2\Delta t), \ldots, V((n-1)\Delta t)\} \to \{I(0), I(\Delta t), I(2\Delta t), \ldots, I((n-1)\Delta t)\}$$

The entropy of the set of currents at times $t = 0, \Delta t, \cdots, (n-1)\Delta t$ is

$$S(\{I\}) = \int \ln\Big[\det\Big|\frac{dI(m\Delta t)}{dV(n\Delta t)}\Big| p[\{V(n\Delta t)\}]\Big]\, dp[\{V(n\Delta t)\}].$$

In the time-discretized approximation,

$$m_i(t) \approx \lim_{K\to\infty} \sum_{k=0}^{K} m_{\infty,i}[V(t-k\Delta t)] \frac{\exp\left(-\frac{k\Delta t}{\tau_i}\right)}{\tau_i} \Delta t$$

$$\frac{dI(m\Delta t)}{dV(n\Delta t)} = \begin{cases} \sum_i \bar{g}_i\, m'_{\infty,i}[V(n\Delta t)]\, \frac{\exp[-\frac{(m-n)\Delta t}{\tau_i}]}{\tau_i} \Big(E_i - V(m\Delta t)\Big)\, \Delta t & \text{if } m > n \\ \sum_i \bar{g}_i \Big[m'_{\infty,i}[V(n\Delta t)]\Big(E_i - V(n\Delta t)\Big)\frac{\Delta t}{\tau_i} - m_i[n\Delta t]\Big] & \\ \quad + \frac{d}{dV(n\Delta t)}\Big[g_{\text{syn}}(n\Delta t)\Big(E_{\text{syn}} - V(n\Delta t)\Big)\Big] & \text{if } m = n \\ 0 & \text{if } m < n \end{cases}$$

Of course, seeking to maximize the entropy of the currents is not a well-posed optimization problem, since the entropy function can grow without bound. Instead, we seek to maximize the entropy of the time-averaged firing rate or response $r(t)$, sampled at time intervals of Δt. Here, the response $r(t)$ is

$$r(t) = \int_{-\infty}^{\infty} h(t-\tau) f\Big[I(\tau)\Big]\, d\tau \tag{2.53}$$

where $h(t)$ is any causal function such that $h(t) = 0$ for $t < 0$ and f is the instantaneous firing rate, defined as some nonlinear, saturating function of the current. The

exact form of the causal function $h(t)$ will play no role in the derivation of the learning rules. The stochastic approximation algorithm leads to the following update rule for changing the parameter q

$$\begin{aligned}\Delta q &= \eta \frac{\partial}{\partial q} \ln\left[\det\left|\frac{dr}{dI}\frac{dI}{dV}\right|\right] \\ &= \eta \frac{\partial}{\partial q} \ln\left[\det\left|\frac{dr}{dI}\right|\right] + \eta \frac{\partial}{\partial q} \ln\left[\det\left|\frac{dI}{dV}\right|\right] \\ &= \eta \,\mathrm{Tr}\left[\left(\frac{dr}{dI}\right)^{-1} \frac{\partial}{\partial q}\left(\frac{dr}{dI}\right)\right] + \eta \,\mathrm{Tr}\left[\left(\frac{dI}{dV}\right)^{-1} \frac{\partial}{\partial q}\left(\frac{dI}{dV}\right)\right], \end{aligned} \tag{2.54}$$

where we have used the fact that

$$\begin{aligned}\ln\big[\det(M+\delta M)\big] - \ln\big[\det(M)\big] &= \ln\big[\det(M^{-1}M + M^{-1}\delta M)\big] \\ &= \ln\big[\det(1 + M^{-1}\delta M)\big] \\ &= \mathrm{Tr}\big(M^{-1}\delta M\big),\end{aligned}$$

valid for a nonsingular square matrix M.

We first examine the second term involving dI/dV in isolation. The change in the parameter q due to this term alone is

$$\begin{aligned}\Delta q &= \eta \,\mathrm{Tr}\left\{\left[g_{\mathrm{leak}} + C\frac{d}{dt}\right]^{-1} \frac{\partial}{\partial q}\left(\frac{dI(m\Delta t)}{dV(n\Delta t)}\right)\right\} \\ &= \eta \sum_m \left[g_{\mathrm{leak}} + C\frac{d}{dt}\right]^{-1} \frac{\partial}{\partial q}\left(\frac{dI(m\Delta t)}{dV(m\Delta t)}\right)\end{aligned}$$

In the limit as $\Delta t \to 0$, the trace is replaced by an integral, and the derivatives by

functional derivatives. For the specific example of the peak conductance $\bar{g}_i$, we have

$$\Delta\bar{g}_i = \eta \int_{-\infty}^{t} \left[g_{\text{leak}} + C\frac{d}{dt}\right]^{-1} \frac{1}{\tau_i}\Big\{m'_{\infty,i}[V(t')]\left[E_i - V(t')\right] - m_{\infty,i}[V(t')]\exp(-\frac{t-t'}{\tau_i})\Big\}\,dt',$$

which, by Taylor expansion of the operator, becomes

$$\begin{aligned}
&= \eta \int_{-\infty}^{t} \Bigg[\frac{1}{g_{\text{leak}}\tau_i}\Big\{m'_{\infty,i}[V(t')]\left[E_i - V(t')\right] - \left[m_{\infty,i}[V(t')] - m_i(t')\right]\Big\} \\
&- \frac{C}{g_{\text{leak}}^2\tau_i}\left(\Big\{m''_{\infty,i}[V(t')]\left[E_i - V(t')\right] - 2\,m'_{\infty,i}[V(t')]\Big\}\left[\frac{d}{dt'}V(t')\right] + \frac{m_i(t') - m_{\infty,i}[V(t')]}{\tau_i}\right) \\
&\qquad + \cdots\Bigg]\,dt'
\end{aligned}$$

The validity of this expansion, of course, depends on the relative magnitudes of g_{leak} and $C\frac{d}{dt}$. Taking the Fourier transform, $\tilde{\mathcal{L}}^{-1} = [g_{\text{leak}} + i\omega C]^{-1}$, so we see that an expansion is valid at low temporal frequencies. We can obviate this approximation by multiplying the right hand side by the operator $\mathcal{L}$. In fact, we can multiply the right hand side by any matrix or linear operator as long as long as the real part(s) of the operator are positive. This condition ensures that the minima of the objective function (in this case, the entropy) remain unchanged.

If only a single peak conductance is varied, then the integral formulation of the learning rule is equivalent to the following differential formulation:

$$\tau_i\frac{\Delta\bar{g}_i}{dt} = \eta\Big\{m'_{\infty,i}[V(t)]\left[E_i - V(t)\right] - \left[m_{\infty,i}[V(t)] - m_i(t)\right]\Big\} \tag{2.55}$$

This is simply the result of taking the time derivative on both sides of the previous

equation and multiplying the right hand side by $\mathcal{L}$.

When more than one parameter is varied simultaneously, then the two formulations are no longer exactly equivalent. Suppose that the stimulus has an intrinsic time scale τ_s. A difference proportional to $\eta^2\tau_s$ evolves between the integral and differential formulation of the learning rule over that time scale, which reflects the coupling of the change in one parameter to the changes in other parameters. Both formulations are stochastic approximations to the entropy integral.

To first order, the learning rule in eq. 2.55 has the same form as the previously described learning rule for quasi-static stimuli. Note, however, that the rate of learning explicitly incorporates the time constant τ_i for each (in)activation variable, such that the time scale for parameter adaptation matches the intrinsic time scale. It is important to note that the parameters for 'slow' (in)activation variables adapt slowly compared to those for 'fast' variables.

The similarity to the quasi-static case really ends when we consider voltage-dependent conductances. In this case, the update rule for the peak conductances is to first order

$$\tau_i[V(t)]\,\frac{\Delta\bar{g}_i}{dt} = \eta\left(\left\{m'_{\infty,i}[V(t)] - m_{\infty,i}[V(t)]\frac{\tau'_i[V(t)]}{\tau_i[V(t)]}\right\}\Big[E_i - V(t)\Big] - \big(m_{\infty,i}[V(t)] - m_i(t)\big)\right).$$

If $\tau_i[V(t)] \propto \left\{m'_{\infty,i}[V(t)]\right\}^{\alpha}$, where $\alpha > 0$ is some exponent and $m_{\infty,i}[V]$ is given by a Boltzmann equation, then

$$\tau_i[V(t)]\frac{\Delta\bar{g}_i}{dt} = \eta\Bigg[\Big((1-\alpha)m'_{\infty,i}[V(t)] + \alpha s_i\Big\{m_{\infty,i}[V(t)]\Big\}^2\Big)\Big[E_i - V(t)\Big] - (m_{\infty,i}[V(t)] - m_i(t)$$

$$= \eta\left\{m_{\infty,i}[V(t)]\frac{s_i}{2}\Big[E_i - V(t)\Big] - (m_{\infty,i}[V(t)] - m_i(t))\right\} \qquad \text{if } \alpha = 1/2.$$

Recall that $\tau_i[V(t)] \propto \left\{m'_{\infty,i}[V(t)]\right\}^{1/2}$ when ion channel opening and closing is governed by a symmetric energy barrier separating the two states of the channel. When $\alpha = 1/2$ and the input is varying slowly, the learning rule will change primarily the peak conductances for those channel types that are fully in the open state ($m_{\infty,i}(V) \approx 1$), rather than giving preference to ion channels that are balanced between the open and closed states.

Returning to the first term of eq. 2.54 on page 135, we compute the variational derivative of the response $r(t)$ with respect to the current I at time τ from eq. 2.53:

$$\frac{\delta r(t)}{\delta I(\tau)} = \begin{cases} f'[I(\tau)]h(t-\tau) & \text{if } \tau \leq t \\ 0 & \text{if } \tau > t \end{cases}$$

The discrete-time matrix corresponding to $\delta r(t)/\delta I(\tau)$ is triangular. The set of lower (or upper) triangular $n \times n$ matrices form a closed group under multiplication, from which it follows that the inverse matrix $[\delta r(t)/\delta I(\tau)]^{-1}$ is also triangular. Note that $f(I)$ needs to be an invertible function for this inverse to be guaranteed to exist. For any two lower (upper) triangular matrices $\mathbf{A}$ and $\mathbf{B}$ of the same size, the trace of the product $(\mathbf{A}.\mathbf{B})$ is $\mathrm{Tr}(\mathbf{A}.\mathbf{B}) = \sum_i A_{ii}B_{ii}$. This allows the the trace in the learning rule

to be evaluated immediately:

$$\mathrm{Tr}\left[\left(\frac{\delta r}{\delta I}\right)^{-1}\frac{\partial}{\partial q}\left(\frac{\delta r}{\delta I}\right)\right] \to \int_{-\infty}^{t}\left[\frac{df}{dI}\right]^{-1}\frac{d^2 f}{dI^2}\,\frac{\partial I(t')}{dq}\,dt'$$

Here we have assumed that the function $h(t)$ does not depend on any of the conductance parameters.

After taking a time derivative, the second term for the peak conductance change reads

$$\frac{d\Delta\bar{g}_i}{dt} = \eta\left[\frac{df}{dI}\right]^{-1}\frac{d^2 f}{dI^2}\,m_i(t)\big[E_i - V(t)\big].$$

We can multiply the right-hand side of this expression by the operator $\mathcal{L}$ from above to match the first term in the learning rule.

To complete the expression for the learning rules for time-varying stimuli, one more constraint is needed. Signals passing through neurons are always corrupted by noise, from synaptic noise due to the stochastic release of neurotransmitter-containing vesicles to the shot-noise of ion transport through ion channels. The noise sets the minimum time scale over which any system must average, lest it amplify the noise instead of the signal. The inverse of this minimum time scale corresponds to the cut-off frequency of the optimal power spectrum of the output, as defined by the mutual information: the spectrum should be flat (decorrelated) up to a frequency f_c, then drop to zero as the signal-to-noise ratio dips below unity.

Since we will not simulate a detailed noise model, we take a simpler, more phenomenological approach: we add a "regularizer" to the objective function we are

seeking to maximize. The objective function in this case is still the entropy of the time series. In the continuous limit and for Gaussian noise, the entropy can be written in terms of the power spectrum $P(\omega)$ as (Papoulis, 1984)

$$S = \ln\sqrt{2\pi e} + 1/N \int \ln P(\omega) d\omega, \tag{2.56}$$

where N is a normalization factor. To limit how fast the voltage changes in time, a term $-\lambda(dV/dt)^2$ is added to the entropy as the objective function. Here λ is a Lagrangian (free) parameter, which will set the effective cut-off frequency. The most important effect of the regularizer $-\lambda(dV/dt)^2$ is that it constrains the variance of $\dot{V} = dV/dt$ to remain finite, such that

$$\begin{aligned} \sigma_{\dot{V}}^2 &= \int \omega^2 P(\omega) d\omega \\ &< \infty. \end{aligned}$$

This constraint ensures that the power spectrum will remain band-limited. By computing the variational derivative of the new objective function with respect to the power spectrum $P(\omega)$, the optimal power spectrum can be shown to scale as

$$P(\omega) \sim \frac{a}{b + \omega^2},$$

where a and b are constants

Incorporating the regularization term into the learning rule for the peak conduc-

tance leads to one more additional term that reads

$$\frac{d\Delta\bar{g}_i}{dt} = -\eta\,(2\lambda)\,m_i(t)[E_i - V(t)]\,\frac{dV}{dt} \tag{2.57}$$

To test the learning rules for time-varying stimuli, we take a one-compartment model of a neuron as in fig. 2.12. We impose an auxiliary condition on the dynamic range of steady-state voltages, as described in section 2.8, in lieu of some arbitrary nonlinear $f(I)$ relationship and simply examine the current or voltage as the neuron's output. The "whitening" of the power spectrum of the voltage time-course is displayed in fig. 2.32:

As in the simulations of section 2.12, six Ca^{2+} and six K^+ conductances were used. If the model neuron is to decorrelate inputs, it should respond more strongly to the transients in the incoming signal rather than the sustained background input. So that the neuron can learn to "adapt" to longer-lasting stimuli by reducing the firing rate, each conductance was associated with both an activation m_i *and* an inactivation variable h_i. The steady-state (in)activation functions are described by Boltzmann functions in the usual way: $h_{\infty,i}(V) = 1/\{1+\exp[(V-V_i)/s_i]\}$ and $m_{\infty,i}(V) = 1/\{1+\exp[-(V-V_i)/s_i]\}$. Initial conditions were such that the activation and inactivation functions for each conductance type overlap. The initial slopes of h_i and m_i were both 10 mV. Time constants were chosen to be voltage-independent, for simplicity. The time constant for all activation variables was initially 10 milliseconds, for inactivation 100 milliseconds.

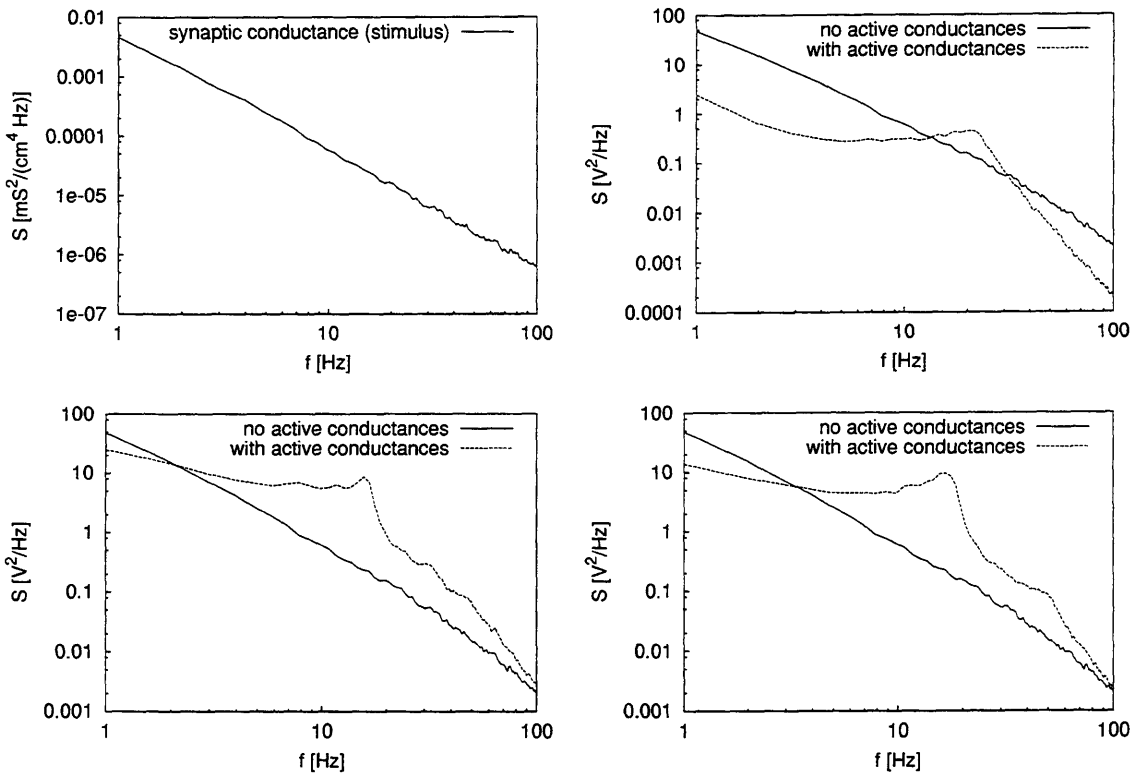


Figure 2.32: A flat output power spectrum indicates decorrelation of statistical dependencies in the input time series. The synaptic conductance given as input to a one-compartment model was Gaussian in amplitude with exponentially decaying correlation in time. The time constant of these input conductance correlations was 250 milliseconds. After adaptation of the time constants, peak conductances, midpoint voltages and slopes of the activation and inactivation curves, the voltage power spectrum is flat up to a cut-off frequency. Three separate runs are shown, each corresponding to 14.6 hours of real time. The initial learning rate was $\eta_0 = 1 \times 10^{-5}$; the learning rate decayed exponentially in time with a time constant of 1.9 hours.

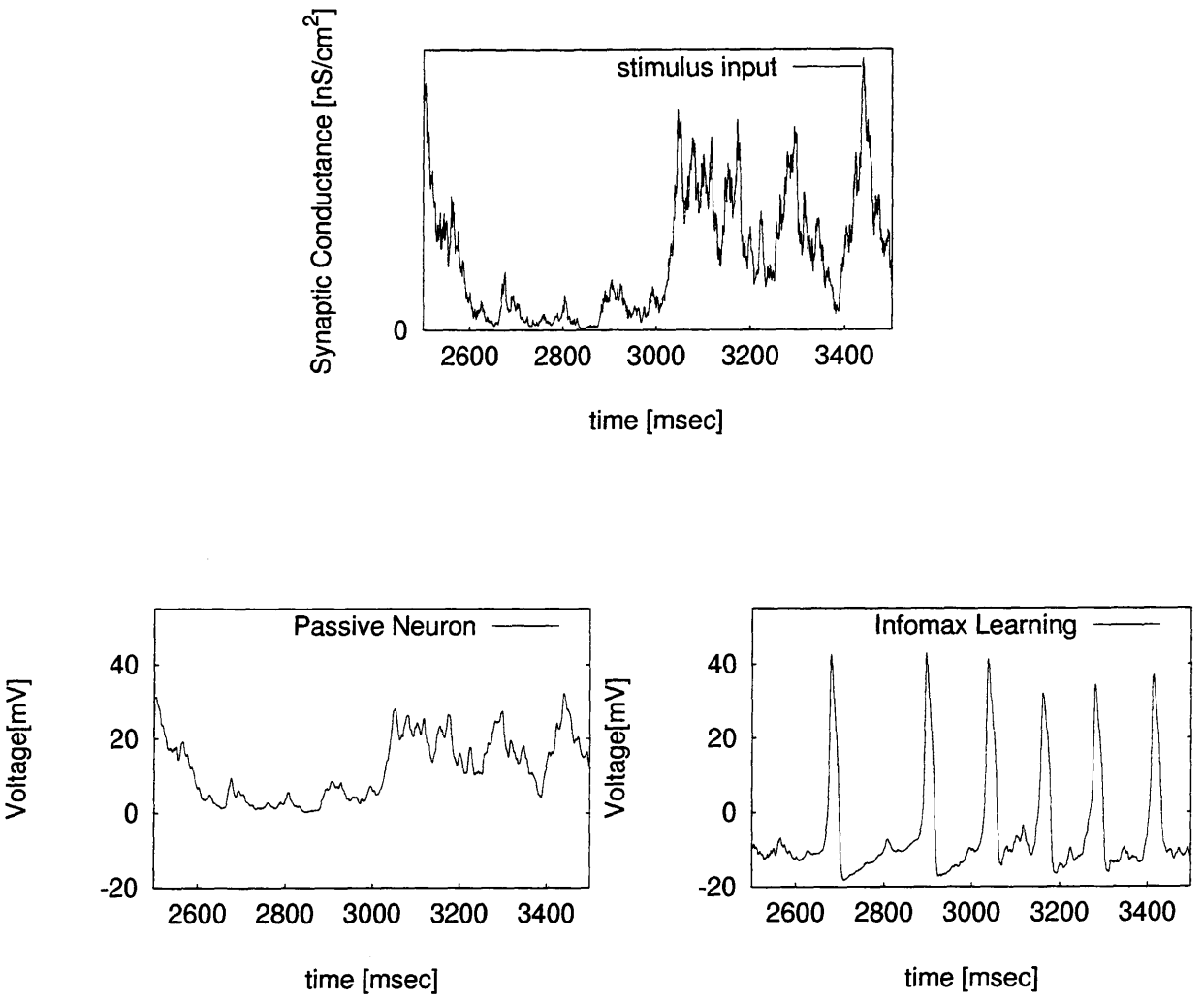


Figure 2.33: Set of voltage traces in response to a time-varying synaptic conductance input before and after information-maximization-based learning of voltage-dependent conductances. Note that the neuron has "learned" to produce spikes, albeit on a longer timescale than standard sodium spikes. The corresponding power spectra for this figure and the next two are shown in fig. 2.32.

