Neural coding and the statistical modeling of neuronal responses

by

Jonathan Pillow

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Center for Neural Science
New York University
May 2005

Eero P. Simoncelli

ACKNOWLEDGMENTS

I want to thank my advisor, Eero Simoncelli, for much inspiration, guidance, and encouragement, and for the boundless energy and enthusiasm he has lavished on the scientific projects we have undertaken together. Thank you also to Liam Paninski, who inspired many of the ideas contained in this thesis and collaborated with me closely on their development. I also owe a huge debt of gratitude to E.J. Chichilnisky, an invaluable mentor and collaborator who generously provided data, advice, and philosophical insight into the nature of losing.

I also want to thank Nava Rubin, who supervised my first research experience at NYU and helped to inspire my interests in vision. I owe many thanks to Bob Shapley, Dan Tranchina, Alex Reyes, and Tony Movshon, who advised me during my time at NYU. Thanks to Sam Feldman and Lynne Kiorpes, for their tireless support. Thank you also to the members of the Simoncelli lab: Nicole Rust, Odelia Schwartz, Alan Stocker, Patrick Hoyer, David Hammond, Zhou Wang, Cynthia Rudin, Rob Dotson, Nicolas Bonnier, and Jose Acosta, and to Valerie Uzzell of the Chichilnisky lab for sharing her data. I am grateful to Stu Greenstein, John Rinzel, Larry Maloney, and Paul Glimcher for much additional encouragement and advice, and to the outside members of my thesis committee, Charlie Peskin and Larry Abbott.

Thank you to Rich Zemel, for introducing me to the field of theoretical neuroscience and for inspiring my first interests in neural coding. I owe a debt of thanks to many outstanding teachers who have inspired and encouraged me, especially John Jensen, David Lomen, and John Stollar. Thank you also to the Flinn Foundation.

Thank you to Joe Holmgren, Jason Salinas, Mike Marziani, Paul Warren and Brady Butterfield, for devoted friendship and loving mockery, and to Erin Cox, for providing a beacon of hope and inspiration. Thank you to my brother Thomas, who has cheered me all along, and to my parents, for their unconditional love and unwavering support, in this as in all other things.

Support for my studies was provided by the National Science Foundation, the NCAA Graduate Fellowship, the Dean's Dissertation Fellowship, and the Howard Hughes Medical Institute.

ABSTRACT

One of the central problems in theoretical neuroscience is that of determining the functional relationship between stimuli (e.g. visual scenes) and the spike responses of neurons. This problem, sometimes referred to as the "neural coding problem", involves the attempt to understand the spiking activity of neurons as a *code* for the sensory stimuli which elicit them. A full understanding of the neural code will mean that we can predict the responses of neurons to arbitrary novel stimuli, and can decode the information contained in spike trains to reveal the stimuli that gave rise to them. Models of the neural code provide insight into the computations performed by the brain, with potential applications for neural prosthetics and the development of novel solutions to engineering problems which the brain solves.

In this thesis, we pursue the development and application of novel statistical tools for modeling the neural code. We focus specifically on models which provide precise quantitative descriptions of the computational behavior of neurons in the early visual pathway.

In the first portion of the thesis, we focus on spike-triggered covariance analysis, a technique for finding a multi-dimensional stimulus subspace in which a neuron computes its response. We apply this technique to data collected in macaque retinal ganglion cells (data provided by EJ Chichilnisky),

and show that it can be used to constrain a model composed of nonlinear spatial subunits and a suppressive temporal feedback signal.

In the second portion of the thesis, we develop a method for fitting neural data using a generalized integrate-and-fire model, which is more powerful, flexible and biophysically realistic than those traditionally used for neural characterization. We show that this model accounts for the detailed timing precision and variability of retinal ganglion cell spike responses, and is useful for decoding spike trains of macaque retinal ganglion cells.

TABLE OF CONTENTS

LIST OF FIGURES

Introduction

The brain is one of the most marvelously complex structures in the known universe, and arguably still one of the most poorly understood. Although the brain was first posited as the primary locus of sensation and cognition nearly two and a half millennia ago[1], it is only within the past century[2] that any understanding of the brain's activity, and its relation to human and animal behavior, has begun to emerge.

It is now universally accepted that the rich panoply of sensory, cognitive and motor capabilities exhibited by humans and animals arises critically from the spiking activity of neurons in the varied structures of nervous system, which includes the brain. One of the overarching goals of the nascent discipline of neuroscience is to provide a comprehensive account of the brain's neuronal activity, and the relationship of this activity to behavior. *Theoretical neuroscience* is a sub-discipline which seeks to provide

[1] Alcmaeon, writing around 450 BCE, performed anatomic dissections of the human brain and optic nerves, hypothesizing that the latter were "light-bearing paths" to the brain (Gross, 1998)

[2] (Adrian, 1926; Hodgkin & Huxley, 1952)

precise descriptions of brain activity and its relationship to behavior using the formal language of mathematics. Inarguably, it is only when we have obtained a mathematical description of the brain's activity that we will be able to say that we understand *how the brain works*. Such a description will mean that we can simulate the brain's activity using a computational model—that we can predict its response to a novel stimulus, its ability to recall a piece of knowledge, or its capacity to learn a novel behavior. The value of such models is not only that we will understand of the brain's operation; understanding the brain's ability to perform such complicated tasks as recognizing a face, navigating in a cluttered environment, or making decisions using multiple sources of information, means that we will understand the computational algorithms used by the brain to perform such tasks, which may suggest powerful engineering solutions to such currently intractable problems.

The neural coding problem

One of the central problems in theoretical neuroscience is that of determining the functional relationship between stimuli (e.g. a visual scene) and the spike responses of neurons. This problem is sometimes referred to as the "neural coding problem". The brain represents information about the outside world using neural spike patterns, and it is by virtue of the brain's ability to transmit and manipulate these patterns that we are able to perceive and interact with our surroundings. We can therefore view the relationship

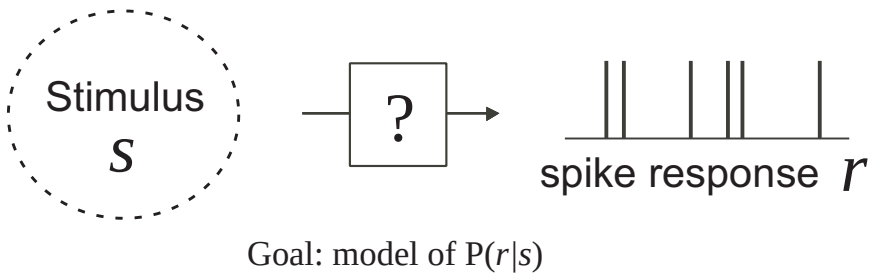

Figure 0.1: The "Neural Coding Problem". Solution consists in finding a model of the probabilistic, functional transformation from stimuli s to spike responses r, or a description of $P(r|s)$.

between neuronal firing patterns and events in the external world as instantiating a *coding relationship*. If we knew how to read the code, then by observing a particular pattern of neural firing, we would know some piece of information about the state of the world. Conversely, if we were given a particular set of events in the world, we would be able to predict the pattern of neuronal activity taking place in the brain.

Generally, we can formalize the neural coding problem as the search for a description of the functional relationship between stimuli s and neuronal spike responses r. Figure 0.1 shows a graphic depiction of this problem. Because the relationship between stimulus and response is probabilistic—repeated presentations of the same stimulus elicit responses which differ unpredictably from trial to trial—we describe the problem as a search for $P(r|s)$, the conditional probability of the response given the stimulus. Learning this distribution from empirical samples (r_i, s_i), i.e. data collected in real neurons, is fundamentally a problem in statistical estimation. The

main focus of this thesis will be the derivation of a set of statistical tools for solving this estimation problem and an illustration of their performance with an application to real data from primate retina, collected in the lab of E. J. Chichilnisky of the Salk Institute.

Before continuing, we should briefly discuss the particular ambitions of our approach. First, we would in principle like know $P(r|s)$ for any possible stimulus s. In the case of visual neurons, this means that we would like to be able to compute the probability of the response for any possible spatio-temporal-chromatic pattern of light impinging on the retina. This is in contrast to approaches which model responses only to some restricted or parametrically varying class of stimuli, such as that provided by a tuning curve. Instead, we consider an approach generally known as *white noise analysis*, a set of techniques for fitting quantitative models of neural function using the measured responses to a set of rapidly varying, stochastic stimuli spanning a large region of a neuron's input space. There has been a recent resurgence of interest in these techniques, partly due to the development of powerful computers capable of real-time random stimulus generation and computationally intensive statistical analysis.