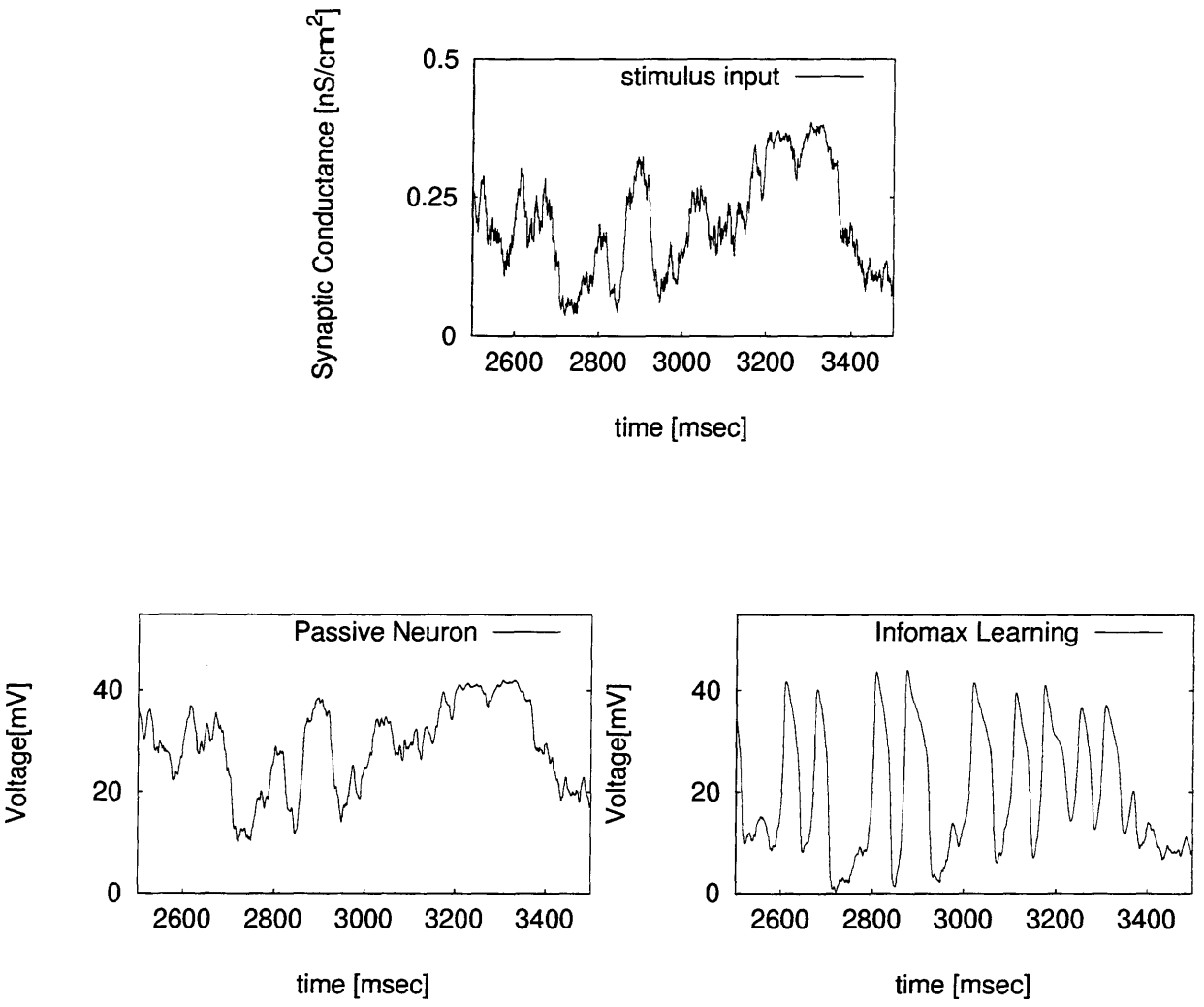
At the start of training the model with random, time-varying synaptic conductance input, all peak values of the voltage-dependent conductances are set to zero. The model thus begins life as a completely passive neuron, a simple RC electrical circuit that low-pass filters and smooths the synaptic conductance input signal. By training the model on time-varying stimuli and applying the learning rules for entropy maximization above, the model starts responding to input by producing spike-like oscillations in the voltage, as shown in fig. 2.33.

To discover why such oscillations occur, we refer back to the stability analysis of the steady state voltage in section 2.8. There, we showed that a fixed point V^* of the voltage dynamics is stable only if

$$\left.\frac{dI}{dV}\right|_{V^*} < 0,$$

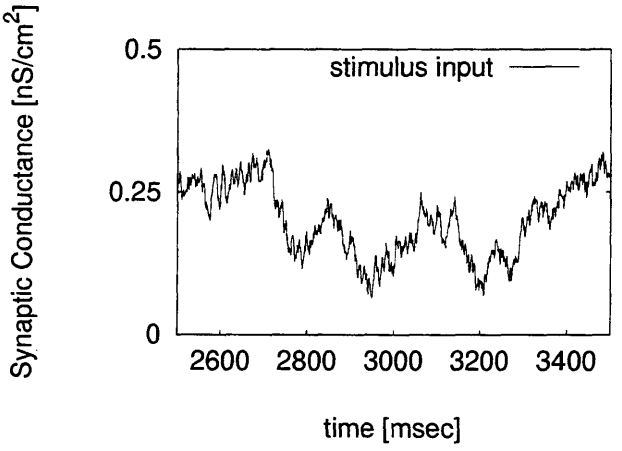
i.e., the total derivative of the current is less than zero. If all sustained inputs elicit stable voltage responses, then the dynamics in response to time-varying inputs will be relaxational, i.e., the voltage always changes in the direction of the stable voltage given by the input at that instant.

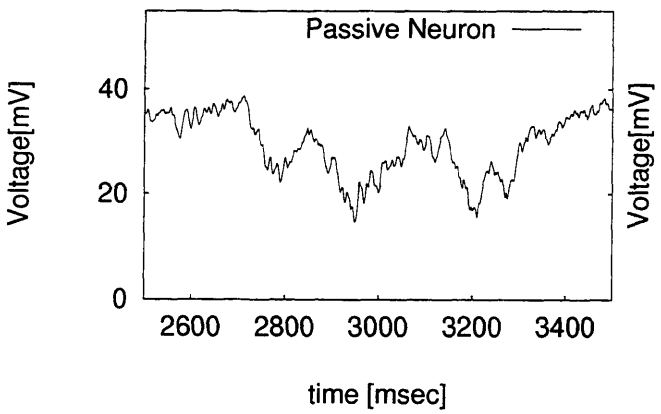
For the sake of argument, suppose that a single, voltage-independent time constant governs all voltage-dependent conductances. Eq. 2.55 of the learning rule is then simply stochastic gradient ascent on $dI/dV[V(t)]$; the learning rule, therefore, pushes the system away from stable, relaxational dynamics. When this derivative becomes zero, the voltage dynamics begin to exhibit oscillations in response to steady input. In the Hodgkin-Huxley model described in section 2.12, a transition of this type occurs

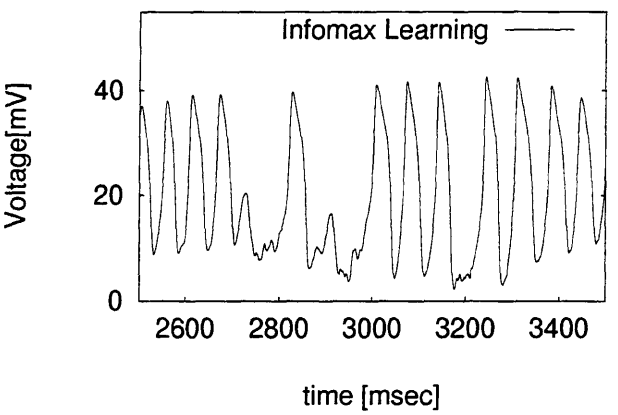


at the soma as the current injected is increased—$dI/dV = 0$ at the soma corresponds to the firing threshold.

The destabilizing character of the entropy maximization learning rules remains unchanged if the time constants of different voltage-dependent conductances are allowed to vary. Since the learning rules for changing the peak conductances are weighted by the time constant, oscillations will arise through the conductances that respond quickly to changing input, rather than the slow conductances. Hence, the oscillations themselves will be fast and "spike"-like.


Synaptic Conductance [nS/cm^2]
0.5
0.25
0
stimulus input
2600
2800
3000
3200
3400
time [msec]


Voltage[mV]
40
20
0
Passive Neuron
2600
2800
3000
3200
3400
time [msec]


Voltage[mV]
40
20
0
Infomax Learning
2600
2800
3000
3200
3400
time [msec]

Chapter 3 Stochastic Resonance

3.1 Noise and the Speed of Information Processing

Maximizing speed is a central objective in the ecological design of a sensory system. Given that the typical time constant of a single neuron is on the order of fifteen milliseconds, it is remarkable that the nervous system can accomplish complex sensory processing tasks within several multiples of this time constant. Perceptual experiments using rapid-fire sequences of visual displays, for instance, reveal that the human visual system can identify complex objects within 70 msec (Subramaniam *et al.*, 1995; Fiser *et al.*, 1996). The vestibulo-ocular reflex (VOR) is even faster, with a response time of about 10 msec. In the hierarchy of cortical processing, each stage is accorded only a limited time in which to process information; the information in the neuronal response, therefore, must be contained in the earliest time segments (Celebrini *et al.*, 1993). Information-theoretic analysis reveals that most of the information can be recovered from the first 50 msec of a visual neuron's response in inferotemporal cortex (Tovée *et al.*, 1993; Tovée, 1994), underlining the speed with which neurons process information. Moreover, the response of a *single* neuron within these first 50 msec typically correlates well with the behavioral decision an awake behaving monkey makes on a perceptual task, based on studies in the motion area MT (Britten *et al.*, 1996). Earlier experiments over longer time windows (two seconds) by Newsome *et*

al. (1989) revealed that a single neuron acting as a putative detector of visual signals typically matches or outperforms the animal in detection accuracy.

Taken together, these results highlight the role of single neurons in the fast and efficient processing of information. While many instances exist in which the cortex does integrate over long times or over many neurons, we will suppose that the sensory system has been optimized so that a single short segment of one neuron's spike train encodes a significant fraction of the total information about the external stimulus. There is a potential downside to such an efficient encoding. If the computational system is not afforded the luxury of averaging over many individual processing units or over long times, any effect of internal noise is magnified. Noise is commonly assumed to be detrimental to computation. However, since neurons typically have thresholds below which they do not respond, a paradoxical effect particular to threshold systems causes this common assumption to be violated: through the phenomenon of stochastic resonance (SR) (Wiesenfeld and Moss, 1995), *increasing* the input noise can maximize performance in systems with fixed thresholds.

Benzi *et al.* (1981) first coined the term stochastic resonance in the context of a model to describe the periodic recurrence of the Earth's ice ages. In this theory, climate changes are the consequence of noise-assisted transitions in a double-well potential system subject to very weak periodic modulation. Noise increases the frequency of hopping across the barrier between potential wells; with too much noise, however, the transitions become independent of the frequency of the periodic modulation. The coherence of the transitions with the driving frequency passes through a

maximum—"the stochastic resonance"—as a function of the noise amplitude.

SR has since been extended to include other physical systems with simpler thresholds responding to subthreshold periodic or aperiodic signals (Bulsara *et al.*, 1994; Stemmler *et al.*, 1995; Collins *et al.*, 1995a). Since spiking nerve cells can be described as threshold devices, the subject of stochastic resonance has recently evoked excitement in the field of computational neuroscience and opened up the possibility that noise could be used in the nervous system to enhance signal detection. SR has been studied experimentally in sensory neurons of the crayfish (Douglass *et al.*, 1993) and cricket (Levin and Miller, 1996) and has found application in models of the auditory nerve fiber (Longtin *et al.*, 1991), visual processing (Stemmler *et al.*, 1995), and cognition (Riani and Simonotto, 1994).

The standard measure of SR has been the signal-to-noise ratio (SNR) of the power spectrum. All spectral quantities, including the SNR, are based on averages over long times. A feature present on average, however, is not necessarily present in a limited sample, such as the single record of a neuron's spike train over the first few hundred milliseconds of its response. Thus the SNR does not at all address whether the system can reliably perform signal detection based on the spike output of a single neuron over a *short* duration.

More appropriate measures that explicitly depend on the sampling time are ones based on information theory: the mutual information between input and output, or, alternatively, the probability of correctly detecting the signal. In this paper, we will examine the existence of stochastic resonance in these information theoretic

quantities using the leaky integrate-and-fire model (Knight, 1972). This model is, without a doubt, the simplest spiking model of a nerve cell that still captures the essential dynamics of real cells (Koch *et al.*, 1995; Mainen *et al.*, 1995). At its heart, the integrate-and-fire model is a simple threshold crossing detector: whenever the membrane voltage crosses the threshold, the cell "fires" and is reset. The leak allows the voltage to decay in the absence of continuous input; as a consequence, a minimum input current is needed for the cell to spike in the absence of noise. The leaky integrate-and-fire model has two thresholds: the first, a fixed voltage threshold; the second, an input threshold that depends on the leak resistance.

For all its simplicity and generality, the integrate-and-fire model cannot be easily reduced to a bistable potential system. If one can define a potential, then all quantities of interest can be elegantly derived from the Kramers' rate for escaping over a potential barrier. The Kramers' rate approach underlies the theory for "stochastic resonance on a circle" (Wiesenfeld *et al.*, 1994; Moss *et al.*, 1994), an abstraction of general excitable systems with deterministic resets. Collins *et al.* (1995b) were able to explicitly reduce the more complicated Fitzhugh-Nagumo model to a bistable potential system, which justified the use of the Kramers' rate formalism. Here we will use a different approach. For the leaky integrate-and-fire model, we will first discuss the fundamental quantity of interest in signal estimation and then outline a set of approximations that will allow us to calculate the information-theoretic quantities directly.

3.2 Fisher Information and Signal Detection

Standard practice in neurophysiology consists of quantifying the neuronal response by summing up the number of spikes fired by a neuron over a fixed time period. The spike count constitutes an estimate of the firing rate of the neuron. Thalamic and cortical neurons typically sum over anywhere from one to several hundred input spikes before emitting a spike (Softky and Koch, 1993); once the threshold of such a neuron is fixed and set to a detection criterion, it can act as a detector of the spike signal from a presynaptic sensory or relay neuron. Consequently, the spike count is the natural measure for a signal detection system made out of neuronal building blocks.

Lord Adrian (1946) first proposed that information in the neuronal code is represented in the gradations of the firing rate, but this is by no means the only assumption one can make: information can, of course, be encoded in the precise timing of individual spikes. From a theoretical perspective, the primary motivation to choose the spike count as the measure of neuronal response is that such a choice allows us to take an analytical approach to the problem of signal detection.

We seek the simplest possible formulation of a signal detection problem; therefore, we will assume that the signal is not varying in time, but is instead constant, and that the noise fluctuations are temporally uncorrelated. More precisely, we consider the detection of a weak, constant signal $s = \Delta\mu$ riding on top of a background input μ in the presence of varying amounts of uncorrelated noise.

By relaxing some of these assumptions, we will be able to extend the arguments

presented below to the case where the signal is slowly varying compared to the time scale of the average interval between spikes (as in Collins *et al.*, 1995a). Periodic inputs will be considered in section 3.5, while the discussion of arbitrarily time-varying inputs will be deferred to sections 3.6 and 3.8.

The noise will lead to a probability distribution for the number of spikes recorded within a fixed time period. As illustrated in fig. 3.1, a weak signal will slightly shift the spike count distribution to higher spike counts. Maximizing the mutual information between the signal and the spike count corresponds roughly to minimizing the overlap between the signal and background spike count distributions. The overlap is a function of how far the means of the two distributions are separated and of how wide the distributions are; changing the input noise will change both the width and the mean of these distributions.

The model of signal detection considered here depends on the ability to estimate with precision the input μ that gave rise to the observed spike count output N. The Cramér-Rao bound sets the theoretical limit on how accurate an unbiased estimate of μ based on sampling the spike count distribution can be. This bound states that the variance between the estimate and the true input μ is always greater than the reciprocal of the Fisher information $J(\mu)$:

$$\mathrm{Var}(\mu_{\mathrm{estimate}}) \geq \frac{1}{J(\mu)}. \tag{3.1}$$

Here the Fisher information $J(\mu)$ is defined as follows (Cover and Thomas, 1991):

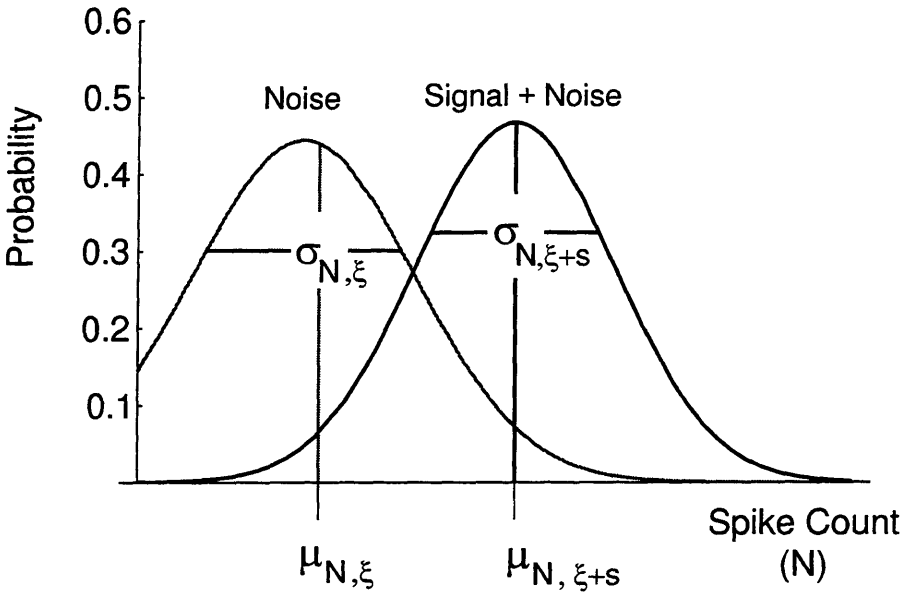


Figure 3.1: The goal in signal detection is to distinguish between a sample response drawn from the background spike count distribution, illustrated in gray, and a response drawn from the signal plus noise distribution, illustrated in black. Possible strategies to improve signal detection include changing the distance between the means of the two distributions and changing the width (variance) of the distributions.

$$J(\mu) = \int \left(\frac{\partial}{\partial \mu} p(N;\mu)\right)^2 / p(N;\mu)\, dN,$$

where $p(N;\mu)$ is the probability of observing N spikes, given an input μ. The lower the variance of the estimate of μ, the finer an input difference $s = \Delta\mu$ an ideal observer of the output can detect.

The Fisher information is closely related to the entropy, or inherent uncertainty, in the distribution $p(N;\mu)$. Maximizing the entropy subject to a fixed set of constraints is equivalent to minimizing the Fisher information (Barron, 1986). In our case, if the input noise amplitude is fixed, then μ is the only free parameter of the spike count distribution. Using these facts, it is possible to write down a simple lower bound for

the Fisher information:

$$J_{\mathrm{LB}}(\mu) = \frac{1}{\sigma_N^2(\mu)}\left(\frac{\partial \mu_N}{\partial \mu}\right)^2, \tag{3.2}$$

where σ_N and μ_N are the standard deviation and mean of the spike count probability distribution, respectively.