Second, we should note that we cannot use the standard tools for empirical density estimation to estimate $P(r|s)$, as the dimensionality of both the stimulus space and the response space are enormous[3]. It is to all in-

[3]Dimensionality of the stimulus is $D = 3$ *color channels* $\times$ m *spatial pixels* $\times$ n *temporal frames*, for a particular discretization of space and time, with m and n each

tents and purposes impossible to collect enough data to directly fit such a high-dimensional probability density. Therefore, we simplify the problem by adopting a strategy which is to represent $P(r|s)$ using a model. The essential features of such a model are that it be simple enough to estimate using extracellular neural data, and that it be flexible enough to capture the essential features of $P(r|s)$. Additionally, the model cannot be a deterministic, but should have a probabilistic formulation that facilitates computing the likelihood of the varying neural response. Finally, it is desirable to have a model which can be interpreted mechanistically, or which can be used to make inferences about the biophysical substrate underlying the neuron's functional properties.

Noted that a fundamental tension persists between these modeling desiderata: it is easy to find models (e.g. Hodgkin-Huxley style) with a clear underlying biophysical interpretation and capable of complex dynamical behaviors, but which are not easily estimated from extracellular data or do not easily admit of probabilistic computations. Similarly, we define models (e.g. linear-rectifying-Poisson) for which it estimation and probabilistic computations are straightforward, but which are so simple as to have limited representational capacity or limited biological plausibility. In this thesis, we examine the estimation and application of two different models which stake out slightly different positions among the tradeoffs imposed

potentially on the order of hundreds to thousands. Dimensionality of the response space depends on the discretization of time and the length of the time window considered.

by these competing desiderata.

Relation to other approaches

Before we proceed, it is useful to provide a brief overview of several alternative approaches within theoretical neuroscience, in order to situate the current work with respect to the contributions from other frameworks for studying the neural code. As outlined above, the goal of this thesis is to explore statistical tools for estimating the encoding distribution $P(r|s)$ for individual neurons, to obtain detailed, quantitative models of the neural response to arbitrary stimuli. This stands in contrast to a more strictly theoretical approach, whose goal has been to understand the evolutionary "design principles" underlying neural coding. This framework seeks to gain a principled statistical understanding of *why* neurons code things the way the do (as opposed to a descriptive explanation of *what* they code). The most important idea to emerge from this perspective so far is Barlowe's "efficient coding hypothesis", which holds that the primary goal of sensory processing is to achieve representations that are *efficient*. Intuitively, this means that the variability in r should be used to optimally preserve information about the variability in s, which we can write formally as maximizing the mutual information between r and s.[4]. One way to derive the optimal encoding $P(r|s)$ for sensory processing (where r is now the response

[4]This is mathematically equivalent to maximizing the entropy of $P(r|s)$ minus the entropy of $P(r)$

of many neurons) is therefore to analyze the statistical properties of natural images, $P(s)$. This distribution turns out to be highly redundant, and it is possible to find transformations which greatly reduce this redundancy by maximizing the entropy of $P(r|s)$[5]. Other approaches in this same vein have examined the empirical distribution of $P(r|s)$ over a wide range of samples from $P(s)$ and have concluded that it is maximally entropic under a particular constraint on $P(r)$. So far, these approaches have generally been restricted to considering a deterministic mapping of the stimulus and a scalar ("rate") representation of the response. More work will be required to relate "*what*" models of the neural code to "*why*" which account the stochasticity and temporal dynamics intrinsic to real neural responses.

Another theoretical framework for studying the neural code involves efforts to model simply the distribution of spike responses, or $P(r)$. Such approaches have generally sought to analyze the statistical properties of

[5]One of the most successful examples of this approach is independent component analysis (ICA), used to provide a theoretical explanation of the shapes of V1 receptive fields. In this approach, assume a deterministic linear mapping of a (vector) stimulus $\vec{s}$ to a (scalar) response r_i for each neuron via a linear kernel $\vec{k}_i$, corresponding loosely to a V1 receptive field; the response of the ith neuron to stimulus $\vec{s}$ is thus $r_i = \vec{k}_i \cdot \vec{s}$. ICA provides an algorithm for finding a set of "receptive fields" $\{\vec{k}_1, \ldots \vec{k}_n\}$ which maximize the entropy of $P(r_1, \ldots, r_n|s)$ averaged over $P(s)$, while holding the "noise entropy" $P(r_1, \ldots r_n)$ fixed. In contrast, the goal considered in thesis would be to estimate $\vec{k}_i$ for each neuron, based on a set of its measured responses to random sample from $P(s)$, which describes *what* the neuron is encoding but not *why*

spike trains (e.g. "whether they can be described as a Poisson process") and to discuss models which can or cannot give rise to the renewal statistics or joint statistics observed in various brain areas. A related but generally non-statistical approach to this same set of issues arises from the dynamical systems perspective, which seeks to find differential equations which capture the dynamical behavior and phase transitions exhibited by neurons. Often these models have the advantage of posing a simple and direct interpretation in terms of underlying biophysical mechanisms. Clearly, the insights provided by these techniques can inform our selection of a statistically accurate and dynamically realistic model for $P(r|s)$.

A third approach which bears importantly on problem of understanding the neural code is the information theoretic approach, which primarily involves finding tools for estimating $I(r, s)$, the mutual information (MI) between stimuli and responses. As noted above, mutual information is a functional of the joint probability distribution $P(r, s)$; it attaches a single number quantifying the meaningful shared variability between stimuli and responses. Although estimating MI from samples presents many difficulties in itself, it is much easier to obtain reliable estimates of this quantity than to estimate the full probability distribution $P(r, s)$. For this reason, MI can provide an important tool for evaluating the success of efforts to model $P(r|s)$. By comparing the MI between s and r with the MI between s and the simulated output of a model, for example, we can determine whether a model captures the same amount of information about the stimulus as

contained in the neural responses themselves.

Finally, it is important to acknowledge the enormous contributions from classical neuroscience and non-statistical modeling to our understanding of neural coding in early visual areas. Retinal ganglion cell responses, which form the domain of application for all the statistical techniques described in this thesis, have already been thoroughly studied using both classical and white-noise stimuli, and have been modeled extensively. Although the goal of the work presented here—to obtain complete, quantitative, functional models of individual retinal ganglion cell response properties that capture detailed differences between neurons and facilitate probabilistic analysis of their responses—differs meaningfully from that of earlier physiological and theoretical studies, there is nevertheless a great debt to the knowledge accumulated through classical studies of retinal stimulus selectivity and classical modeling of linear and nonlinear retinal response properties.

Part I: Spike-triggered covariance analysis of Retinal Responses

The thesis can be divided roughly into three parts. The first covers spike-triggered covariance (STC), an idea for finding a reduced-dimensional description of $P(r|s)$ using the variance of the spike-triggered stimulus ensemble (i.e. the subset of $P(r, s)$ for which $r = 1$). This idea was first proposed by Brenner et al (Brenner, Bialek, & Steveninck, 2000) and further developed by (Schwartz, Chichilnisky, & Simoncelli, 2002) as a tool for constraining a multi-dimensional nonlinear model of an individual neuron's

response.

In Chapter 1, we derive spike-triggered covariance analysis from first principles and develop a space-time separable analysis which amounts to performing STC in a subspace which preserves only spatial information or temporal information about the stimulus.

This work has been conducted in collaboration with Eero Simoncelli and E. J. Chichilnisky of the Salk Institute, and was presented as a poster at the Annual Society for Neuroscience Meeting 2003 and the Computation and Systems Neuroscience meeting in March 2004.

Part II: Neural characterization with a generalized integrate-and-fire model

This section comprises Chapters 2 and 3, which describe statistical procedures for fitting a generalized integrate-and-fire neuron to extracellular neural data.

The work in Chapter 2 was conducted in collaboration with Eero Simoncelli, and was presented as a talk at the Computational Neuroscience Meeting (CNS) 2002, and published in the journal *Neurocomputing* (Pillow & Simoncelli, 2003).

The work in Chapter 3 was conducted jointly with Liam Paninski and Eero Simoncelli. It expands upon the problem addressed in Chapter 2, with the addition of an explicitly probabilistic model and a more general form of recurrent dependence on the spike train history.

This work was presented orally at both the Neural Information Processing Society (NIPS) Meeting 2002 and the Computation and Neural Systems (CoSyNe) Meeting 2003. It was published in the NIPS proceedings 2003 (Pillow, Paninski, & Simoncelli, 2004), and an expanded version of the paper was recently published in Neural Computation (Paninski, Pillow, & Simoncelli, 2004).

Part III: Application of the Generalized Integrate-and-Fire Model to Retinal Responses

This work presented in Chapter 4 entails a comprehensive application of the model described in Chapter 3 to the spike responses of retinal ganglion cells (RGCs) to temporal white noise stimuli, using data collected by Valerie Uzzell in the lab of E. J. Chichilnisky. We show that the model gives a simple, thorough and intuitive account of "neural precision", a much-discussed phenomenon involving high repeatability in the timing of spiking onset during multiple repeats of a stimulus. We also demonstrate the utility of a probabilistic model of the neural code with an application to the decoding of RGC responses.