As we shall show below, the mutual information between input and output and the probability of correct detection in the discrimination between two alternatives can both be expanded in powers of $J_{\mathrm{LB}}(\mu)$, underlining the fundamental nature of this quantity.

Neural modelers have previously applied the Fisher information to address the coding of stimulus information across many neurons (Paradiso, 1988; Seung and Sompolinsky, 1993). We proceed in the next section to calculate the mean μ_N and standard deviation σ_N of the spike count distribution as a function of the mean and variance of the input, which we will subsequently use to show that $J_{\mathrm{LB}}(\mu)$ passes through a SR peak as a function of the noise amplitude.

3.3 The Spike Count Distribution and the First Passage Time

The behavior of the subthreshold membrane voltage V of the leaky integrate-and-fire model is given by

$$C\frac{dV}{dt} = -\frac{V}{R} + I, \tag{3.3}$$

where C is the membrane capacitance, R the membrane resistance, and I the input current. If the input current is composed of a deterministic part μ and a white noise component ξ, we can rewrite eq. 3.3 as a Langevin equation:

$$\dot{V} = -\frac{V}{\tau} + \mu + \xi(t), \tag{3.4}$$

where $\tau = RC$ is a time constant, and we have set C to unity.

By definition, white noise is uncorrelated in time and obeys a relation of the type $\langle \xi(t)\xi(\tau) \rangle = 2\,D\,\delta(t-\tau)$, where D is a constant known as the noise amplitude.

The threshold (spiking) voltage θ is an absorbing boundary of the process: once the threshold θ is crossed, the voltage V is immediately reset to some set voltage V_0. A threshold crossing in this model is equivalent to one output spike of the unit.

The reset implies that the leaky integrate-and-fire unit "forgets" about its past history as soon as it crosses its threshold. Since the noise is uncorrelated, any correlations between successive threshold crossings will be due to the deterministic part of

the input alone. If μ is constant, then spike intervals are statistically uncorrelated. In this case, the mean μ_N and variance $\sigma_{\rm N}^2$ of the spike count distribution are completely determined by the moments of the probability distribution of individual intervals between spikes. Suppose that we observe the train of spikes from an integrate-and-fire unit over a period of T milliseconds. In all the examples to follow, we will take this time period to be $T = 200$ milliseconds, which corresponds roughly to the time scale of visual perception. A standard result from statistical renewal theory (Cox, 1962) states that, in the asymptotic limit, the mean and variance of the spike count become

$$\mu_{\rm N} = \frac{1}{\mu_{\rm FPT}}\, T \tag{3.5a}$$

$$\sigma_{\rm N}^2 = \frac{\sigma_{\rm FPT}^2}{\mu_{\rm FPT}^3}\, T, \tag{3.5b}$$

where $\mu_{\rm FPT}$ is the mean interspike interval (first passage time) and $\sigma_{\rm FPT}$ is the standard deviation of the interspike interval (equivalent to the width of the interspike interval distribution). Serial correlations between spike intervals caused by a time-varying input μ will typically increase $\sigma_{\rm N}^2$ by a multiplicative factor.

The first equation (eq. 3.5a) is strictly true under the "equilibrium" assumption, in which the start of the statistical process occurred long before the period of observation. In other words, we imagine that the model neuron had been spiking continuously at the background rate before spikes were being counted. If the signal is a small perturbation of the continuous background process, we are justified in using the statistical equilibrium result. Both expressions are exact if the interspike interval probability distribution is exponential, which will be approximately true if the mean

interspike interval μ_{FPT} is much longer than the characteristic time scale τ. Correction terms to σ_{N}^2 become important only if σ_{FPT} tends to zero; these corrections will be of order unity.

We have thus reduced the problem to calculating μ_{FPT} and σ_{FPT}^2. Given that the voltage of the integrate-and-fire unit is V at time t, determining the time until the next threshold crossing (or spike) is classically known as the first passage time (FPT) problem. The FPT distribution for the stochastic differential equation 3.4 was first solved in integral equation form by Siegert (1951). The problem and its solution in integral transform or integral equation form have been rediscovered many times in the context of neuronal models (for an overview, see Ricciardi, 1977).

The random noise fluctuations will lead to a probability distribution of voltages at time t for an integrate-and-fire unit that was initially at voltage V_i. This probability distribution is given by a Fokker-Planck equation associated with the Langevin equation 3.4:

$$\frac{\partial P(V,t;V_i,0)}{\partial t} = \frac{\partial}{\partial V}\left\{\left(-\frac{V}{\tau}+\mu\right)P\right\} + \frac{1}{2}\frac{\partial}{\partial V}\left\{\sigma^2\frac{\partial P}{\partial V}\right\},$$

which describes the evolution of the probability that the voltage is V at time t, given that the voltage was V_i at time $t = 0$.

The Fokker-Planck equation separates the input into two components: a mean current μ and variance of the current σ^2, both of which have units of inverse time. For future use, we note that we can define the dimensionless quantities $\hat{\mu} = \mu\tau/\theta$ and $\hat{\sigma} = \sigma\sqrt{\tau}/\theta$.

The probability of first passage of the voltage across the threshold θ in the time $[t, t+\Delta t)$, starting from voltage V_i, is

$$F(V_i,t) = \frac{d}{dt}\left[1 - \int_{-\infty}^{\theta} P(V,t;V_i,0)\,dV\right],$$

which satisfies the adjoint to the Fokker-Planck equation:

$$\frac{\partial F(V_i,t)}{\partial t} = \left\{-\frac{V_i}{\tau} + \mu\right\}\frac{\partial F}{\partial V_i} + \frac{1}{2}\sigma^2\frac{\partial^2 F}{\partial V_i^2}. \tag{3.5}$$

An integral transform and a change of variables transforms this equation into the confluent hypergeometric equation. In terms of the Laplace transform, $\tilde{F}(V_i,s) = \int_0^\infty \exp(-st)\,F(V_i,t)\,dt$, the solution is:

$$\tilde{F}(V_i,s) = U(\frac{s\tau}{2},\frac{1}{2},\frac{(\mu\tau - V_i)^2}{\sigma^2\tau})/U(\frac{s\tau}{2},\frac{1}{2},\frac{(\mu\tau-\theta)^2}{\sigma^2\tau}), \tag{3.6}$$

where $U(a,b,z)$ are confluent hypergeometric functions of the second kind and the solution has been constrained to satisfy the boundary conditions

$$\tilde{F}(\theta,s) = 1$$

$$\lim_{V_i\to-\infty}\tilde{F}(V_i,s) = 0.$$

Ricciardi (1977) writes eq. 3.6 in terms of parabolic cylinder functions.

Unfortunately, inverting the Laplace transform of eq. 3.6 analytically is generally

intractable. However, we can derive the mean and variance of the first passage time directly from the Laplace transform:

$$\mu_{\mathrm{FPT}} = -\left.\frac{d\tilde{F}(V_i, s)}{ds}\right|_{s=0} \tag{3.7a}$$

$$\sigma^2_{\mathrm{FPT}} = \left.\frac{d^2\tilde{F}(V_i, s)}{ds^2}\right|_{s=0} - \left(\left.\frac{d\tilde{F}(V_i, s)}{ds}\right|_{s=0}\right)^2 \tag{3.7b}$$

Computing second or higher moments, although straightforward, is not for the faint of heart. The results for the mean μ_{FPT} and variance σ^2_{FPT} of the first passage time distribution are given in the appendix A1.

3.4 The Stochastic Resonance

Written in terms of σ^2_{FPT} and μ_{FPT}, the lower bound $J_{\mathrm{LB}}(\mu)$ of the Fisher information is:

$$J_{\mathrm{LB}}(\mu) = \left(\sigma^2_{\mathrm{FPT}}\,\mu_{\mathrm{FPT}}\right)^{-1}\left(\frac{\partial \mu_{\mathrm{FPT}}}{\partial \mu}\right)^2 T. \tag{3.7}$$

We compute $J_{\mathrm{LB}}(\mu)$ from eq. 3.7 for $\mu\tau/\theta = 0.7$ and $T = 200$ msec and display the result as a function of the noise standard deviation in fig. 3.2a. An optimal noise level exists that maximizes $J_{\mathrm{LB}}(\mu)$; the peak in $J_{\mathrm{LB}}(\mu)$ is termed the stochastic resonance.

More realistic models of spiking membranes also lead to similar "resonant" behavior in $J_{\mathrm{LB}}(\mu)$, so this effect is by no means peculiar to the leaky integrate-and-fire model. The result for the original Hodgkin-Huxley model of the squid giant

axon (Hodgkin and Huxley, 1952) is displayed in fig. 3.2c; in this graph, the input current was set to 70% of the threshold current for sustained firing of spikes. Numerous differences exist between the leaky integrate-and-fire model and the Hodgkin-Huxley model, from the difference in the firing rate as a function of the input current (displayed in the right half of fig. 3.2), to the fact that the Hodgkin-Huxley model does not have a fixed voltage threshold for spiking. Nonetheless, both models display this form of stochastic resonance.

Since the expressions (A1) and (A2) for μ_{FPT} and σ^2_{FPT} are rather unwieldy, more insight can be gained by examining the asymptotic behavior of these quantities for subthreshold input currents μ in the weak and strong noise limits. As the noise variance σ^2 increases, $J_{\mathrm{LB}}(\mu)$ makes a smooth transition from the asymptotic behavior in the weak noise limit to the strong noise limit.

To simplify the notation, we switch to dimensionless quantities $\hat{\mu} = \mu\tau/\theta$ and $\hat{\sigma} = \sigma\sqrt{\tau}/\theta$. In these units, the threshold input at which the leaky integrate-and-fire unit will just barely cross the voltage threshold in the absence of noise is $\hat{\mu} = 1$. If $\hat{\mu} < 1$, then $\hat{\mu}$ is simply the normalized voltage that the integrate-and-fire unit reaches in the absence of noise. Define $\hat{\Delta} = 1 - \hat{\mu}$ as the effective dimensionless distance to the threshold.

We will study the behavior of the leaky integrate-and-fire unit in the range $0 < \hat{\Delta} \leq 1$. In effect, changing $\hat{\mu}$ in this range is equivalent to changing the threshold $\hat{\Delta}$. The Fisher information thus depends only on the threshold and the noise amplitude; no reference is made either explicitly or implicitly to the size of the signal to be

detected. We shall see in section 3.6 how $J_{\rm LB}$ couples to the signal strength in the small-signal expansion of various signal detection quantities.

To investigate the limit of weak noise, we assume that $\hat{\Delta}^2 \gg \hat{\sigma}^2$. Using Stirling's approximation for $n!$ in the hypergeometric series of eqs. A1 and A2, the method of steepest descents yields the asymptotic expansions in powers of $\hat{\Delta}^2/\hat{\sigma}^2$:

$$\mu_{\rm FPT} = \tau\frac{\sqrt{\pi}\hat{\sigma}}{\hat{\Delta}}\exp\Big(\frac{\hat{\Delta}^2}{\hat{\sigma}^2}\Big)\Big[1+\frac{1}{2}\frac{\hat{\sigma}^2}{\hat{\Delta}^2}+O\Big(\frac{\hat{\sigma}^4}{\hat{\Delta}^4}\Big)\Big] \tag{3.8a}$$

$$\sigma^2_{\rm FPT} = \mu^2_{\rm FPT} - 2\tau\Big[\gamma+\ln(2)+\ln\Big(\frac{\hat{\Delta}}{\hat{\sigma}}\Big)-\frac{3}{4}\Big(\frac{\hat{\sigma}^2}{\hat{\Delta}^2}\Big)-O\Big(\frac{\hat{\sigma}^4}{\hat{\Delta}^4}\Big)\Big]\mu_{\rm FPT}, \tag{3.8b}$$

where γ is the Euler-Mascheroni constant. Alternatively, we can use the integral representation of the confluent hypergeometric function (Abramowitz and Stegun, 1970) to derive the same expressions.

In the limit of $\hat{\Delta}^2 \gg \hat{\sigma}^2$, the Fisher information lower bound becomes

$$J_{\rm LB}(\hat{\mu}) = \frac{4T\hat{\Delta}^2}{\tau\hat{\sigma}^4}\,\frac{1}{\mu_{\rm FPT}}\left\{1-\pi^{-\frac{1}{2}}\left[2\gamma+2\ln(2)+\ln\Big(\frac{\hat{\Delta}}{\hat{\sigma}}\Big)\right]\frac{\hat{\Delta}}{\hat{\sigma}}\exp\Big(-\frac{\hat{\Delta}^2}{\hat{\sigma}^2}\Big)\right\}^{-1}.$$

Restoring the dimensional units, we have, to first order:

$$\boxed{J_{\rm LB}(\mu) = \frac{4T}{\sqrt{\tau^3\pi}}\,\frac{\Delta^3}{\sigma^5}\exp\Big(-\frac{\Delta^2}{\sigma^2\tau}\Big),} \tag{3.8}$$

in which Δ has units of voltage. As σ increases, eq. 3.8 rises quickly to a peak in $J_{\rm LB}$

and then subsequently falls off slowly in a power-law fashion.

Note that eq. 3.8 is of the same form as the standard SR expression for the SNR in the adiabatic approximation of the periodically modulated bistable potential (McNamara and Wiesenfeld, 1989; Jung, 1995; Gingl *et al.*, 1995):

$$\mathrm{SR} \propto \sigma^{-\alpha} \exp\left(-\beta \frac{\Delta^2}{\sigma^2}\right), \tag{3.9}$$

where $\alpha = 4$ and $\beta = 1/2$ for standard stochastic resonance. The noise-induced "resonance" or maximum in J_{LB} occurs when

$$\frac{\hat{\Delta}^2}{\hat{\sigma}^2} = \frac{\alpha}{2\beta} = \frac{5}{2}. \tag{3.10}$$

Caution should be exercised in interpreting results based on asymptotic expansions. A very incorrect use of the asymptotic expansions would predict that the firing rate $f(\mu) \propto 1/\mu_{\mathrm{FPT}}$ has a stochastic resonance! With that in mind, we should remark that the value of $\hat{\Delta}^2/\hat{\sigma}^2$ in eq. 3.10 is only marginally larger than unity, whereas the asymptotic expansion is valid in the regime $\hat{\Delta}^2/\hat{\sigma}^2 \gg 1$. However, the method of converging factors (Miller, 1952) can be used to improve and extend the range of validity of the asymptotic expansions. These converging factors can be used to compute the corrections to J_{LB} in eq. 3.8 to high accuracy in the vicinity of the optimal noise level σ_{max}; as an example, we give the converging factor for μ_{FPT} in appendix A1. A similar, but rather unwieldy, expression exists for the converging factor in the asymptotic expansion of σ^2_{FPT}.

Alternatively, we can compare the prediction of eq. 3.10 to the numerical solution of the full nonasymptotic equations. The optimal noise level $\sigma_{\max}$ for the leaky integrate-and-fire unit is plotted in fig. 3.3a as a function of the steady state voltage distance to the threshold. The position of the maximum was computed numerically from the full equations for μ_N and σ_N^2. Since the stochastic resonance expression is based on an asymptotic expansion, we plot in fig. 3.3b the "error" resulting from using eq. 3.10 instead of the full numerical solution to predict the optimal noise level. We note from the scale of the y-axis in fig. 3.3b that the quality of the approximation is quite good.

Now that the behavior of $J_{\rm LB}(\mu)$ in the limit of weak noise ($\hat{\Delta} \gg \hat{\sigma}$) has been established, let us consider the case where the noise is strong. In the limit in which $\hat{\sigma}^2 \gg 1 \geq \hat{\Delta}$, we have

$$\mu_{\rm FPT} = \tau \left[\sqrt{\frac{\pi}{\hat{\sigma}^2}} - \frac{(2\hat{\mu}-1)}{\hat{\sigma}^2} \right]$$
$$\sigma^2_{\rm FPT} = \Big(2\tau \ln(2)\Big)\mu_{\rm FPT} - \nu\tau^2 \Big[\frac{(2\hat{\mu}-1)}{\hat{\sigma}^2}\Big],$$

where $\nu = \pi - 2\ln(2) = 1.7553\ldots$.

In this limit, $J_{\rm LB}$ tends to

$$J_{\rm LB}(\mu) = \frac{2}{\pi \ln(2)} \frac{T}{\sigma^2}, \tag{3.11}$$

which is independent of τ, θ, and μ.

We remark that the σ^{-2} tail of the resonance peak in $J_{\rm LB}$ is much shallower than

the predicted σ^{-5} tail from eq. 3.8. The same σ^{-2} scaling (Jung, 1995) is commonly observed in systems showing stochastic resonance in response to periodic signals.

Since eq. 3.11 is independent of all parameters other than σ and T, we can conclude immediately that the relationship $J_{\mathrm{LB}}(\mu) \propto T/\sigma^2$ will hold also when we take the limits $\mu \to \infty$ or $\tau \to \infty$. For instance, it is easy to show that the Fisher information lower bound for a perfect integrate-and-fire unit with zero leak is $J_{\mathrm{LB}}(\mu) = T/\sigma^2$, which holds true for all possible noise amplitudes. Zero leak implies that no input threshold exists; the leaky integrate-and-fire unit will approximate the perfect integrator as long as the input is above the threshold ($\mu\tau > \theta$). Not surprisingly, if the input is above the threshold, $J_{\mathrm{LB}}(\mu)$ is a monotonically decreasing function of the noise amplitude—the optimal noise value is zero. In this case, noise will hurt any signal discrimination task.

We have assumed that no limit exists on the number of threshold crossings within the time period T. Appendix A.4 explores the asymptotic behavior of $J_{\mathrm{LB}}(\mu)$ when this assumption no longer holds.

3.5 Resonance versus Linearization

In their seminal work, McCullough and Pitts (1943) proposed a model of a neuron as a binary quantization device in which the neuron was either firing or not firing. Like all models for neurons, the McCullough-Pitts neuron has a threshold; adding noise to an input below threshold pushes the neuron above threshold, raising the average firing rate. Since the nature of the current threshold differs in the bistable

McCullough-Pitts neuron and the leaky integrate-and-fire unit, it is illustrative to compare the two models.

A McCullough-Pitts neuron is defined as having the following response $f(\mu)$ to an input μ and a noise value ξ:

$$f(\mu) = \begin{cases} 0 & \text{if } \mu + \xi < \theta, \\ 1 & \text{if } \mu + \xi > \theta. \end{cases}$$

Adding noise to the input of a quantization device is known in engineering as "dithering," or stochastic linearization. If we take the average of N presentations for each input μ to the McCullough-Pitts neuron, the graph of the mean $\langle f(\mu) \rangle$ will be a smoothed (dithered) version of the quantization step, as illustrated in fig. 3.4a. Increasing the noise amplitude in the leaky integrate-and-fire model also progressively linearizes the firing rate curve, smearing out the threshold nonlinearity, as illustrated in fig. 3.4b.

Since the spike count μ_N over a fixed time window is an estimate of the firing rate $f(\mu)$, we can rewrite the Fisher information (lower bound) for both models as

$$J_{\text{LB}}(\mu) = \frac{1}{\sigma^2_{f(\mu)}} \left[\frac{\partial}{\partial \mu} \langle f(\mu) \rangle \right]^2. \tag{3.12}$$

A strikingly similar formula holds for the spike output signal-to-noise ratio (SNR) in response to a sinusoidal input signal $\mu(t) = \mu + \Delta\mu \sin(\omega_0 t)$. Using an elegant and

simple argument by Gammaitoni (1995), we will show that

$$\mathrm{SNR}(\omega_0) \propto \frac{1}{f(\mu)} \left[\frac{\partial}{\partial \mu} f(\mu)\right]^2. \tag{3.13}$$

We will justify this last expression in more detail, but note first that the two equations 3.12 and 3.13 become identical in the limit of low firing rates! Stochastic resonance is usually defined in terms of the SNR, but we will argue that the more fundamental definition of SR in neuronal models is in terms of the Fisher information.