This work was conducted in collaboration with Liam Paninski, Valerie Uzzell, E. J. Chichilnisky and Eero Simoncelli. It was presented orally at the Society for Neuroscience Meeting (SFN) 2004, and has been submitted for publication.

CHAPTER 1

Characterization of macaque retinal ganglion cell responses using spike-triggered covariance

Light responses of retinal ganglion cells (RGCs) exhibit a variety of nonlinear features, including an accelerating contrast-response function, nonlinear pooling of spatial information, and dynamic gain adjustment. Here we describe a spike-triggered covariance (STC) analysis which allows us to characterize such nonlinearities in RGC responses to stochastic stimuli. Moreover, we introduce a space-time separable variant of STC analysis, or "subspace STC", which allows us to separately examine the influence of spatial and temporal nonlinearities on the RGC response. This analysis provides a reduced-dimensional description of the stimulus features that drive a neuron's response, and allow us to constrain

a functional model of a neuron's input-output properties. Using these constraints, we fit a model which contains identical spatially-shifted subunits, a nonlinear combination rule and a divisive temporal feedback signal. We show that this model quantitatively predicts responses to novel stimuli much better than a classical LNP model characterized using reverse correlation.

Introduction

One of the central problems in sensory neuroscience is that of characterizing the relationship between sensory stimuli and the spike responses of neurons. This is sometimes called the "neural coding problem", as the spike trains of sensory neurons can be considered a code used to convey information about external stimuli to the brain. This general problem has been investigated in a large number of sensory areas using a wide variety of stimuli and experimental preparations.

Recently, much work has considered an explicitly statistical setting for studying the neural coding problem (Bialek, Rieke, Steveninck, & Warland, 1991; Simoncelli, Paninski, Pillow, & Schwartz, 2004; Paninski, 2003; Sharpee, Rust, & Bialek, 2004; Arcas & Fairhall, 2003). In this setting, the problem can be considered to be that of characterizing the probabilistic

relationship between stimuli and spike trains, or $P[r(t)|s(t)]$, where $r(t)$ is the response and $s(t)$ is the relevant stimulus at time t, respectively. Generally, the goal of such approaches is to describe $P[r|s]$ using a model, where the model describes the probability of spiking at a fixed moment in time as a function of the stimulus $f(s)$. These *encoding models*, so-called because they express the conditional probability underlying the encoding process (the mapping of a fixed stimulus to a stochastic spike response), provide a full description of a neuron's input-output properties. That is, they can be used to make quantitative predictions about the neuron's response to arbitrary novel stimuli.

Much work on the neural coding problem has focused on retinal ganglion cells (RGCs), which form the output layer of the retina. RGC spike trains encode all visual information which is transmitted to the brain, and there is much evidence to suggest that these spike trains contain a highly compressed representation of the visual signal (Balboa & Grzywacz, 2000; Ruderman & Bialek, 1994; Smirnakis, Berry, Warland, Bialek, & Meister, 1997). This makes RGCs a natural focal point for the study of neural coding, and clarifies the importance of developing models which can capture the functional transformations carried out by RGCs.

RGCs are a relatively well-studied class of neurons, and much is known about their physiological response properties. They have localized receptive fields and quasi-linear sensitivity to contrast, which makes them relatively easy to characterize qualitatively using classical stimuli. Nonlinearities in

RGC responses have also been well documented, including a rectifying, accelerating contrast-response function, several forms of contrast-gain control and nonlinear pooling of spatial information (Hochstein & Shapley, 1976; J. D. Victor & Shapley, 1979b; J. D. Victor, 1987; Benardete, Kaplan, & Knight, 1992; Chander & Chichilnisky, 2001). Qualitative models of these phenomena have been proposed, and have proven quite successful in explaining the types of linear and nonlinear behaviors observed in RGC responses. (see (Meister & Berry, 1999) for a review).

Nevertheless, RGC response properties exhibit substantial heterogeneity across species and cell types. Here we describe a set of tools for characterizing the response properties of individual macaque retinal ganglion cells, using extracellularly recorded responses to spatiotemporal white noise stimuli. One advantage of this approach is that, in addition to building a quantitative model of an individual neuron's response, we assess the influence of nonlinearities in the context of a rich spatiotemporally varying stimulus, rather than for a restricted set of hand-tuned or parametrically varying stimuli. In this paper, we illustrate how to investigate the response properties of a neuron using spike-triggered covariance analysis. We attempt to develop intuition about what kind of model to use and demonstrate how to fit such a model using the constraints obtained from STC.

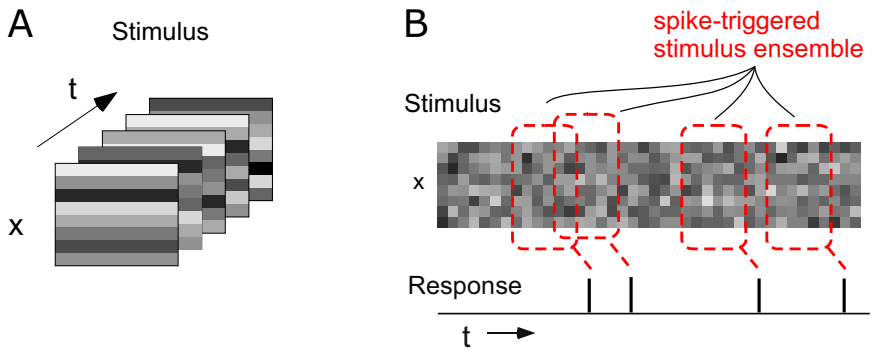

Figure 1.1: Stimulus and spike-triggered ensemble. **(A)** Depiction of the stochastic stimulus used to characterize retinal responses. The stimulus consisted of 1-dimensional bars, each of whose intensity was drawn randomly from a fixed-variance Gaussian distribution on each frame. **(B)** The spike-triggered stimulus ensemble consists of a set of stimulus "chunks" preceding each spike, illustrated here by the red boxes. These chunks represent the spatiotemporal stimulus region which causally influences the generation of each spike. In this example, the spike-triggered stimulus is shown to be a portion of the stimulus 6 frames long, extending from 8 frames before a spike to 2 frames before a spike, and spanning 8 bars in spatial extent. (See Methods for more details on how this window was selected for real cells). Each member of this (example) spike-triggered stimulus ensemble is therefore a 48-dimensional vector (each pixel of the 6 x 8 stimulus chunk is a dimension). Spike-triggered analysis proceeds generally by analyzing the statistical differences between the spike-triggered stimulus ensemble and the "raw" stimulus ensemble.

1.1 Neural Characterization

Figure 1.1 illustrates the experimental paradigm used in this experiment. Fig. 1.1A depicts the stimulus (flickering bars), and fig. 1.1B shows a two-dimensional representation of this stimulus (1D space, 1D time) alongside an example of the RGC spike response.

In the following analysis, we assume that the response $r(t)$ at any moment in time t is causally determined by some space-time portion of the

stimulus[1]. We denote the relevant stimulus portion by the vector $\vec{s}(t)$. (Note that the vectors associated with adjacent time bins have considerable overlap in their entries). The collection of all such stimulus vectors $\{\vec{s}(t)\}_{t\in[0,T]}$ comprises the "raw stimulus ensemble", and the subset of vectors $\{\vec{s}(t_i)\}$ associated with spike times $\{t_i\}$ is labeled the "spike-triggered stimulus ensemble", outlined in red in fig. 1.1B.

We can now restate the general problem as that of characterizing $P[r(t)=1|\vec{s}(t)]$, the probability of a spike occurring at time t given the associated space-time stimulus $\vec{s}(t)$. A straightforward solution to this problem would be to simply apply Bayes' rule to obtain:

$$P[r(t)=1|\vec{s}(t)] = \frac{P[r(t)=1, \vec{s}(t)]}{P[\vec{s}(t)]}, \tag{1.1}$$

where the numerator of the right hand side is the distribution of the spike-triggered stimulus ensemble and the denominator is the distribution of the

[1]Note that this assumption is not technically correct, given that the probability of spiking may also depend on the recent spike history of the neuron, as for example during the refractory period. This simplifying assumption is nevertheless quite useful in making the problem more tractable. It is equivalent to assuming that the neural response is an inhomogeneous Poisson process, which may be adequate to capturing the statistics of spike responses on a coarse time scale in many brain areas. A more general approach to the neural coding problem considers probabilistic models which condition on both the stimulus and spike train history; this approach has been pursued in several recent papers (Berry, Warland, , & Meister, 1997; Paninski et al., 2004; Pillow, Paninski, Uzzell, Simoncelli, & Chichilnisky, 2005)

raw stimulus ensemble. We know the denominator to be Gaussian, because the raw stimuli were drawn from a Gaussian distribution. The characterization problem therefore reduces to that of estimating the distribution of the spike-triggered stimuli.