Why does the slope of $\langle f(\mu) \rangle$ play such a prominent role in these formulae? In short, because the ability to detect signal differences will depend primarily on the expected difference in spike counts, as long as we can neglect the variability in the output:

$$\begin{aligned} \Big\langle \mu_N(\mu_2) - \mu_N(\mu_1) \Big\rangle &= \Big(f(\mu_2) - f(\mu_1) \Big)\, T \\ &\approx \left[\frac{\partial}{\partial \mu} f(\mu)\right] \Delta\mu\, T, \end{aligned}$$

where $\Delta\mu = \mu_2 - \mu_1$ is the input difference and the brackets indicate the statistical average. A similar reasoning holds for the response to sinusoidal input: Imagine that we replace the sinusoid by a square wave switching between input levels $\mu + B$ and $\mu - B$. The periodic component of the leaky integrate-and-fire unit's spike train is determined by the average number of spikes fired during the 'up' phase of the square wave after subtracting the number of spikes fired during the 'down' phase. The power

$S(\omega_0)$ in the spike train at the driving frequency of the sinusoidal input is thus

$$S(\omega_0) \propto \left[\frac{\partial}{\partial \mu}\langle f(\mu)\rangle\right]^2,$$

as long as both the amplitude and frequency of the sinusoidal modulation are small (Gammaitoni, 1995).

For the McCullough-Pitts neuron, $\langle f(\mu)\rangle$ is simply the convolution of the probability density function p_ξ of the noise with a step function. Evaluated at a distance Δ from the threshold, the slope of $\langle f(\mu)\rangle$ is identical to the probability density function p_ξ at that point. Note that if the noise is Gaussian, the probability density function $p_\xi(\Delta) \propto \sigma^{-1}\exp(-\Delta^2/\sigma^2)$ is *already of the SR type* (eq. 3.9) as a function of σ! In other words, the value of the slope (probability density function) first increases and then decreases as the noise amplitude grows. Sandwiched between the threshold nonlinearity and the saturating nonlinearity, no other behavior is possible for the slope.

We can thus immediately write down the power $S(\omega_0)$ in the output of a McCullough-Pitts neuron in response to a periodic input at a frequency ω_0:

$$S(\omega_0) \propto \left[\frac{\partial}{\partial \mu}\langle f(\mu)\rangle\right]^2 = \frac{1}{2\pi\sigma^4}\exp\Big(-\frac{\Delta^2}{\sigma^2}\Big).$$

The slope dictates and guarantees the existence of stochastic resonance in the McCullough-Pitts neuron. This has led Gammaitoni (1995) to categorize SR in such systems as a special case of stochastic linearization.

However, the slope of the firing rate in the leaky integrate-and-fire model typically does *not* display "resonant" behavior. If the effective distance to the threshold $\hat{\Delta}$ is greater than $1/2$, the slope $f'(\mu)$ never reaches a maximum. Even for $\hat{\Delta} < 1/2$, the maximum is typically not pronounced. Of course, one can always introduce a saturating nonlinearity into the firing rate to force a maximum in the slope $f'(\mu)$, but firing rates at the SR optimum are typically nowhere near firing rate saturation (see section 3.8). Thus, while stochastic linearization and firing rate saturation together are sufficient conditions for SR, saturation is not a necessary condition.

For the integrate-and-fire model, SR in the Fisher information $J_{\mathrm{LB}}(\mu)$ arises through the conjunction of two effects: stochastic linearization—the increase in the slope of the firing rate curve—and the noise-induced increase in the spike count variance. Adding noise to the input comes with a price: the noise in the output increases as well. Thus an increase in the separation in mean firing rates does *not* imply an increase in the discriminability or Fisher information. Returning once again to eq. 3.2, we note that the numerator is affected by stochastic linearization, whereas the denominator measures the "noise" in the output. For the leaky integrate-and-fire model with $\hat{\Delta} > 1/2$, both numerator and denominator always increase as a function of the input noise σ; eventually, though, the denominator will dominate. As a consequence, $J_{\mathrm{LB}}(\mu)$ will first rise and then fall as a function of the input noise variance σ^2; this is the origin of the SR peak.

For sinusoidal inputs, it is no longer sufficient just to consider the power $S(\omega_0)$ in the output signal. To observe stochastic resonance, we need to divide $S(\omega_0)$ by

the background (noise) spectrum $N(\omega_0)$—this is, by definition, the SNR. If we can approximate the integrate-and-fire unit's spike train as a Poisson point process with $\sigma^2_{\mathrm{FPT}} = \mu^2_{\mathrm{FPT}}$, the background spectrum $N(\omega)$ is given by a very simple expression (Lukes, 1961):

$$N(\omega) = f(\mu) + 2\pi f(\mu)^2 \delta(\omega).$$

For subthreshold background inputs, the leaky integrate-and-fire unit's unmodulated spike train *is* very nearly Poisson, as indicated by eq. 3.8b. Systematic corrections to $N(\omega)$ can be computed from eq. 3.6, although we will not pursue these corrections here. Armed with the expressions for signal and noise spectra, we can write SNR $= S(\omega_0)/N(\omega_0) \propto [f(\mu)]^{-1}\ [f'(\mu)]^2$, which is eq. 3.13. Once again, the increase in the signal power due to stochastic linearization is counteracted by the increase in the noise power.

In the Poisson approximation, the variance $\sigma^2_f(\mu)$ of the firing rate is equal to the firing rate, so SNR $\propto J_{\mathrm{LB}}(\mu)$. Whether the question is to determine the Fisher information for a constant input μ or the SNR for a sinusoidal input $\mu(t)$, the answers to both are equivalent, at least to first order. (A second-order approximation for the SNR will introduce a weak dependence of the SNR on the driving frequency ω_0 of the sinusoidal signal; the first-order approximation is called *non-dynamical* SR (Gammaitoni *et al.*, 1995; Gingl *et al.*, 1995).) Certainly the simpler and arguably the more fundamental question in SR of threshold crossing models is the Fisher information.

To describe SR using the McCullough-Pitts model, we need to postulate how the

output noise behaves; a common assumption is that the output spike train is Poisson. In contrast, a spiking model neuron represents an implicit model for the variance in the firing rate and the background power spectrum. Note also that the effect of input noise on the firing rate function in a spiking model is generally more complicated than the simple picture given by the McCullough-Pitts description: adding uncorrelated noise to the leaky integrate-and-fire model is not equivalent to convolving the firing rate function with the probability density function for the noise.

Nonetheless, the McCullough-Pitts description is a valid abstraction of several important models of neuronal spiking, including the original Hodgkin-Huxley model for spikes in the squid giant axon. Unlike in the leaky integrate-and-fire model, the firing rate in the Hodgkin-Huxley model undergoes a first-order transition to a non-zero firing frequency as a function of the input current. The essence of a first-order transition, also known as a Hopf bifurcation, is captured in the step function of the McCullough-Pitts description. For the Hodgkin-Huxley model, the first-order transition in the firing rate is shown in fig. 3.2d. Note the large jump in the firing rate and the subsequent shallow slope of the curve and compare the graph to the corresponding one for the McCullough-Pitts neuron without noise in fig. 3.4a.

However, the McCullough-Pitts model is seriously flawed for other models of neuronal spiking, including the leaky integrate-and-fire model. These models do not have the bistable character of the original Hodgkin-Huxley model; instead, they commonly

undergo a second-order transition, as typified by a piecewise linear firing rate function:

$$f(\mu) = \begin{cases} c\left(\mu - \mu_\theta\right) & \text{if } \mu > \mu_\theta \\ 0 & \text{if } \mu \leq \mu_\theta, \end{cases}$$

where c is a constant and μ_θ is the input current threshold. Here, the firing rate function is continuous at the threshold μ_θ, but its derivative is not. For real cortical cells in slice, a piecewise linear model tends to be more appropriate than a bistable model for describing the adapted firing rate discharge in response to current injection (Mason and Larkman, 1990). The same behavior has been found in recordings from the intact network (Ahmed *et al.*, 1995): cortical neurons *in vivo* also respond linearly to increasing the current.

A linear firing rate function can arise in the absence of noise through the confluence of two effects: First, adding a fast transient potassium current ("A-current") (Connor and Stevens, 1971; Hille, 1992) can transform the bistable Hodgkin-Huxley model into a model with a second-order transition.

The theory of local bifurcations, when applied to sets of stiff differential equations that describe the evolution of the membrane potential during spikes, predicts that the firing rate behaves as $c\sqrt{\mu - \mu_\theta}$ near the threshold μ_θ for all non-hysteretic second-order transitions, except at degenerate points (Strogatz, 1991). When neurophysiologists plot the inverse of the first interspike interval, the instantaneous firing rate can oftentimes be fitted by a square-root-like function. The sustained firing rate, however, is subject to adaptation, the second effect mentioned above. Firing

rate adaptation, a complex, nonlinear, Ca^{2+}-dependent mechanism that opposes high sustained firing rates, progressively linearizes the square root behavior of the firing rate.

For a piecewise linear firing rate model, the slope $f'(\mu)$ will increase indefinitely under stochastic linearization. To determine whether $J_{\mathrm{LB}}(\mu)$ has a stochastic resonance, we need to have a model for the output noise; the firing rate $f(\mu, \sigma)$ alone is not sufficient information. It is reasonable, though, to assume the scaling

$$\sigma^2_{\mathrm{FPT}}(\mu, \sigma) \underset{\sigma \to 0}{\longrightarrow} \mu^2_{\mathrm{FPT}}(\mu, \sigma)$$

$$\sigma^2_{\mathrm{FPT}}(\mu, \sigma) \underset{\sigma \to \infty}{\longrightarrow} \chi\, \mu^{\rho}_{\mathrm{FPT}}(\mu, \sigma)$$

for the variance of the FPT distribution, where $\rho \leq 2$ and χ is a positive real number ($\chi = 1$ for $\rho = 2$). For the leaky integrate-and-fire unit, $\rho = 1$ and $\chi = 2\tau \ln(2)$.

In the limit $\sigma \to 0$, as σ^2_{FPT} approaches μ^2_{FPT} from below, the FPT distribution tends to the exponential distribution $P(\mathrm{FPT} = t) \propto \exp(-t/\mu_{\mathrm{FPT}})$. In this case, the spike count distribution is *necessarily* Poisson, for which we can write the Fisher information as:

$$J_{\mathrm{Poisson}}(\mu) = \left(2T \frac{\partial \sqrt{f(\mu, \sigma)}}{\partial \mu} \right)^2 \tag{3.14}$$

Equation 3.14 can be used as a straightforward estimate of $J_{\mathrm{LB}}(\mu)$ in its own right, but, more importantly, if $J_{\mathrm{Poisson}}(\mu)$ passes through a maximum, then so will $J_{\mathrm{LB}}(\mu)$.

(A sketch of the argument is given in appendix A3.)

Under the assumptions of appendix A3, the piecewise linear model will exhibit SR, but a quadratic model $f(\mu, \sigma = 0) \propto (\mu - \mu_\theta)^2$ will not. A power-law firing rate of this type is not entirely unreasonable and has been proposed in at least one serious model for the behavior of cortical cells in primary visual cortex (Carandini and Heeger, 1994; Carandini *et al.*, 1995).

3.6 Discriminability, Mutual Information, Detection Probability, Information Capacity, and Coherence

In section 3.5, the signal-to-noise ratio in response to periodic inputs was shown to be equivalent to the Fisher information lower bound $J_{\mathrm{LB}}(\mu)$. We will now show how other signal detection quantities can be approximated by power series in the Fisher information lower bound $J_{\mathrm{LB}}(\mu)$. In the first three cases, we assume a background current input μ, against which a fixed signal $s = \Delta\mu$ is to be observed.

The discriminability, or Mahalonobis distance, d' is a normalized measure of how far apart two probability distributions are. If we consider the leaky integrate-and-fire unit's spike count distribution with mean $\mu_N(\mu)$ and variance $\sigma_N^2(\mu)$, the relationship between d' and $J_{\mathrm{LB}}(\mu)$ can be expressed as

$$d' = 2\frac{\left|\mu_N(\mu + \Delta\mu) - \mu_N(\mu)\right|}{\left(\sigma_N(\mu + \Delta\mu) + \sigma_N(\mu)\right)} \approx \Delta\mu\sqrt{J_{\mathrm{LB}}(\mu)},$$

which is valid as long as the difference between $\mu_N(\mu + \Delta\mu)$ and $\mu_N(\mu)$ is small.

We now compute the mutual information between the spike count N and the presence or absence of the (weak) signal against a background input μ. For simplicity, suppose that the signal $s = \Delta\mu$ is likely to be present with probability $p = 1/2$. The mutual information is

$$I(s; N) = H(N) - H(N|s), \tag{3.15}$$

where

$$H(N) = -\sum_{N=0}^{\infty} p(N) \log p(N), \; p(N) = \frac{1}{2}p(N|\mu + \Delta\mu) + \frac{1}{2}p(N|\mu)$$

and

$$\begin{aligned} H(N|s) &= \frac{1}{2}H(N|\mu + \Delta\mu) + \frac{1}{2}H(N|\mu) \\ &= -\frac{1}{2}\sum_{N=0}^{\infty}\Big(p(N|\mu + \Delta\mu)\log_2 p(N|\mu + \Delta\mu) + p(N|\mu)\log_2 p(N|\mu)\Big). \end{aligned}$$

For two Gaussians of a given variance and different means $\mu_N^{(1)}$ and $\mu_N^{(2)}$, the mutual information between the two distributions will be tightly approximated by the following upper bound

$$I \geq 1 - \log_2\left\{1 + \exp\left[-\frac{(\mu_N^{(1)} - \mu_N^{(2)})^2}{4\sigma_N^2}\right]\right\}.$$

The spike count distributions will tend to Gaussians as $T \to \infty$. As long as $T > \mu_{\mathrm{FPT}}$, we can use a Gaussian approximation for the spike count distribution to obtain

$$I \approx 1 - \log_2\left\{1 + \exp\left[-\frac{(\Delta\mu)^2}{4} J_{\mathrm{LB}}(\mu)\right]\right\} \tag{3.16}$$

If $\Delta\mu$ is small compared to all other quantities, the mutual information is approximately

$$I \approx \frac{(\Delta\mu)^2}{8 \ln 2} J_{\mathrm{LB}}(\mu). \tag{3.17}$$

An approximation using continuous Gaussians is often hard to justify rigorously, because the underlying spike count distributions are discrete and typically highly skewed. Nonetheless, the error in the approximation turns out not to be severe. The quality of the approximation can be assessed in fig. 3.5, where we compare it to the mutual information computed by simulating the original stochastic differential equation 3.4 for the leaky integrate-and-fire unit. Note that, since detection, in this case, is a "yes-no" decision, we have at most $\log_2(2) = 1$ bits of information.

A measure closely related to the mutual information between the presence or absence of a signal and the spike count is the probability of correct detection in the discrimination between two alternatives. In a two-alternative forced-choice (2AFC) test, the subject or animal is presented with two stimuli in random order, in which one stimulus contains the signal, the other does not. The probability of correct detection

based on the spike count output of a neuron is

$$P(d) = \sum_{N=0}^{\infty} p(N|\mu)p(N' > N|\mu + \Delta\mu) + \frac{1}{2}\sum_{N=0}^{\infty} P(N|\mu)p(N|\mu + \Delta\mu), \tag{3.18}$$

where $p(N|\mu)$ is the probability of observing N spike counts in the distribution without the signal and $p(N' > N|\mu + \Delta\mu)$ is the probability of observing more than N spike counts in the presence of the signal.

Once again, we can use the Gaussian approximation to derive an approximate expression for $P(d)$:

$$\begin{aligned} P(d) \approx & \frac{1}{2}\,\mathrm{erfc}\left(-\frac{\Delta\mu}{\sqrt{2}}\sqrt{J_{\mathrm{LB}}(\mu)}\right) \\ & - \frac{1}{\sqrt{\pi}}\left(1 - \frac{1}{\sqrt{2}}\right)\left(\frac{\Delta\mu}{\sqrt{2}}\sqrt{J_{\mathrm{LB}}(\mu)}\right)\exp\left(-\frac{\Delta\mu^2}{2}J_{\mathrm{LB}}(\mu)\right) \end{aligned} \tag{3.19}$$

The leading order term is

$$P(d) = \frac{1}{2} + \frac{\Delta\mu}{2\sqrt{\pi}}\sqrt{J_{\mathrm{LB}}(\mu)}.$$

If, instead, one stimulus, rather than two, is presented, the leading order term in the probability of correct detection is

$$P(d) = \frac{1}{2} + \frac{\Delta\mu}{2\sqrt{2}\sqrt{\pi}}\sqrt{J_{\mathrm{LB}}(\mu)}.$$

The results for the probability of detection in a two-alternative forced-choice test are presented in fig. 3.6, where we once again compare simulation results to our asymptotic theory. We compare the behavior of $P(d)$ and the mutual information I as a function of both μ and σ in fig. 3.7, which plots the results of eqs. 3.16 and 3.19.

The relationships between $J_{\mathrm{LB}}(\mu)$ and the quantities described in section 3.6 are summarized in table 3.1.

Table 3.1 Small Signal Relationships

Discriminability d'	Information I	2AFC Detection $P(d)$
$\Delta\mu\sqrt{J_{\mathrm{LB}}(\mu)}$	$\frac{(\Delta\mu)^2}{(8\ln 2)}J_{\mathrm{LB}}(\mu)$	$\frac{1}{2}+\frac{\Delta\mu}{2\sqrt{2\pi}}\sqrt{J_{\mathrm{LB}}(\mu)}$

One can generalize the expression for the mutual information to the case of multiple stimuli. Suppose that inputs μ are drawn randomly from some probability density function $p(\mu)$, where μ is no longer restricted to a discrete set of values, but can vary continuously. Then Shannon's mutual information between μ and the spike count N is

$$I(\mu,N)=\int \log_2\left\{\sqrt{\frac{J_{\mathrm{LB}}(\mu)}{2\pi e}}\right\}p(\mu)d\mu-\int \log_2\left[p(\mu)\right]p(\mu)d\mu, \qquad (3.20)$$

provided that the Gaussian approximation is valid.

By taking the functional derivative of eq. 3.20 and taking into account the La-

grangean constraint that $\int p(\mu)\, d\mu = 1$, it is easy to show that the probability distribution $p(\mu)$ that maximizes $I(\mu, N)$ is

$$p(\mu) = \frac{\sqrt{J_{\mathrm{LB}}(\mu)}}{\int_{-\infty}^{\infty} \sqrt{J_{\mathrm{LB}}(\mu)}\, d\mu}.$$

The maximum in the information is known as the capacity and was first derived in terms of the mean and variance of a given FPT distribution by Stein (1967). Since the capacity should be a finite number, the range of possible firing rates must be bounded; to achieve this constraint, we consider the case where each spike of the leaky integrate-and-fire unit is followed by a strict refractory period t_{ref}.

Substituting the optimal probability distribution into eq. 3.20, we can write the capacity C as:

$$C = \log_2 \left[\int_{-\infty}^{\infty} \sqrt{\frac{J_{\mathrm{LB}}(\mu)}{2\pi e}} d\mu \right] \tag{3.21}$$

To compute the capacity, the distribution of inputs is changed for each fixed noise amplitude so that the information transfer through the leaky integrate-and-fire unit is maximal: a pure translation of the input distribution is equivalent to changing the input threshold or adding a bias input; changing the synaptic weights for different inputs can effect the required multiplicative changes. As fig. 3.8 illustrates, increasing the noise amplitude will always decrease the capacity. Not shown is this surprising fact: Suppose we fix the probability distribution to be optimal for some noise level σ. If σ is now varied, but $p(\mu)$ remains fixed, then $I(\mu, N)$ is still a decreasing function

of σ. Therefore, tuning the thresholds or the input distribution so that the leaky integrate-and-fire unit reaches capacity does not imply that the system is tuned to a stochastic resonance. Only when the input probability distribution and thresholds are constrained to remain *non-optimal* will SR occur.

Lastly, we point out that a signal-to-noise ratio can be defined for continuously varying, non-periodic input signals by using the measure of coherence. How to define the SNR is not as clear-cut as in the periodic signal case and depends on the nature of the signal processing problem (Carter, 1985). The coherence $\gamma^2(\omega)$ between the input signal μ and the spike count N, however, is well defined as the power spectral measure quantifying the degree of linearity in the system. The coherence simply reduces to a number between zero and one. The ratio of the linear to the nonlinear component in neuronal signal transduction, $\gamma^2(\omega)/(1-\gamma^2(\omega))$, is then one possible definition of SNR (Carter, 1985). Collins *et al.* (1995a) computed the coherence γ^2 for the Fitzhugh-Nagumo model of neural activity as the frequency of the input μ's temporal variation tends to zero:

$$\lim_{\omega\to 0}\gamma^2(\omega) = \frac{\left\{\int\left[\left(\mu-\langle\mu\rangle\right)\left(f(\mu)-\langle f(\mu)\rangle\right)\right]p(\mu)d\mu\right\}^2}{\int\left(\mu-\langle\mu\rangle\right)^2 p(\mu)d\mu \quad \int\left(f(\mu)-\langle f(\mu)\rangle\right)^2 p(\mu)d\mu}, \tag{3.22}$$

where $\langle\cdots\rangle$ indicates the average over μ, $f(\mu)$ is the firing rate in response to μ, and the input μ varies extremely slowly in time. Suppose that the variance in μ over time

is very small. Then a Taylor expansion around $\langle\mu\rangle$ yields

$$\lim_{\omega\to 0} \gamma^2(\omega) \approx \left\langle \left(\mu - \langle\mu\rangle\right)^2 \right\rangle J_{\mathrm{LB}}(\langle\mu\rangle)$$

as a first order approximation. To be precise, we have implicitly assumed a time scale T for the variations in μ here, where T is large; otherwise, one needs to employ an empirical formula to compute the denominator of eq. 3.22, as did Collins *et al.* (1995b).