Unfortunately, however, this observation does not directly lead to a tractable solution for estimating $P[r|\vec{s}]$, due to the so-called "curse of dimensionality". The stimulus ensemble occupies a very high-dimensional space (with dimensionality determined by the number of elements in $\vec{s}(t)$) , so it is impossible to obtain enough samples to estimate the spike-triggered stimulus distribution empirically[2]. Instead, we turn to a set of statistical tools for dimensionality reduction, in hopes of making the problem more tractable. The hope is that there will be only a small number of dimensions along which the distribution of the spike-triggered stimuli differ from the distribution of the raw stimuli. If we can discover these relevant dimensions, then we need only estimate the empirical distribution in the relevant subspace. In this way, we constrain a model which operates only on a particular subspace and still be confident of capturing the functional properties of the neural response.

[2]Consider, for example, that for an n-dimensional stimulus, we would need 2^n stimuli just to have a sample from each orthant of stimulus space

1.2 One-Dimensional Models and the STA

Reverse correlation, or spike-triggered averaging, represents an approach to system identification which goes back at least thirty years in neuroscience (deBoer & Kuyper, 1968; Marmarelis & Naka, 1972). However, rather than conceive of the spike-triggered average (STA) as the linear term in a series approximation to the neural functional (as in Volterra/Wiener series expansions), we can view it as a tool for dimensionality reduction. Mathematically, the STA is defined as the average stimulus preceding a spike:

$$\text{STA} = \sum_i \vec{s}(t_i), \tag{1.2}$$

where we let $\{t_i\}$ denote the set of spike times. Since our goal is to find a compact statistical representation of $P[r = 1, \vec{s}]$ (the numerator in eq. 1.1), the STA (i.e. mean of this distribution) is an obvious starting point.

Geometrically, we can view the STA as the difference between the means of the spike-triggered and the raw stimulus ensembles (the latter being zero, since we take the raw stimuli to be zero-mean Gaussian white noise). The STA con therefore be viewed as a vector which defines an axis or "direction" in stimulus space. Clearly, if the mean of the spike-triggered stimuli is different from that of the raw stimuli, then the probability of spiking varies as a function of the stimulus projection onto the STA. We can therefore approximate the full distribution $P[r|\vec{s}]$ using the one-dimensional distribution $P[r|x]$, where $x = STA \cdot \vec{s}$. This is simply the probability of spiking conditioned only on the stimulus projection onto the STA. We can

estimate this distribution quite easily using:

$$P[r|x] = \frac{\hat{P}[r=1,x]}{\hat{P}[x]}, \tag{1.3}$$

where the numerator and denominator are empirical estimates (e.g. histograms) of the *projected* spike-triggered and raw stimulus distributions.

We can also consider an alternative motivation for this characterization procedure. Imagine a neuron that computes its response by first linearly filtering the stimulus and then firing with some probability determined by a nonlinear function of filter output. We will refer to this model as the "one-dimensional LNP" model. It consists of a single linear filter (L), followed by a static nonlinearity (N), which converts the filter output to an instantaneous probability of firing, followed by Poisson spike generation (P). This model is well-known in the literature[3]. Mathematically, it can be written as:

$$P[r|\vec{s}] = f(\vec{k} \cdot \vec{s}); \tag{1.4}$$

where $\vec{k}$ is the linear filter and f is the nonlinearity.

[3]This model is sometimes referred to as an "LN cascade". We add the "P" in "LNP" to make explicit the assumption that the output of the "N" stage must be converted into a spike train via a Poisson process. This model includes most simple quasi-linear filtering models of neural response, such as the "difference-of-Gaussians" model of retinal ganglion cells, or the Gabor-filter model of V1 simple cells. A procedure for recovering the linear filter in such a model using sinusoidal stimuli was first demonstrated in retinal ganglion cells by (Enroth-Cugell & Robson, 1966)