Computing $\gamma(\omega)$ for arbitrary frequency ω would allow one to compute the total information in the spike train from Shannon's formula $I \sim \int \log_2[1 + \mathrm{SNR}(\omega)]\, d\omega$. Compared to the information in the firing rate over a fixed time scale T, the information in the spike train over all possible time scales can potentially be significantly larger. However, the information in the timing of spikes is much more difficult to compute analytically. We will return to this point in section 3.8.

3.7 Reducing the Threshold

Adding noise is certainly not the only way to improve signal detection. One of the obstacles to signal detection, the input (current) threshold, can be removed quite simply: turn subthreshold signals into suprathreshold signals by adding a constant bias term that raises the background level of input. In other words, instead of changing the input noise variance σ^2, we change the background μ, which we here call the bias. (The input signal $s = \Delta\mu$ that is to be detected remains fixed throughout.) From

another perspective, raising μ is equivalent to reducing the input current threshold.

From eqs. 3.8 and 3.10, we deduce that the maximum of J_{LB} scales as

$$J_{\mathrm{LB,\,max}}(\mu) \propto \frac{1}{(\theta - \mu\tau)^2}.$$

In other words, as the bias input μ approaches the current threshold, the Fisher information increases rapidly. Note, though, that we do not mean to imply that $J_{\mathrm{LB,\,max}}(\mu)$ diverges as the threshold is approached: since the (optimal) input noise amplitude tends to zero, we must take into account correction factors to eq. 3.5b for the spike count variance σ_{N}^2; σ_{N}^2 will approach a constant, preventing $J_{\mathrm{LB,\,max}}(\mu)$ from diverging.

Above the input threshold, stochastic resonance is no longer an effect: the noise amplitude can no longer be tuned to some nonzero value to improve signal detection performance. In fact, given the choice to either set the noise amplitude or the bias, the optimal strategy for signal detection is always to change the bias.

If we apply a bias to effectively reduce the threshold to zero, we will be interested in a different question than stochastic resonance, but one that is similar: does an optimal bias exist for the detection of weak, constant signal differences in the presence of a fixed, non-zero amplitude of the noise? The answer is, in general, no. As illustrated in fig. 3.9a for the leaky integrate-and-fire model, the information in the firing rate is typically a monotonically increasing and saturating function of the bias μ.

For very weak signals, the mutual information between signal and spike count, as defined in section 3.6, is linearly proportional to the lower bound of the Fisher

information $J_{\mathrm{LB}}(\mu)$. In section 3.5, we explored the relationship between the firing rate slope $f'(\mu)$ and the Fisher information $J_{\mathrm{LB}}(\mu)$. (The slope is proportional to the average spike output ΔN in response to an input signal $s = \Delta\mu$.) Recall that $J_{\mathrm{LB}}(\mu)$ can possess a maximum as a function of the noise variance σ^2 in cases where $f'(\mu)$ is a strictly increasing function of the noise variance. As a function of the bias μ, the converse can exist: $J_{\mathrm{LB}}(\mu)$ can be a monotonically increasing and saturating function, whereas $f'(\mu)$ possesses a maximum. We illustrate this last point by comparing the two panels of fig. 3.9. With the noise amplitude set as in fig. 3.9, $J_{\mathrm{LB}}(\mu)$ and, hence, the information $I(\mu)$ do not pass through a maximum as a function of the bias μ; instead, these quantities increase and then saturate at some level. In contrast, the average difference in spike counts $\Delta N = f'(\mu)\Delta\mu T$ in fig. 3.9b does show a peak as a function of μ for the same fixed noise amplitude as in fig. 3.9a.

There are, however, instances in which an optimal bias *does* exist. If the firing rate jumps suddenly as a function of the input, signal differences that straddle the firing rate discontinuity are most easily detected. Neural models with a first-order ("hard") input threshold, such as the McCullough-Pitts, Fitzhugh-Nagumo, and Hodgkin-Huxley models, thus possess optimal bias levels. Also, increasing the bias too much can push a neuron to its maximal saturating firing rate, leading to a decrease in $J_{\mathrm{LB}}(\mu)$; from an information transmission perspective, such extremes are best avoided. We consider firing rate saturation in appendix A.4.

In eq. 3.11 we showed that $J_{\mathrm{LB}}(\mu)$ is bounded asymptotically by T/σ^2. As a consequence, the amount of information transmitted during a fixed time period T and in

the presence of a fixed noise amplitude can still be quite low, regardless of the bias, as shown in fig. 3.9a. Note from the same figure that it is not absolutely necessary for the leaky integrate-and-fire unit to be operating above the threshold for the information or Fisher information to be close to its asymptotic maximum. The information $I(\mu)$ already approaches saturation for bias levels μ below the input current threshold. We conclude that the only effective way of improving signal detection for a fixed noise amplitude is to average over longer times T.

Reducing the input threshold to zero is, of course, a trivial approach to the problem of signal detection. Other biological or computational constraints may exist that force the bias or background input to remain constant. One computational model (Stemmler *et al.*, 1995) uses SR to enhance the detection of weak signals when the bias, in fact, becomes negative. Many computations are not possible in the absence of input thresholds or other nonlinearities in signal transduction (Hopfield, 1994). In addition, adding a bias will change the firing rate encoding of previously suprathreshold inputs more strongly than uncorrelated noise. Inspection of fig. 3.4b confirms that white noise will only weakly change the firing rate for inputs $\mu\tau/\theta \gg 1$ that are far above threshold, while changing the firing rate for inputs below threshold more strongly.

3.8 Spike Times or Firing Rate?

In computing a spike count or firing rate, we average over individual spikes, thereby discarding any information contained in the spike times. The longer the time period T over which the spike train is observed, the more dramatic is the loss of information

by ignoring spike times: the information in a spike interval code grows *exponentially* faster than the information in a firing rate code as a function of T (Stein, 1967). For this reason alone, recent attention has focused on the role of noise in changing the information carried in the interspike intervals (Zador and Bulsara, 1996; Levin and Miller, 1996).

What are the merits of using the individual spike times instead of the mean firing rate in the detection of very weak signals? Here we need to distinguish between periodic and aperiodic inputs.

Weak signals that are periodic are most naturally encoded in the intervals between spikes, not in the firing rate. Experimental evidence that noise improves the periodicity of spike timing in such an interval code has been found for crayfish mechanoreceptors (Douglass *et al.*, 1993), sensory cells in the shark (Braun *et al.*, 1994), and the cricket cercal system (Levin and Miller, 1996).

Suppose, however, that the input is *not* periodic. We note first that, unlike in a digital computer or (possibly) olfactory cortex, no "clock" is known to exist in neocortex. Without a clock, only the time of a spike relative to the last spike can matter. Therefore, any spike time code in neocortex will, in fact, be a spike interval code. Information theory predicts that the optimal spike interval code will be Poisson, i.e., that the intervals (or first passage times) will be exponentially distributed. Recall that additive noise and a constant, subthreshold input current μ to the leaky integrate-and-fire unit will also lead to (nearly) exponentially distributed intervals. Note that the spike train in a Poisson firing rate code is statistically indistinguishable from the

spike train in an optimal spike interval code.

Computing the information in the spike times for an arbitrary signal distribution can be a daunting task. Zador and Bulsara (1996), for instance, use Monte Carlo methods to numerically compute the (conditional) entropy of the first passage time distribution, thereby directly calculating the mutual information between a random Gaussian input signal and the output spike train of a leaky integrate-and-fire unit. Such a purely numerical approach is not hampered by restrictive assumptions about the nature of correlations in the spike train.

There is one simpler model of spike train encoding for which the information in the spike times has been calculated analytically in a perturbation expansion by Bialek *et al.* (1993): take the superposition of a random Gaussian signal $\mu(t)$ and random Gaussian noise $\xi(t)$, choose a threshold level Δ, and assign a spike to each time the wave-form of signal plus noise crosses Δ. Note that the hallmarks of the leaky integrate-and-fire unit, namely the exponential decay of the membrane potential and the reset to zero potential after each threshold crossing, are absent. For the simpler model, DeWeese (1995) computed the maximal information rate R in the threshold crossings (spike times) of a random Gaussian signal $\mu(t)$ in the presence of random Gaussian noise with variance σ as:

$$R = \frac{f(\langle\mu\rangle)\Delta^2}{2\ln 2\,\sigma^4}\left\langle\left(\mu - \langle\mu\rangle\right)^2\right\rangle, \tag{3.23}$$

where $f(\langle\mu\rangle)$ is the mean frequency of threshold crossings when the signal μ is not changing. The same threshold crossing model was the basis for Gingl *et al.*'s (1995) theory of nondynamical SR—the expression they derived for the SNR is the same as DeWeese's information rate up to a constant of proportionality. We point out that the firing rate $f(\mu)$ is described by a different function than in the leaky integrate-and-fire model; in fact, the firing rate of the simple threshold crossing model is infinite when the Gaussian noise is uncorrelated in time. Nevertheless, let us compare eq. 3.23 for the threshold crossing model to eq. 3.17 for the leaky integrate-and-fire unit after dividing by T:

$$R = \frac{f(\mu)\Delta^2}{2\ln 2\,\sigma^4}\,(\Delta\mu)^2\,. \tag{3.24}$$

The concordance between the two expressions for the information rate may be surprising at first: should not the information rate in the spike times be much higher? But remember that eq. 3.17 is based on a *weak signal* expansion. A signal is weak when either the amplitude $\Delta\mu$ or the observation period T is so small that only a few spikes are fired within T in response to the stimulus. When the number of spikes is low, firing rates and spike times are nearly equivalent: both the information in a firing rate code and the information in a spike interval code start out linear in T. A spike interval code, however, will remain linear in T as the number of spikes increases. Note that the natural extension of eq. 3.24 to time-varying inputs consists of replacing $(\Delta\mu)^2$ by $\left\langle\left(\mu - \langle\mu\rangle\right)^2\right\rangle$, which will hold for all time intervals T.

To show that the stimulus information in the firing rate is nearly equal to the

information in the spike intervals in the signal detection limit, assume that our optimal mean firing rate code and an unspecified spike interval code are indistinguishable in terms of the firing rate. Further, suppose that we insist that the model neuron achieve a certain level of accuracy in the detection of a weak signal in a two-alternative forced-choice (2AFC) test. A system that correctly chooses the signal over the background with probability $P(d) = 0.75$ is said to be at the signal detection threshold; in contrast, a system that guesses randomly has a probability $P(d) = 0.5$ of being correct. The signal detection threshold is the criterion at which the system (here, the model neuron) begins to detect the signal "reliably."

How many spikes does an optimally noise-tuned integrate-and-fire unit fire in response to a signal $s = \Delta\mu$ that is correctly detected 75% of the time? If detection is based on the spike count in a $T = 200$ msec time window, the average number of spikes can be estimated from eqs. 3.19 and A1. The result is displayed in fig. 3.10 as a function of the background input μ. The figure shows that the decision is based on an average difference in the number of spikes that is no greater than one single spike.

Signal detection is thus *not* the result of averaging over many spikes, even though spikes are produced by a predominantly stochastic mechanism. As such, SR stands in contrast to other applications of noise in which the addition of input noise is followed by an averaging over time or space.

Since a firing rate code at the signal detection threshold differs little in information content from an equivalent interval code, other considerations, such as how decodable and reliable the code is, will play a role in deciding which representation of information

is preferable. We point out that encoding information in an interval code has one potential disadvantage:

A sophisticated interval code is more prone to noise distortion than a simple-minded firing rate code. Distortion here quantifies the "distance," or error, between the original input signal and the signal reconstructed by decoding the output (see chapter13, Cover and Thomas, 1991). Shifting a single spike by a few milliseconds can change dramatically the meaning of an encoded message in a spike interval code, much more so than in a firing rate code. Thus, while the information content is higher in an interval code, the minimum information needed to successfully transmit signals within a prescribed error bound is also higher. In the limit of low firing rates, the problem of distortion is compounded by a lack of error correction, since any type of correction would require the transmission of additional intervals.

We thus conjecture that the encoding of very weak signals in single neurons will involve the firing rate, since the timing of spikes is less likely to matter than whether a spike occurred at all.

3.9 Conclusion

The input signal to a simple neuron model such as the leaky integrate-and-fire unit can be constant, periodic, or arbitrarily varying in time. Regardless of the nature of the signal, the relevant information quantities have the same asymptotic scaling properties as a function of the noise amplitude. Whether we wish to determine the mutual information between input and spike output, the signal-to-noise ratio for

sinusoidal input, or the probability of correctly detecting a constant signal within a limited time—for subthreshold inputs, these quantities are given by the class of equations for SR:

$$\text{SNR}, d', I, P(d) \propto \sigma^{-\alpha} \exp\left(-\beta \frac{\Delta^2}{\sigma^2}\right),$$

where σ is the standard deviation of the input noise, and Δ is an effective input current threshold. The exponents α and β depend on the quantity on the left-hand side and may differ for different models of neuronal spiking.

A model for neuronal spiking need not possess a true voltage threshold to exhibit SR; an input current threshold for sustained spiking is sufficient. The Hodgkin-Huxley model is one example of a model without a true voltage threshold. SR can still exist for periodic signals that lie completely above the input current threshold, as shown by Bulsara *et al.* (1994) for the non-leaky integrate-and-fire model. SR in the mutual information for non-periodic inputs, however, seems to occur only for signals with a component below the input current threshold (as was also observed by Zador and Bulsara, 1996).

With the SR expression and other results stated or derived in sections 3 and 3.6, we can summarize the results for the mutual information, for instance, in these scaling relationships:

(1) Doubling the signal difference quadruples the mutual information.

(2) Averaging over twice as long a time leads to a two-fold increase in the mutual information.

(3) For subthreshold inputs, the maximum information is inversely proportional to the squared distance of the steady-state voltage from the threshold voltage.

(4) Above threshold, doubling the noise variance σ^2 halves the information.

Whether the nervous system can utilize any stochastic effect to its advantage is primarily constrained by the limited time available for processing. In the case of "classical" stochastic resonance in response to periodic signals, the nervous system must be able to estimate whether a periodic signal is present without resorting to power spectral measures and averaging over long times. The question of whether efficient estimation methods for such signals exist is an open question.

Experimental studies, most of which have focused on sinusoidal input signals, have shown that artificial manipulation of the noise level can lead to stochastic resonance in neuronal systems. As yet, no experimental proof exists as to whether any part of the nervous system in any animal actually uses noise to enhance information transmission.

Noise will always be present, of course, in any fully functioning nervous system. Can organisms use SR by adapting neuronal thresholds and other parameters to match the pre-existing and inescapable noise present in the biological hardware and the stimulus? We have argued to the contrary: fully optimizing the threshold of a leaky integrate-and-fire unit in the presence of a fixed noise amplitude implies adapting the threshold past the SR optimum for that noise amplitude. For instance, the retinae of both vertebrates and invertebrates possess elaborate adaptation mecha-

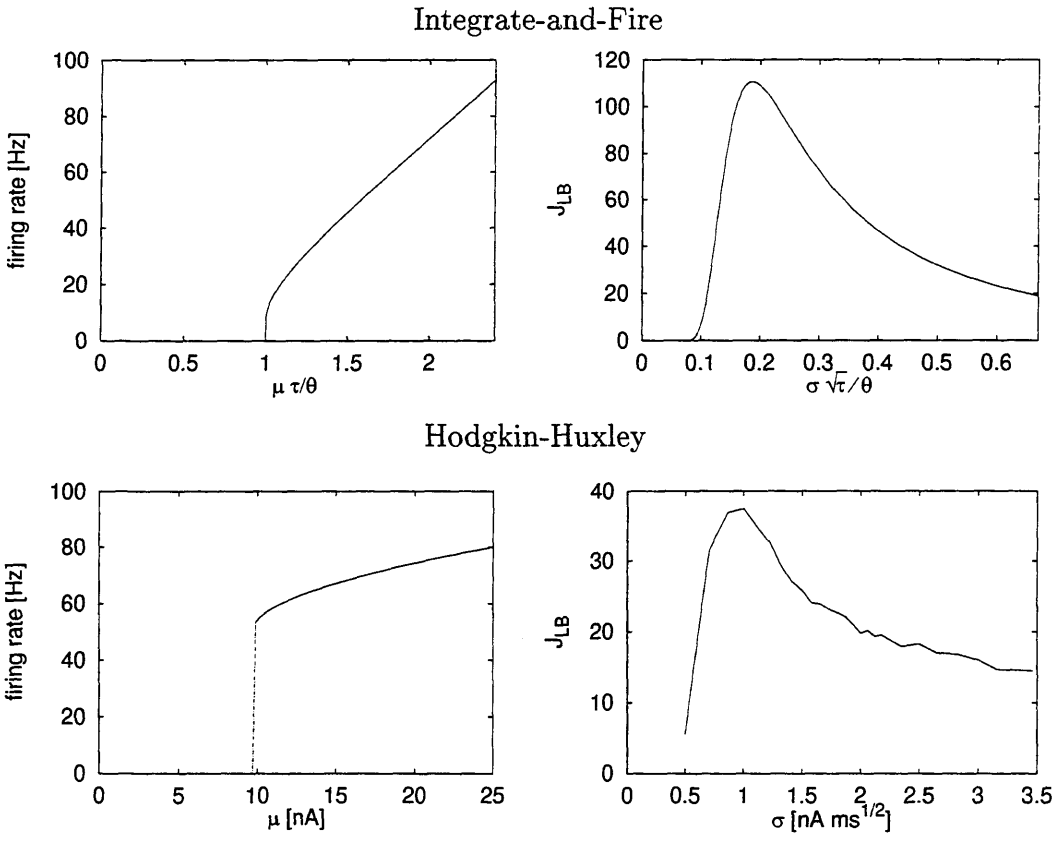
nisms to optimize information transfer under different light levels. At very low light levels, the noise σ^2 is dominated by quantum fluctuations (photon noise), which is simply proportional to the mean luminance and can hence be manipulated experimentally. But because of retinal adaptation, no SR effect is observed in the detection of weak, constant stimulus contrasts based on the output of retinal ganglion cells (Barlow and Levick, 1969). Thus we expect biology to use SR when there are computational or biophysical constraints on the neuronal thresholds and the distribution of inputs.

The true test of SR in biological systems lies in whether the system can "learn" to generate the proper internal noise when the threshold and the distribution of input signals are fixed. Regulating the noise could be achieved by changing the probability of transmitter release at synapses within the neuronal network. It is well-known that synapses in the hippocampus (archicortex) operate in an all-or-none fashion (Bekkers and Stevens, 1990; Hessler *et al.*, 1993); the probability of failure typically lies between 70-95% each time the synapse is activated. Long term potentiation (LTP) is known to change this synaptic transmission probability (Bolshakov and Siegelbaum, 1995). Even without synaptic failure, the recruitment of associative fibers from other cortical neurons can lead to a change in the noise level, because the firing of cortical cells, even at high rates, is known to be notoriously variable (Softky and Koch, 1993). When large or extended stimuli in visual or auditory frequency space are present, the associative fibers will convert the irregular firing of cells in other cortical regions into noise in the input current to the cell.

Temperature-related noise sources, on the other hand, are unlikely to be subject

to direct control by a biological system. These noise sources originate from within the neuronal membrane itself and include thermal current noise, conductance fluctuations, and the activity of electrogenic ion pumps (for a recent review, see Longtin and Hinzer, 1996). We therefore hypothesize that learning the appropriate noise amplitude to achieve SR will occur at the network level.

Furthermore, the biological system must be able to adapt the noise level to changing conditions. Any real world system will be faced with a range of input signals, not just one single constant signal whose presence the system needs to detect. One crucial test of whether the nervous system actively uses stochastic resonance occurs when the statistics of the inputs change. Does the noise level change as the input probability distribution changes? For instance, additional information about the context of the input can bias the input statistics. A computational system using stochastic resonance will encode this additional information by changing the noise amplitude. In this case, the traditional line between noise and information is blurred: the noise becomes the information.


Integrate-and-Fire
firing rate [Hz]
μ τ/θ
J_LB
σ √τ/θ
Hodgkin-Huxley
firing rate [Hz]
μ [nA]
J_LB
σ [nA ms^1/2]

The right-hand column displays the sustained firing rate as a function of the input in the absence of noise for the leaky integrate-and-fire model (fig. 3.2b) and the standard Hodgkin-Huxley model for a patch of squid giant axon at $T = 6.7^o$ Celsius (fig. 3.2d). The left-hand column displays the Fisher information lower bound $J_{\rm LB}(\mu)$ as a function of the root-mean-square noise amplitude for an input μ that is equal to 70% of the threshold input. For the integrate-and-fire model (fig. 3.2a), $J_{\rm LB}(\mu)$ is based on the analytical expressions of eqs. A1 and A2 for the variance and mean of the first passage time (interspike interval) distribution. For the Hodgkin-Huxley model (fig. 3.2c), a spike count histogram ($N = 10,000$ trials of 200 ms) was derived from simulations of the Hodgkin-Huxley equations using the simulation package NEURON (Hines, 1993). These histograms were subsequently used to estimate $J_{\rm LB}(\mu)$. In both cases, $J_{\rm LB}(\mu)$ was renormalized by the square of the threshold input to facilitate the comparison between integrate-and-fire and Hodgkin-Huxley models. Both models display stochastic resonance: tuning the noise amplitude maximizes $J_{\rm LB}(\mu)$. Note the quick rise to the peak and slow subsequent decay, the characteristic hallmarks of stochastic resonance. Of the differences between the integrate-and-fire model and the Hodgkin-Huxley model, these are the most important in this context:

(1) The firing rate undergoes a first-order transition to nonzero firing in the Hodgkin-Huxley model. In addition, the firing rate curve is hysteretic (not shown).

(2) The threshold in the Hodgkin-Huxley model is not constant, but a function of the voltage history of the neuronal membrane.

The variable threshold characteristics are captured in the following model for the threshold:

$$\theta = \theta_0 + \Delta\theta(t)$$

$$\frac{d}{dt}\Delta\theta(t) = -\frac{\Delta\theta(t)}{\tau_{\Delta\theta}} + \lambda\Big(V - V_0\Big),$$

where λ is a positive constant and the term $\lambda\Big(V - V_0\Big)$ increases the threshold during prolonged membrane voltage excursions away from the reset voltage V_0 that do not lead to a spike (threshold crossing).

Figure 3.2:

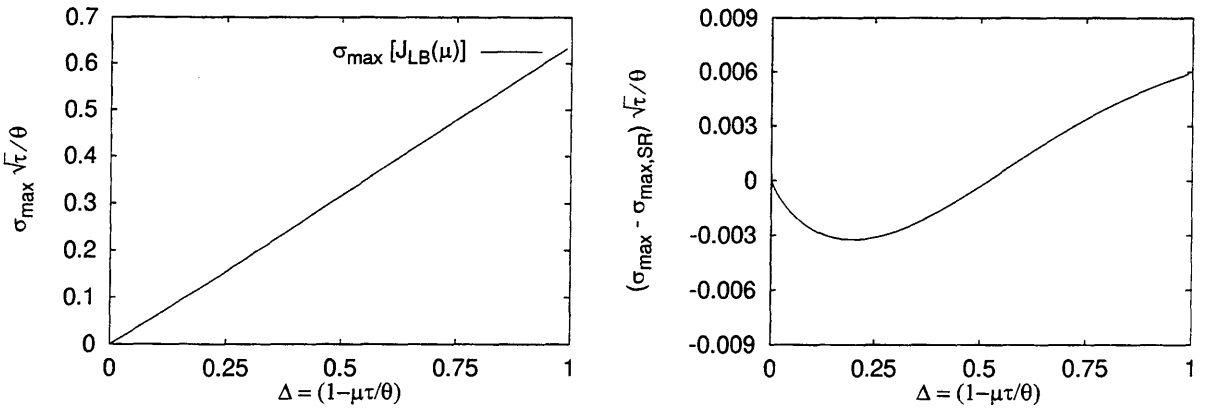


Figure 3.3: Figure (a) displays the optimal noise level for the Fisher information lower bound J_{LB} (solid line) as a function of the normalized distance Δ from the background voltage to the threshold voltage. $\sigma_{max}\sqrt{\tau}/\theta$ plotted on the y-axis represents a root-mean square voltage, where the normalization $\sqrt{\tau}/\theta$ stems from the Fokker-Planck eq. 3.5 (see text for details). Equation 3.10 predicts that the optimal root-mean square voltage of the noise is $V_{\mathrm{rms}} \approx 0.6\Delta$, or roughly 60% of the additional voltage input needed to reach threshold in the absence of noise. Figure (b) displays the absolute difference between the true optimal noise level for $J_{LB}(\mu)$ and the noise level predicted by the stochastic resonance equation 3.10, which is based on an asymptotic expansion. Note that the scale on the y-axis is two orders of magnitude smaller than in fig. (a), indicating that the deviation of σ_{max} from perfect linearity is minimal.

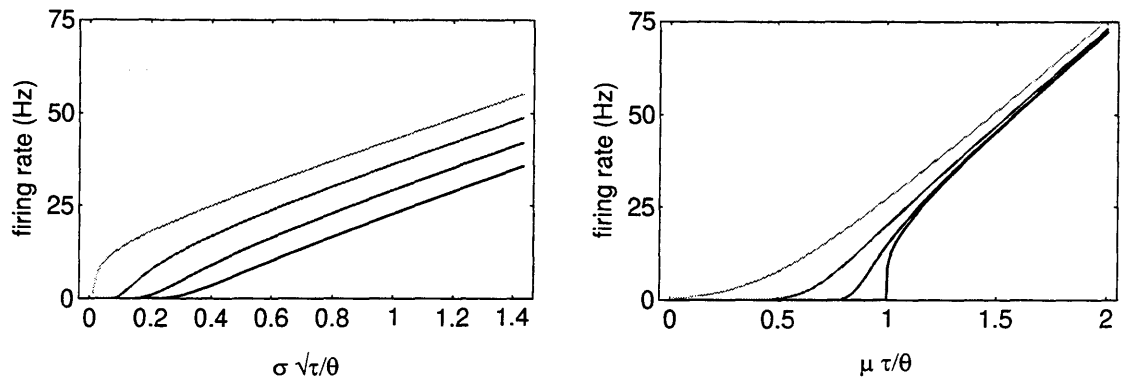


Figure 3.4: In fig. (a),the output of a McCullough-Pitts neuron, represented by a step function between zero and one, is dithered by varying amounts of Gaussian noise, which progressively smoothes out the threshold nonlinearity as the noise variance σ^2 increases. Black represents $\sigma^2 = 0$. Successive gray levels indicate increasing noise variances: $\sigma^2 = 0.1$, $\sigma^2 = 0.3$, and $\sigma^2 = 0.6$. In fig. (b), the firing rate ("f-I" curve) of a leaky integrate-and-fire unit is displayed for different noise variances $\sigma^2\tau = 0$, $\sigma^2\tau = 8\ 10^{-3}\theta^2$, $\sigma^2\tau = 5\ 10^{-2}\theta^2$, and $\sigma^2\tau = 2\ 10^{-1}\theta^2$.

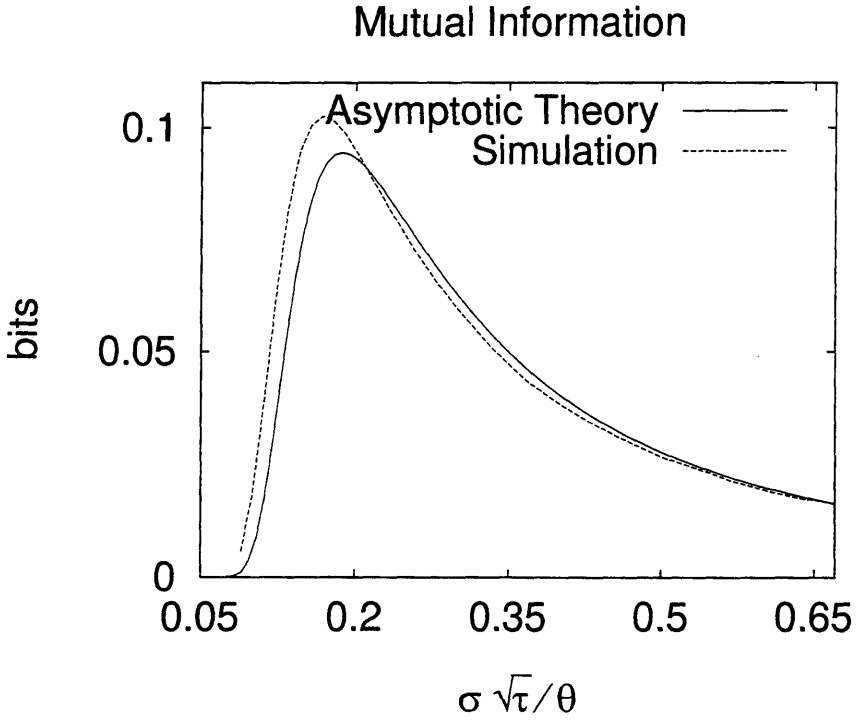


Figure 3.5: The mutual information between the signal and spike count of a leaky integrate-and-fire unit for a background stimulus set at 0.7 θ and a signal stimulus at 0.77 θ, as measured over $T = 200$ msec. The optimal noise level is termed the "stochastic resonance." For each point of the dotted graph, the stochastic differential equation of eq. 3.4 was simulated 100,000 times for a duration of 200 ms. An $O(\Delta t^2)$ algorithm was used with a time step of $\Delta t = 0.1$ ms.

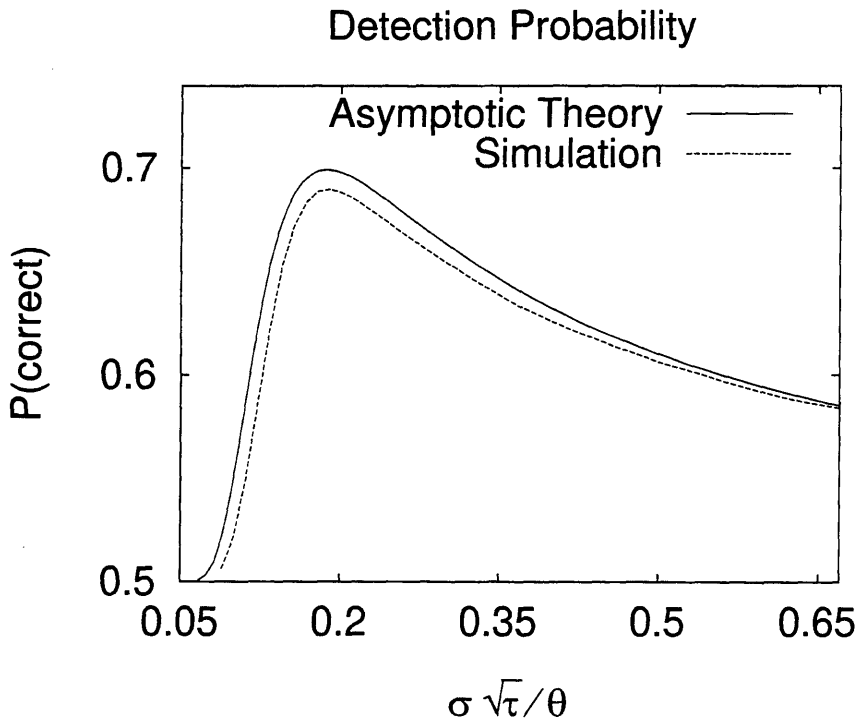


Figure 3.6: The probability of correct detection in the discrimination between the noisy background and the signal also has a stochastic resonance as a function of the standard deviation of the noise. Detection is based on a two-alternative forced choice test, in which the observer decides between a spike count arising from the background input and another spike count from the signal input. We remark that the two-alternative forced choice test implies two "looks" at the information, instead of a single "look," as in the case of the mutual information shown in fig. 3.5. The dashed graph (simulation) plots the result of eq. 3.18, the solid line (theory) the result of eq. 3.19. All parameters are the same as for fig. 3.5. Since the mutual information and detection probability are closely related measures, the optimal noise levels are similar for both. (The peak of the theoretical curves must occur at the identical noise level.)

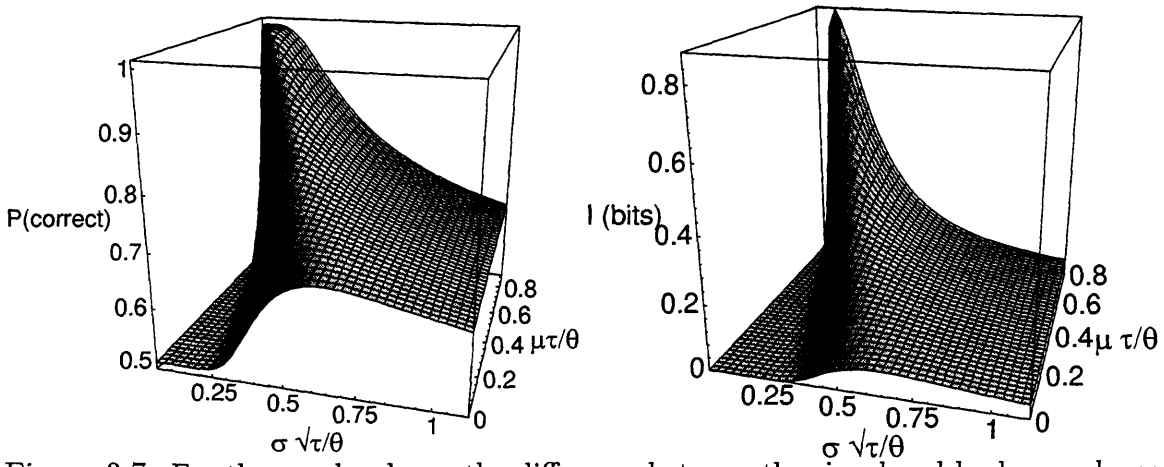


Figure 3.7: For the graphs above, the difference between the signal and background was $\Delta\mu = 0.1\ \theta/\tau$. ($T = 200$ msec). The surfaces were calculated from eq. 3.16 for the mutual information between input and spike count for fig. (a) and from eq. 3.19 for the probability of correct detection in a two-alternative forced-choice test for fig (b). The mutual information is much more sharply peaked than the detection probability because I is directly proportional to the Fisher information lower bound, while $P(d)$ is proportional to its square root.

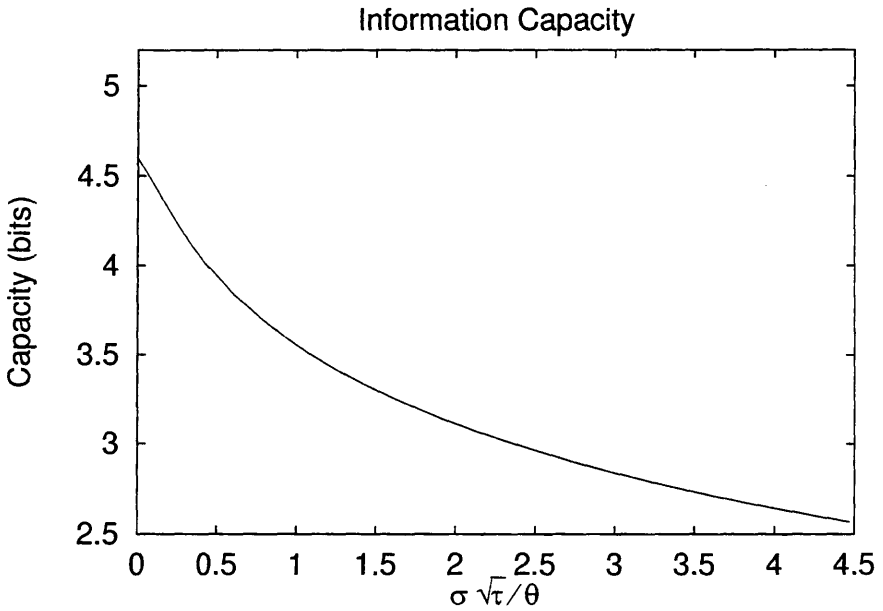


Figure 3.8: The capacity shown here is the maximum amount of information a leaky integrate-and-fire unit with a refractory period of $t_{\text{ref}} = 5$ milliseconds can transmit in $T = 200$ milliseconds. To compute the integral in eq. 3.21 numerically, Wan and Tuckwell's (1982) perturbation results for the mean and standard deviation of the FPT distribution were used for inputs far above threshold $\mu\tau > 8\theta$. For the neuron to transmit at capacity, the input stimulus ensemble (i.e., the possible mean input currents μ) must be optimized for each noise level σ. In the limit of zero noise, the capacity is simply $C = \log_2\left[T/(t_{\text{ref}}\sqrt{e})\right]$. Note that the capacity is a steadily decreasing function of the rms noise amplitude.

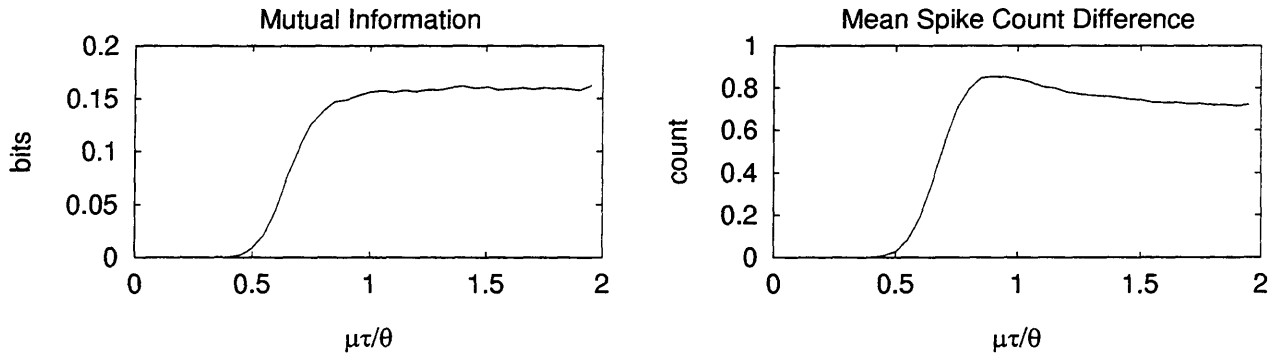


Figure 3.9: Panel (a) displays the mutual information between the input signal and spike count as a function of the bias (or background) input μ for a fixed input noise amplitude. The input signal is $s = \Delta\mu\tau/\theta = 0.07$ and the duration T over which the spike train is observed is 200 msec, the same as in figs. 3.5 and 3.6. The standard deviation of the noise is set to $\sigma\sqrt{\tau}/\theta = 0.16$, an arbitrary choice. For reference, this noise level is optimal for a background input $\mu\tau/\theta = 0.75$. The mutual information measure was computed from the results of simulating the stochastic differential equation 3.4 ($N = 2,000,000$ trials for each data point). At this fixed noise level, the mutual information is a monotonically increasing and saturating function of the bias level. Panel (b) displays the expected difference between the number of spikes in response to the signal and the response to the background. This difference is proportional to the slope $f'(\mu)$ of the firing rate curve. In contrast to the mutual information, the slope is a non-monotonic function of the bias level.

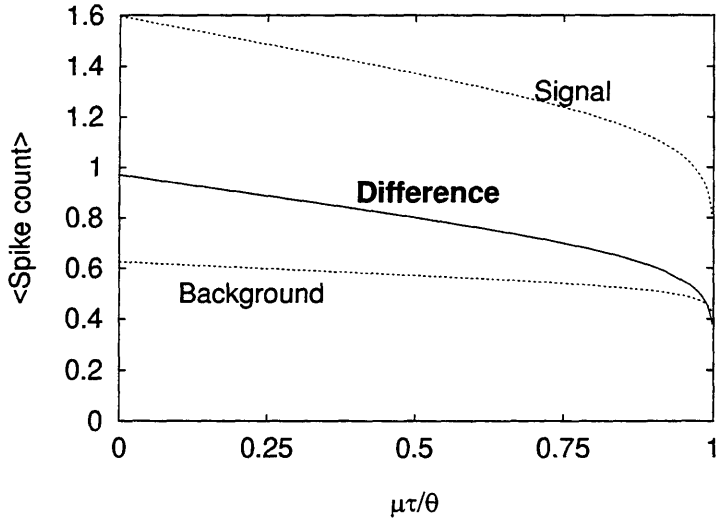


Figure 3.10: For each value of μ, the signal difference $\Delta\mu$ was calculated that would lead to a detection probability $P(d) = 0.75$ provided that the noise is tuned to its optimal value $\sigma^2_{\max} = 2/5(\theta - \mu\tau)^2/\tau$ (eq. 3.10). The expected spike counts T/μ_{FPT} for $T = 200$ msec are computed from eq. A1 for both the signal and the background input. The average difference between the two is displayed as the solid line.

Bibliography

Numbers in brackets are the page numbers where the article is referenced.

Abramowitz, M. and Stegun, I. A. 1970. *Handbook of Mathematical Functions AMS-55.* National Bureau of Standards, 1970. [161]

Adrian, E. D. 1946. *The Physical Background of Perception.* Oxford: Clarendon Press, 1946. [151]

Aggarwal, S. K. and MacKinnon, R. 1996. Contribution of the S4 segment to gating charge in the *Shaker* K^+ channel. *Neuron* 16:1169–1177. [13]

Ahmed, B., Anderson, J. C., Douglas, R. J., Martin, K. A. C. and Nelson, J. C. 1994. Polyneural innervation of spiney stellate neurons in cat visual cortex. *Journal of Comparative Neurology* 341:39–49. [90]

Ahmed, B., Anderson, J. C., Douglas, R. J., Martin, K. A. C. and Nelson, C. 1995. Comparison of current-discharge relationships of pyramidal neurons from the visual-cortex—in vitro and in the anesthetized cat. *J. Physiol. London* 485P:P10–P11. [171]

Almers, W. 1978. Gating currents and charge movements in excitable membranes. *Rev. Physiol. Biochem. Pharmacol.* 82:96–190. [14]

Atick, J. J. 1992. Could information theory provide an ecologocial theory of sensory processing? *Network* 3:213–251. [7, 29]

Atick, J. J. and Redlich, A. N. 1992. What does the retina know about natural scenes. *Neural Comp.* 4:196–210. [130]

Avery, R. B. and Johnston, D. 1996. Multiple channel types contribute to the low-voltage-activated calcium current in hippocampal CA3 pyramidal neurons. *J. Neurosci.* 16:5567–5582. [51]

Baddeley, R. J. and Hancock, P. J. B. 1991. A statistical analysis of natural images matches psychophysically derived orientation tuning curves. *Proc. Roy. Soc. B* 246:219–223. [7]

Barlow, H. B. and Levick, W. R. 1969. Three factors limiting the reliable detection of light by retinal ganglion cells of the cat. *J. Physiol.* 200:1–24. [191]

Barron, A. R. 1986. Entropy and the central limit theorem. *The Annals of Probability* 14:336–342. [153]

Barron, A. R. 1993. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Info. Theory* 39:930–945. [103]

Becker, W. 1989. Metrics. In: *The neurobiology of saccadic eye movements.* Wurtz, R. H., Goldberg, M. E., eds., pp. 13-67. New York: Elsevier, 1989. [97]

Bekkers, J. M. and Stevens, C. F. 1990. Presynaptic mechanism for long-term potentiation in the hippocampus. *Nature* 346:724–729. [191]

Bell, A. J. and Sejnowski, T. J. 1995. An information maximization approach to blind separation and blind deconvolution. *Neural Comp.* 7:1129–1159. [44, 64, 130]

Benzi, R., Sutera, A. and Vulpiani, A. 1981. The mechanism of stochastic resonance.

J. Phys. A 14:L453–L457. [148]

Bialek, W., DeWeese, M., Rieke, F. and Warland, D. 1993. Bits and brains: Information flow in the nervous system. *Physica A* 200:581–593. [185]

Bolshakov, V. Y. and Siegelbaum, S. A. 1995. Regulation of hippocampal transmitter release during development and long-term potentiation. *Science* 269:1730–1734. [191]

Braun, H. A., Wissing, H., Schafer, K. and Hirsch, M. C. 1994. Oscillation and noise determine signal-transduction in shark multimodal sensory cells. *Nature* 367:270–273. [184]

Britten, K. H., Newsome, W. T., Shadlen, M. N., Celebrini, S. and Movshon, J. A. 1996. A relationship between behavioral choice and the visual responses of neurons in macaque MT. *J. Vis. Neurosci.* 13:87–100. [147]

Bulsara, A. R., Lowen, S. B. and Rees, C. D. 1994. Cooperative behavior in the periodically modulated Wiener process - noise-induced complexity in a model neuron. *Phys. Rev. E* 49:4989–5000. [149, 189]

Carandini, M., Heeger, D. J. and Movshon, J. A. 1995. Linearity and gain control in V1 simple cells. In: *Cerebral Cortex vol X. Cortical Models.* Jones, E. G., Ulinski, P. S., eds. New York: Plenum Press, 1995. [173]

Carandini, M. and Heeger, D. J. 1994. Summation and division by neurons in primate visual cortex. *Science* 264:1333–1336. [173]

Carter, G. C. 1985. Coherence and time delay estimation. In: *Les Houches 1985 Traitement du Signal.* Lacoume, J. -L., Durrani, T. S., Stora, R., eds., pp. 515-

572. NY: North-Holland, 1985. [179]

Celebrini, S., Thorpe, S., Trotter, Y. and Imbert, M. 1993. Dynamics of orientation coding in area v1 of the awake primate. *Vis Neurosci* 10:811–825. [147]

Collins, J. J., Chow, C. C. and Imhoff, T. T. 1995a. Stochastic resonance without tuning. *Nature* 376:236–238. [149, 152, 179]

Collins, J. J., Chow, C. C. and Imhoff, T. T. 1995b. Aperiodic stochastic resonance in excitable systems. *Phys. Rev. E* 52:R3321–R3324. [150, 180]

Connor, J. A., Walter, D. and McKown, R. 1977. Neural repetitive firing: Modifications of the hodgkin-huxley axon suggested by experimental results from crustacean axons. *Biophys. J.* 18:81–102. [10, 116]

Connor, J. A. and Stevens, C. F. 1971. Prediction of repetitive firing behaviour from voltage clamp data on an isolated neurone soma. *J. Physiol.* 213:31–53. [171]

Cover, T. M. and Thomas, J. A. 1991. *Elements of Information Theory.* New York: Wiley-Interscience, 1991. [92, 130, 152, 188]

Cox, D. R. 1962. *Renewal Theory.* Great Britain: Methuen, 1962. [156]

De Waard, M., Liu, H., Walker, D., Scott, V. E. S., Gurnett, C. A. and Campbell, K. P. 1997. Direct binding of G-protein $\beta\gamma$ complex to voltage-dependent calcium channels. *Nature* 385:446–450. [26]

DeWeese, M. R. 1995. Optimization principles for the neural code. Ph. D. thesis, Princeton. [185]

Douglas, R., Koch, C., Mahowald, M., Martin, K. and Suarez, H. 1995. Recurrent excitation in neocortical circuits. *Science* 269:981–985. [91]

Douglas, R. J. and Martin, K. A. C. 1991. A functional microcircuit for cat visual cortex. *J. Physiol.* 440:735–769. [90]

Douglass, J. K., Wilkens, L., Pantazelou, E. and Moss, F. 1993. Noise enhancement of information transfer in crayfish mechanoreceptors by stochastic resonance. *Nature* 365:337–340. [149, 184]

Fiser, J. , Subramaniam, S. and Biederman, I. 1996. The effect of changing size and spatial frequency content of gray-scale object images in RSVP identification tasks. *Inv. Opht. & Vis. Sci.* 37:S178. [147]

French, R. J., Prusak-Sochaczewski, E., Zamponi, G. W., Becker, S., Kularatna, A. S. and Horn, R. 1996. Interactions between a pore-blocking peptide and the voltage sensor of the sodium channel: An electrostatic approach to channel geometry. *Neuron* 16:407–413. [13]

Gammaitoni, L. 1995. Stochastic resonance and the dithering effect in threshold physical systems. *Phys. Rev. E* 52:4691–4698. [166, 167]

Gammaitoni, L., Marchesoni, F. and Santucci, S. 1995. Stochastic resonance as a *Bona Fide* resonance. *Phys. Rev. Lett.* 74:1052–1055. [169]

Gingl, Z., Kiss, L. B. and Moss, F. 1995. Non-dynamical stochastic resonance: Theory and experiments with white and various coloured noises. *Il Nuovo Cimento* 17D:795–802. [162, 169]

Goldstein, S. A. N. 1996. A structural vignette common to voltage sensors and conduction pores: Canaliculi. *Neuron* 16:717–722. [14]

Guckenheimer, J. and Holmes, P. 1983. *Nonlinear Oscillations, Dynamical Systems,*

and Bifurcations of Vector Fields. New York: Springer, 1983. [107]

Hebb, D. O. 1949. *The Organization of Behavior.* New York: Wiley, 1949. [58]

Helmchen, F., Imoto, K. and Sakmann, B. 1996. Ca buffering and action potential-evoked Ca signaling in dendrites of pyramidal neurons. *Biophys. J.* 70:1069–1081. [51]

Herlitze, S., Garcia, D. E., Mackie, K., Hille, B., Scheuer, T. and Catterall, W. A. 1996. Modulation of Ca^{2+} channels by G-protein $\beta\gamma$ subunits. *Nature* 280:258–262. [23]

Hessler, N. A., Shirke, A. M. and Malinow, R. 1993. The probability of transmitter release at a mammalian central synapse. *Nature* 366:569–572. [191]

Hille, B. 1992. *Ionic Channels of Excitable Membranes.* Sunderland, MA: Sinauer Associates, 1992. [14, 171]

Hines, M. 1993. NEURON—a program for simulation of nerve equations. In: *Neural Systems: Analysis and Modeling.* Eeckman, F., ed., pp. 127-136. Kluwer Academic Publishers, 1993. [194]

Hodgkin, A. L. and Huxley, A. F. 1952. A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physiol. (Lond.)* 117:500–544. [8, 17, 21, 160]

Hofmann, F., Biel, M. and Flockerzi, V. 1994. Molecular basis for Ca^{2+} channel diversity. *Ann. Rev. Neurosci.* 17:399–418. [12]

Hopfield, J. J. 1994. Neurons, dynamics, and computation. *Physics Today* 47:40–46. [183]

Johnston, D. and Wu, S. M. -S. 1995. *Foundations of Cellular Neurophysiology.* Cambridge, MA: MIT Press, 1995. [19, 54]

Jung, P. 1995. Stochastic resonance and optimal design of threshold detectors. *Phys. Lett. A* 207:93–104. [162, 164]

Kisvarday, Z. F., Martin, K. A. C., Freund, T. F., Magloczky, Z., Whitteridge, D. and Somogyi, P. 1986. Synaptic targets of HRP-filled layer III pyramidal cells in the cat striate cortex. *Exp. Brain Res.* 64:541–552. [90]

Knight, B. W. 1972. Dynamics of encoding in a population of neurons. *J. Gen. Physiol.* 59:734–766. [150]

Koch, C., Bernander, Ö. and Douglas, R. J. 1995. Do neurons have a voltage or a current threshold for action potential initiation? *J. Comput. Neurosci.* 2:63–82. [150]

Lamb, T. D. and Pugh Jr., E. N. 1992a. A quantitative account of the activation steps involved in phototransduction in amphibian photoreceptors. *J. Physiol.* 449:719–758. [24]

Lamb, T. D. and Pugh Jr., E. N. 1992b. G-protein cascades: Gain and kinetics. *Trends in Neurosci.* 15:291–298. [24]

Larsson, H. P., Baker, O. S., Dhillon, D. S. and Isacoff, E. Y. 1996. Transmembrane movement of the *Shaker* K^+ channel S4. *Neuron* 16:387–397. [13]

Laughlin, S. 1981. A simple coding procedure enhances a neuron's information capacity. *Z. Naturforsch.* 36c:910–912. [36]

Laurent, G., Seymour-Laurent, K. J. and Johnson, K. 1993. Dendritic excitability

and a voltage-gated calcium current in locust nonspiking local interneurons. *J. Neurophysiol.* 69:1484–1498. [212]

Lehky, S. R. and Sejnowski, T. J. 1990. Neural model of stereoacuity and depth interpolation based on a distributed represenation of stereo disparity. *J. Neurosci.* 10:2281–2299. [4]

LeMasson, G., Marder, E. and Abbott, L. F. 1993. Activity-dependent regulation of conductances in model neurons. *Science* 259:12915–1917. [43, 64, 215]

Levin, J. and Miller, J. P. 1996. Stochastic resonance in the cricket cercal system. *Nature* 380:165–168. [149, 184]

Linsker, R. 1992. Local synaptic learning rules suffice to maximize mutual information in a linear network. *Neural Comp.* 4:691–702. [44]

Longtin, A. , Bulsara, A. and Moss, F. 1991. Time-interval sequences in bistable systems and the noise-induced transmission of information by sensory neurons. *Phys. Rev. Lett.* 67:656–659. [149]

Longtin, A. ánd Hinzer, K. 1996. Encoding with bursting, subthreshold oscillations, and noise in mammalian cold receptors. *Neural Computation* 8:215–255. [192]

Lukes, T. 1961. Sequences of stochastic pulses. *Proc. Phys. Soc. (London)* 78:153–168. [169]

Lytton, W. W. and Sejnowski, T. J. 1991. Simulations of cortical pyramidal neurons synchronized by inhibitory interneurons. *J. Neurophysiol.* 66:1059–1079. [54]

Ma, J. Y., Li, M., Catterall, W. A. and Scheuer, T. 1994. Modulation of brain Na^+ channels by a g-protein-coupled pathway. *Proc. Natl. Acad. Sci. USA* 91:12351–

12355. [23]

Mainen, Z. F., Joerges, J., Huguenard, J. R. and Sejnowski, T. J. 1995. A model of spike initiation in neocortical pyramidal neurons. *Neuron* 15:1427–1439. [21, 150]

Mason, A. and Larkman, A. 1990. Correlations between morphology and electrophysiology of pyramidal neurons in slices of rat visual cortex. *J. of Neurosci.* 10:1415–1428. [171]

McCullough, W. S. and Pitts, W. 1943. A logical calculus of ideas immanent in nervous activity. *Bull. Math. Biophys.* 5:115–133. [164]

McGuire, B. A., Hornung, J. -P., Gilbert, C. D. and Wiesel, T. N. 1984. Patterns of synaptic input to layer 4 of cat striate cortex. *J. Neurosci.* 4:3021–3033. [90]

McNamara, B. and Wiesenfeld, K. 1989. Theory of stochastic resonance. *Phys. Rev. A* 39:4854–4869. [162]

Miller, J. C. P. 1952. A method for the determination of converging factors, applied to the asymptotic expansions for the parabolic cylinder functions. *Proc. Camb. Phil. Soc.* 48:243–254. [162]

Moss, F., Pierson, D. and O'Gorman, D. 1994. Stochastic resonance: Tutorial and update. *Intern. J. Bifurc. and Chaos* 4:1383–1398. [150]

Newsome, W. T., Britten, K. H. and Movshon, J. A. 1989. Neuronal correlates of a perceptual decision. *Nature* 341:52–54. [148]

Olshausen, B. A. and Field, D. J. 1996. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 381:607–609. [38]

Papoulis, A. 1984. *Probability, Random Variables, and Stochastic Processes.* McGraw-

Hill, 1984. [140]

Paradiso, M. A. 1988. A theory for the use of visual orientation information which exploits the columnar structure of striate cortex. *Biol. Cybern.* 58:35–49. [154]

Pinsker, M. S. 1964. *Information and information stability of random variables and processes.* San Francisco: Holden-Day, 1964. [42]

Radons, G. 1993. On stochastic dynamics of supervised learning. *J. Phys. A: Math. Gen.* 26:3455–3461. [46]

Riani, M. and Simonotto, E. 1994. Stochastic resonance in the perceptual interpretation of ambiguous figures: A neural network model. *Phys. Rev. Lett.* 72:3120–3123. [149]

Ricciardi, L. 1977. *Diffusion Processes and Related Problems.* Berlin: Springer, 1977. [157, 158]

Ripley, B. D. 1996. *Pattern Recognition and Neural Networks.* Cambridge, England: Cambridge University Press, 1996. [55]

Ruderman, D. L. 1995. Statistics of natural images. *Network* 5:517–548. [7, 29]

Schuster, H. G. 1992. Learning by maximizing the information transfer through nonlinear noisy neurons and noise breakdown. *Phys. Rev. A.* 46:2131–2138. [44, 64]

Seoh, S. -A., Sigg, D., Papazian, D. M. and Bezanilla, F. 1996. Voltage-sensing resiudes in the S2 and S4 segments of the *Shaker* K channel. *Neuron* 16:1159–1167. [13]

Seung, H. S. and Sompolinsky, H. 1993. Simple models for reading neuronal popula-

tion codes. *Proc. Natl. Acad. Sci. USA* 90:10749–10753. [1, 4, 154]

Siegert, A. J. F. 1951. On the first passage time probability problem. *Phys. Rev.* 81:617–623. [157]

Softky, W. R. and Koch, C. 1993. The highly irregular firing of cortical cells is inconsistent with temporal integration of random EPSPs. *J. Neurosci.* 13:334–350. [151, 191]

Somers, D. C., Nelson, S. B. and Sur, M. 1995. An emergent model of visual cortical orientation selectivity. *J. Neurosci.* 13. [91]

Stein, R. B. 1967. The information capacity of nerve cells using a frequency code. *Biophys. J.* 7:797–826. [31, 35, 178, 184]

Stemmler, M. 1996. A single spike suffices—the simplest form of stochastic resonance in model neurons. *Network* 7. [32]

Stemmler, M., Usher, M. and Niebur, E. 1995. Lateral interactions in primary visual cortex: A model bridging physiology and psychophysics. *Science* 269:1877–1880. [91, 149, 183]

Strogatz, S. H. 1991. *Nonlinear Dynamics and Chaos.* Reading, MA: Addison-Wesley, 1991. [107, 171]

Subramaniam, S., Biederman, I., Kalocsai, P. and Madigan, S. F. 1995. Accurate identification, but chance recognition memory, of RSVP object pictures. *Inv. Opht. & Vis. Sci.* 36:S377. [147]

Tovée, M. J. 1994. How fast is the speed of thought? *Current Biology* 4:1125–1127. [147]

Tovée, M. J., Rolls, E. T. and Treves, A. 1993. Information encoding and the responses of single neurons in the primate temporal visual-cortex. *J. Neurophysiol.* 70:640–654. [147]

Tsypkin, Y. Z. 1971. *Adaptation and Learning in Automatic Systems.* NY: Academic Press, 1971. [44]

Wan, F. Y. M. and Tuckwell, H. C. 1982. Neuronal firing and input variability. *J. Theor. Neurobiol.* 1:197–218. [198]

White, E. L. 1989. *Cortical Circuits: Synaptic Organization of the Cerebral Cortex — Structure, Function and Theory.* Boston: Birkhauser, 1989. [90]

Wiesenfeld, K., Benz, S. P. and Booi, P. A. A. 1994. Phase-locked oscillator optimization for arrays of josephson junctions. *J. Appl. Phys.* 76:3835–3846. [150]

Wiesenfeld, K. and Moss, F. 1995. Stochastic resonance and the benefits of noise: From ice ages to crayfish and SQUIDs. *Nature* 373:33–36. [148]

Yang, N., George Jr., A. L. and Horn, R. 1996. Molecular basis of charge movement in voltage-gated sodium channels. *Neuron* 16:113–122. [13]

Yang, N. and Horn, R. 1995. Evidence for votlage-dependent S4 movement in sodium channels. *Neuron* 15:213–218. [13]

Yuste, R. and Tank, D. W. 1996. Dendritic integration in mammalian neurons, a century after Cajal. *Neuron* 16:701–706. [52]

Zador, A. and Bulsara, A. R. 1996. Threshold detection of wideband signals: A noise-controlled maximum in the trans-information. *Phys. Rev. E* 54:2185–2188. [184, 185, 189]

Appendix A Appendix

A.1 Simple Model for Ca^{2+}-dependent Learning

Many nerve cells do not spike, but instead release neurotransmitter in a graded-response fashion to stimulation. Such cells include many sensory receptors, such as the hair cells of the auditory-vestibular system and the photoreceptors of the retina, and several classes of inhibitory interneurons in insects (Laurent *et al.*, 1993). Suppose that for these non-spiking cells the amount of neurotransmitter release depends nonlinearly on the concentration of free Ca^{2+} within the cell body. Calcium enters the cell through a set of voltage-dependent Ca^{2+} conductances; for simplicity we assume that these conductances are noninactivating. Furthermore, we will ignore the transient behavior of the Ca^{2+} ion channels, and focus on the steady state.

The inward Ca^{2+} current I_{Ca} at steady state is the sum of currents through the different Ca^{2+} channel types:

$$I_{Ca} = \sum_j \bar{g}_j m_{\infty,j}(V)\,(E_{Ca} - V) = \sum_j \frac{\bar{g}_j\,(E_{Ca} - V)}{1 + \exp(-s_j(V - V_j))}, \tag{A.1}$$

where $\bar{g}_j$, V_j, and s_j are the peak conductance, voltage at half-activation, and slope of the activation curve, respectively, for Ca^{2+} channel type j.

Ca^{2+} entering the cell is either removed through ion pumps or bound to buffer molecules such as calmodulin. As a consequence, the free Ca^{2+} concentration within the cell will depend on the Ca^{2+} inward current as follows:

$$\frac{\partial \left[Ca^{2+}\right]}{\partial t} = -k \left[Ca^{2+}\right] + \gamma I_{Ca}, \tag{A.2}$$

where γ converts current into a concentration, and k is a constant that summarizes the action of ion pumps and buffer molecules. (To be specific, $\gamma = 1/(zF\mathcal{V})$, with z the charge on a Ca^{2+} ion, $\mathcal{V}$ the volume of the cell, and F Faraday's constant converting charge into moles.) The rate of neurotransmitter release will be some nonlinear, saturating function of the concentration of free calcium. For instance, if we treat the neurotransmitter or the vesicles containing the neurotransmitter as simply another buffering mechanism for Ca^{2+}, then one possible description of the rate R of neurotransmitter release is

$$R\Big(\left[Ca^{2+}\right]\Big) = R_0 \quad \frac{\left[Ca^{2+}\right]}{\left[Ca^{2+}\right] + K_d}$$

Here, R_0 is the maximal rate of neurotransmitter release. and K_d measures how well the neurotransmitter binds to calcium. The release rate R is the analogue of the firing rate f for a spiking neuron.

Now, suppose that a range of different stimuli drive the cell, leading to a range of

voltages and a range of responses. What happens in this case if the cell tries to maximize the entropy of the responses R by changing solely the calcium conductance parameters? If the calcium current constitutes only a very small fraction of the current entering the cell, then the distribution of voltages can remain fairly constant, even if the response R changes dramatically as a function of input. This case corresponds to the model of feedforward learning in section 2.6. The resulting learning rules for adapting the various calcium conductances are *non-local*, i.e., changing one Ca^{2+} conductance requires knowledge of the states of all other Ca^{2+} conductances.

The situation becomes no simpler even when the calcium current constitutes a more sizeable fraction of the total current. Because of the presence of voltage-dependent Na^+ and K^+ conductances, calcium conductances will typically not be the only conductances contributing to the nonlinear voltage dependence of the current:

$$I_{\text{tot}}(V) = I_{\text{Ca}^{2+}}(V) + I_{\text{Na}+,\text{K}^+,\text{Cl}^-,\text{other}}(V)$$

Information maximization will require the derivative dI/dV. In section 2.7, we use the trick of grouping all voltage-dependent conductances together. With this approach, we use the physical constraint of charge conservation to our advantage: the current flowing through the set of all voltage-dependent conductances must equal the current discharged through a remaining linear (Ohmic) conductance, simplifying the total derivative dI/dV. If the adaptation mechanism only changes a subset of the voltage-dependent conductances, such as the Ca^{2+} conductances, then the derivative dI/dV remains nontrivial and the resulting learning rules are still non-local.

A plausible adaptation mechanism must, therefore, be able to affect all voltage-dependent conductances. Unlike the learning rule for homeostasis (i.e., maintaining a constant firing rate) by LeMasson *et al.* (1993), a purely Ca^{2+}-dependent mechanism is not feasible for information maximization.

A.2 Notes on the Hodgkin-Huxley Spiking Model

The first-order differential equations for the gating variables m_i and h_i in Hodgkin-Huxley models contain a dependent time constant $\tau_i(V)$:

$$\tau_i(V)\frac{dm_i}{dt} = m_{\infty,i}(V) - m_i$$

If $\tau_i(V) =$ constant, then a Fourier series analysis of the periodically spiking Hodgkin-Huxley system becomes tractable. Another motivation to consider voltage-independent time constants is that, in this case, the transitions between simple ion channel states follows the function $[d/dV m_{\infty,i}(V)]]^\gamma$ with exponent $\gamma = 1$, as described in section 2.2. This is the functional form required for information maximization learning.

Assume that the time constants are voltage-independent. If $V(t)$ is a periodic function, then, after an initial transient, $m(t)$ will also be periodic with the same period, but generally with a phase delay, as can be shown by Fourier decomposition. Suppose that $m(t)$ is slaved to $V(t)$: in other words, changing $m(t)$ will have a small effect on $V(t)$, whereas changing $V(t)$ will have a huge effect on $m(t)$. In other words, assume $V(t)$ is some fixed periodic function. Defining $\omega = 2\pi/T$, we can write down the

Fourier series for $m_\infty[V(t)]$:

$$m_\infty[V(t)] = \sum_{n=-\infty}^{\infty} c_n \exp(i\,\omega n t)$$

where

$$c_n = \frac{1}{T}\int_0^T m_\infty[V(t)] \exp(-i\,\omega n t)\, dt$$

We can substitute this expression into the integral equation for $m(t)$ (with $m(0) = 0$) to get:

$$m(t) = \exp\left(-\frac{t}{\tau}\right) \int_0^t \sum_n c_n \exp(i\omega n t')\ \exp\left(\frac{t'}{\tau}\right) \frac{dt}{\tau}$$

Interchanging summation and integration (assuming uniform convergence)

$$= \sum_n \frac{c_n}{1 + i\,\omega n \tau} \exp(i\omega n t) - \frac{1}{\tau} \exp\left(-\frac{t}{\tau}\right)$$

As $t \to \infty$, the only terms that remain are

$$\begin{aligned} &= \sum_n \frac{c_n}{1 + i\,\omega n \tau} \exp(i\,\omega n t) \\ &= \sum_n \frac{c_n}{\sqrt{1 + (\omega n \tau)^2}} \exp\left[i\,\omega n t - i \arctan(\omega n \tau)\right] \end{aligned} \tag{A.3}$$

By taking the limit of $t \to \infty$, we can drop the assumption of slaving. Once ω is

known, this is the only self-consistent solution. Equation A.3 shows that the Fourier components of $m(t)$ lag behind those of $m_\infty(t)$ with the frequency-dependent phase delay of $\arctan(\omega n\tau)$, and are dampened by an amplitude factor $\left[1+(\omega n\tau)^2\right]^{-\frac{1}{2}}$. We can use this result to compute other averages, such as $\langle m_i(V)^2\rangle$:

$$\langle m_i(V)^2\rangle = \sum_{n=-\infty}^{\infty} \frac{c_{-n}\, c_n}{1+(\omega n\tau)^2}$$

Using the fact that

$$\sum_{n=-\infty}^{\infty} c_{-n}\, c_n = \frac{1}{T}\int_0^T \left\{m_{\infty,i}[V(t)]\right\}^2 dt \tag{A.4}$$

$$c_0\, c_0 = \left\{\frac{1}{T}\int_0^T m_{\infty,i}[V(t)]\, dt\right\}^2, \tag{A.5}$$

we get the following relationships in the limit

$$\lim_{\tau\to 0}\langle m_i(V)^2\rangle = \langle m_{\infty,i}(V)^2\rangle$$

$$\lim_{\tau\to\infty}\langle m_i(V)^2\rangle = \langle m_{\infty,i}(V)\rangle^2$$

Since $\langle m_{\infty,i}(V)^2\rangle \geq \langle m_{\infty,i}(V)\rangle^2$, if we replace

$$s_i\left(\langle m_{\infty,i}(V)\rangle - \langle m_{\infty,i}(V)^2\rangle\right)$$

by

$$s_i\left(\langle m_i(V)\rangle - \langle m_i(V)^2\rangle\right),$$

then we will overestimate the derivative, unless $V(t)$ is constant or τ is zero. Since any biophysical implementation of such a learning mechanism must be based on the fluctuations of individual ion channels, we will compare the results above to the net rate of forward/backward transitions of ion channels. Averaged over one cycle, the rate of backward transitions must be *equal* to the rate of forward transitions, else the system is not in quasi-steady-state. The rate of backward transitions is:

$$\begin{aligned} f_C &= \frac{1}{T}\int_0^T \beta_i[V(t)]\, m_i[V(t)]\, dt \\ &= \frac{1}{T}\int_0^T \beta_i[V(t)]\, m_{\infty,i}[V(t)]\, dt - \frac{1}{T}\int_0^T \beta_i[V(t)]\, \tau[V(t)]\, \frac{dm}{dt}\, dt, \end{aligned}$$

whereas the forward rate of transitions is:

$$\begin{aligned} f_O &= \frac{1}{T}\int_0^T \alpha_i[V(t)]\Big\{1 - m_i[V(t)]\Big\}\, dt \\ &= \frac{1}{T}\int_0^T \alpha_i[V(t)]\Big\{1 - m_{\infty,i}[V(t)]\Big\}\, dt + \frac{1}{T}\int_0^T \alpha_i[V(t)]\, \tau[V(t)]\, \frac{dm}{dt}\, dt, \end{aligned}$$

Of course,

$$\frac{1}{T}\int_0^T \beta_i[V(t)]\, m_{\infty,i}[V(t)]\, dt = \frac{1}{T}\int_0^T \alpha_i[V(t)]\Big\{1 - m_{\infty,i}[V(t)]\Big\}\, dt$$

by definition of $m_{\infty,i}$. Thus, to show that the average rates are equal, we subtract

$$f_O - f_C = \frac{1}{T}\int_0^T \Big\{\alpha_i[V(t)] + \beta_i[V(t)]\Big\}\tau[V(t)]\, \frac{dm}{dt}\, dt.$$

Using $\alpha_i[V(t)] + \beta_i[V(t)] = \tau[V(t)]^{-1}$

$$= \frac{1}{T}\int_0^T \frac{dm}{dt}\,dt$$

$$= 0,$$

and, thus, $f_O = f_C$. For $\tau =$ constant, we can rewrite the average rate as

$$f_O, f_C = \langle m_{\infty,i}(V)\rangle - \langle m_{\infty,i}(V)^2\rangle + \frac{1}{T}\int_0^T m_{\infty,i}[V(t)]\,\tau\frac{dm}{dt}\,dt$$

In terms of the Fourier series for $m_{\infty,i}[V(t)]$, and the corresponding Fourier series we found for $m_i[V(t)]$, we have

$$\frac{1}{T}\int_0^T m_{\infty,i}[V(t)]\,\tau\frac{dm}{dt}\,dt = \sum_{n=-\infty}^{\infty} i\,c_{-n}\,c_n\frac{\omega n\tau}{1+i\,\omega n\tau}$$

$$= \sum_n c_{-n}\,c_n\frac{(\omega n\tau)^2}{1+(\omega n\tau)^2} + i\sum_n c_{-n}\,c_n\frac{\omega n\tau}{1+(\omega n\tau)^2}$$

If we take the limit $\tau \to \infty$, we have

$$\frac{1}{T}\int_0^T m_{\infty,i}[V(t)]\,\tau\frac{dm}{dt}\,dt = \sum_{n\neq 0} c_{-n}c_n$$

From eqs. A.4 and A.5, we can write

$$\frac{1}{T}\int_0^T m_{\infty,i}[V(t)]\,\tau\frac{dm}{dt}\,dt = \langle m_{\infty,i}[V(t)]^2\rangle - \langle m_{\infty,i}[V(t)]\rangle^2$$

Thus, in this limit,

$$f_O, f_C = \langle m_{\infty,i}(V) \rangle - \langle m_{\infty,i}(V) \rangle^2$$

In the opposite limit $\tau \rightarrow 0$, the integral $\int_0^T m_{\infty,i}[V(t)]\, \tau \frac{dm}{dt}\, dt$ tends to zero, so

$$f_O, f_C = \langle m_{\infty,i}(V) \rangle - \langle m_{\infty,i}(V)^2 \rangle.$$

In either limit, in fact,

$$f_O, f_C = \langle m_i(V) \rangle - \langle m_i(V)^2 \rangle.$$

A.3 Mean and Variance of the First Passage Time Distribution

The first moment and variance of the first passage time distribution are given by:

$$\mu_{\mathrm{FPT}} = \frac{\tau}{2} \left\{ \sum_{n=1}^{\infty} \frac{x^{2n}}{n \left(\frac{1}{2}\right)_n} - 2\sqrt{\pi} x \sum_{n=0}^{\infty} \frac{\left(\frac{1}{2}\right)_n x^{2n}}{\left(\frac{3}{2}\right)_n n!} \right\} \Bigg|_{(\mu\tau - V_0)/\sqrt{\sigma^2 \tau}}^{(\mu\tau - \theta)/\sqrt{\sigma^2 \tau}} \tag{A1}$$

$$\sigma^2_{\mathrm{FPT}} = \frac{\tau^2}{4}\Bigg\{ 2(\gamma + 2\ln(2)) \sum_{n=1}^{\infty} \frac{x^{2n}}{n\,(\frac{1}{2})_n} + 2\sum_{n=2}^{\infty}\left(\sum_{m=1}^{n-1}\frac{1}{m}\right)\frac{x^{2n}}{n\,(\frac{1}{2})_n} - 2\sqrt{\pi}x\Big[2\gamma\sum_{n=1}^{\infty}\frac{(\frac{1}{2})_n\,x^{2n}}{(\frac{3}{2})_n\,n!} - 2\,e^{x^2}\sum_{n=1}^{\infty}\left(\sum_{m=1}^{n}\frac{1}{m}\right)\frac{(-x)^{2n}}{(\frac{3}{2})_n}\Big] - \left[\sum_{n=1}^{\infty}\frac{x^{2n}}{n\,(\frac{1}{2})_n} - 2\sqrt{\pi}x\sum_{n=0}^{\infty}\frac{(\frac{1}{2})_n\,x^{2n}}{(\frac{3}{2})_n\,n!}\right]^2\Bigg\}\Bigg|_{(\mu\tau-\theta)/\sqrt{\sigma^2\tau}}^{(\mu\tau-V_0)/\sqrt{\sigma^2\tau}} \tag{A2}$$

The formulae use the Pochhammer symbol notation:

$$(a)_n = a\,(a+1)\,(a+2)\ldots(a+n-1) \quad n \geq 1$$

$$(a)_0 = 1.$$

The asymptotic expansion with converging factor for μ_{FPT} as $\sigma \to 0$ is

$$\mu_{\mathrm{FPT}} = \tau\Bigg\{\sqrt{\pi}\,\frac{\exp(x^2)}{x}\left[\sum_{n=0}^{\lfloor x^2\rfloor}\frac{(1/2)_n}{(x^2)^n} + C_{\lceil x^2\rceil}\,\frac{(1/2)_{\lceil x^2\rceil}}{(x^2)^{\lceil x^2\rceil}}\right]\Bigg\}\Bigg|_{(\mu\tau-\theta)/\sqrt{\sigma^2\tau}}^{(\mu\tau-\theta)/\sqrt{\sigma^2\tau}} + \frac{\tau}{2}\Bigg\{-\ln(x^2) + \sum_{m=0}^{\lfloor x^2\rfloor}\frac{(1/2)_m}{(-x^2)^m}\Bigg\}\Bigg|_{(\mu\tau-V_0)/\sqrt{\sigma^2\tau}}^{(\mu\tau-\theta)/\sqrt{\sigma^2\tau}} \tag{A3a}$$

where $\lfloor x \rfloor$ is the closest integer smaller or equal to x and $\lceil x \rceil$ is the closest integer larger or equal to x. $.C_{\lceil x^2\rceil}$, the converging factor, is a series in $1/x^2$. Defining

$y = x^2 - \lceil x^2 \rceil$, the converging factor is

$$C_{\lceil x^2 \rceil} = (\frac{1}{6} + y) - \frac{1}{x^2}(\frac{2}{3}y^3 - \frac{1}{3}y^2 - \frac{1}{6}y + \frac{13}{540}) + O(\frac{1}{x^4}). \qquad \text{(A3b)}$$

A similar expression, albeit more complicated, holds for the asymptotic expansion of σ^2_{FPT}.

A.4 Note on the Effect of Firing Rate Saturation

Eq. 3.4 governing the behavior of the leaky integrate-and-fire unit allows the firing rate to grow without bound as either $\sigma \to \infty$ or $\mu \to \infty$. Since signal detection occurs in the regime of low firing rates far removed from saturation, we are justified in neglecting the physical limits on very high firing rates.

If the firing rate saturates very strongly and quickly as a function of the input, however, the leaky integrate-and-fire unit effectively converts into a bistable device. For bistable devices, a "resonance" in the transmitted information occurs because all inputs are mapped onto the same output in both the low and high noise limits. The SR peak in the lower bound to the Fisher information J_{LB}, however, exists already in the absence of saturation.

It is a simple matter to explore the effect of firing rate saturation on the SR peak by introducing a "hard refractory period" after each spike. In this model, the voltage V is held at the reset voltage V_0 for a refractory period t_{ref} after each spike. The variance σ^2_{FPT} of the first passage time remains unchanged, but the average interspike interval

μ_{FPT} is effectively lengthened by t_{ref}. In terms of the original Fisher information J^0_{LB}, the effect of saturation leads to

$$J_{\mathrm{LB}}(\mu) = \left(\frac{\mu_{\mathrm{FPT}}}{t_{\mathrm{ref}} + \mu_{\mathrm{FPT}}}\right) J^0_{\mathrm{LB}}(\mu).$$

As a consequence, as $\sigma \to \infty$,

$$J_{\mathrm{LB}} \propto \frac{1}{\sigma^3},$$

which implies that the asymptotic tail of the SR peak decays faster than the $1/\sigma^2$ behavior expected in the absence of saturation.

A.5 The Poisson Estimate of the Fisher Information

If the spike train of a neuron is Poisson, the Fisher information is predicted by eq. 3.14. Even if the spike train is not Poisson, a SR peak in the Fisher information for the rate-equivalent Poisson process will be mirrored by a peak in the original process. We sketch a rough argument why.

Assume that both $\mu_{\mathrm{FPT}}(\mu, \sigma)$ and $\sigma^2_{\mathrm{FPT}}(\mu, \sigma)$ for $\sigma > 0$ are twice differentiable positive definite functions that are concave in both μ and σ. The concavity requirement ensures that the $\mu^2_{\mathrm{FPT}}(\mu, \sigma)$ and $\sigma^2_{\mathrm{FPT}}(\mu, \sigma)$ intersect at most twice as a function of

σ for given μ. The firing rate $f(\mu,\sigma)$ describes a threshold system if there exists a threshold input such that, for $\mu < \mu_\theta$, $f(\mu) = 0$ in the absence of noise ($\sigma = 0$).

For $\mu < \mu_\theta$, insist that

$$\sigma^2_{\mathrm{FPT}}(\mu,\sigma) \underset{\sigma\to 0}{\longrightarrow} \mu^2_{\mathrm{FPT}}(\mu,\sigma)$$

$$\sigma^2_{\mathrm{FPT}}(\mu,\sigma) \underset{\sigma\to \infty}{\longrightarrow} \chi\, \mu^\rho_{\mathrm{FPT}}(\mu,\sigma)$$

where $\rho < 2$, and χ is a positive real number.

Now if $f(\mu,\sigma)$ can be bounded by $M\left|\mu+\sigma\right|^{2-\varepsilon}$, where M and ε are real, positive numbers ($\varepsilon < 2$), then $J_{\mathrm{Poisson}}(\mu)$ has a maximum as a function of σ for $\mu < \mu_\theta$. In particular, a piecewise linear $f(\mu,\sigma)$ can be bounded in such a fashion.

There exists a σ_0 for each μ, such that $\sigma^2_{\mathrm{FPT}}(\mu,\sigma_0) = \mu^2_{\mathrm{FPT}}(\mu,\sigma_0)$. We can further guarantee that $\sigma^2_{\mathrm{FPT}}(\mu,\sigma) < \mu^2_{\mathrm{FPT}}(\mu,\sigma)$ for $\sigma < \sigma_0$ by invoking the concavity of these functions and the asymptotic limit of $\sigma \to 0$. We conclude then that

$$J_{\mathrm{LB}}(\mu) < J_{\mathrm{Poisson}}(\mu) \text{ for } \sigma > \sigma_0,$$

$$J_{\mathrm{LB}}(\mu) > J_{\mathrm{Poisson}}(\mu) \text{ for } \sigma < \sigma_0.$$

Since $J_{\mathrm{LB}}(\mu)$ approaches $J_{\mathrm{Poisson}}(\mu)$ as $\sigma \to 0$, $J_{\mathrm{LB}}(\mu)$ will possess a maximum if $J_{\mathrm{Poisson}}(\mu)$ does.