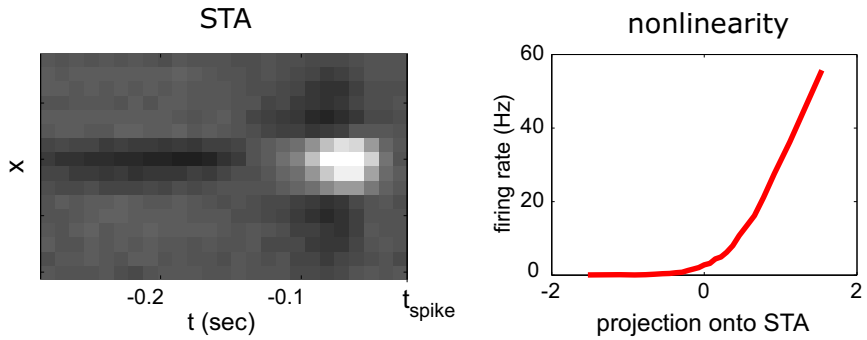

Figure 1.2: The spike-triggered average (STA) and point nonlinearity recovered for a typical ON retinal ganglion cell. The STA has center-surround spatial organization typical for RGCs, with a bright central region flanked by two dark regions on the vertical axis, and biphasic temporal organization along the horizontal axis. The nonlinearity (right) shows the instantaneous spike rate of the neuron as a function of the stimulus projection onto the STA. Together, the STA and nonlinearity provide a one-dimensional LNP model (Linear-nonlinear-Poisson) of the neuron's response.

For a one-dimensional LNP model, the STA serves as an estimator for the linear filter $\vec{k}$. A simple mathematical result due to (Bussgang, 1952) establishes that if:

1. the raw stimulus distribution $\{\vec{s}\}$ is spherically symmetric, *and*
2. the expected value of $f(\vec{k} \cdot \vec{s})$ under the distribution $P[\vec{s}]$ is not zero (i.e. the expected STA is not the zero vector)

then the STA gives an unbiased estimate of $\vec{k}$. A simple geometric proof of this result is presented in (Chichilnisky, 2001). Once we have obtained this estimate for $\vec{k}$, it is simple to estimate the nonlinearity f using the procedure described above (1.3).

Figure 1.2 shows the result of this characterization procedure applied

to a sample ON retinal ganglion cell. The STA (left) has spatiotemporal structure like that commonly observed in RGCs, with center-surround spatial structure (a bright region flanked by two dark regions) along the vertical axis and biphasic (contrast reversing) temporal structure along the horizontal axis.

The right plot shows the recovered nonlinearity for this cell, which gives the instantaneous firing rate as a function of the stimulus projection onto the STA. The nonlinearity rectifies strongly for stimuli with a negative projection onto the STA. It is interesting to note that representing this neuron's response with a Volterra/Wiener series expansion, which involves approximating this nonlinearity with a polynomial, would require a large number of terms to capture the shape of this nonlinearity accurately.

So far, we have discussed two motivations for using the STA to characterize neural responses, which are subtly different. Under the first, it is a handy tool for obtaining a one-dimensional approximation to the full (possibly multidimensional) neural coding distribution $P[r|\vec{s}]$. Under the second, it provides a simple, unbiased estimator of the true model responsible for generating neural responses. In the remainder of this paper, we will investigate which of these perspectives is more correct by examining whether we need to use a multidimensional representation of $P[r|\vec{s}]$ to accurately model the neural response of RGCs.

1.3 Multi-Dimensional Models and STC Analysis

Spike-triggered covariance (STC) analysis is an idea for moving beyond the mean of the spike-triggered ensemble, which is necessary for characterizing models that operate on more than one dimension of stimulus space. STC analysis in its modern form was introduced by (Brenner et al., 2000), was subsequently developed by (Schwartz et al., 2002), and has since been applied to characterizing sensory responses by several groups (e.g. (Touryan, Lau, & Dan, 2002; Rust, Schwartz, Movshon, & Simoncelli, 2004)). We review the technique briefly here.

The intuition underlying STC analysis is simple. We wish to examine the variance of the spike-triggered stimulus ensemble, and determine if there are any dimensions along which this variance differs from the variance of the raw stimulus ensemble. We can easily compute the variance of the spike-triggered stimuli along any direction in stimulus space, simply by left- and right-multiplying the spike-triggered covariance matrix (the covariance of the spike-triggered stimulus ensemble) with a unit vector $\vec{u}$ which points in this direction. That is,

$$\mathrm{var}_i(\vec{u} \cdot \vec{s}(t_i)) = \vec{u}^T \Lambda_1 \vec{u}, \tag{1.5}$$

where $\{s(t_i)\}$ is the set of spike-triggered stimuli, and Λ_1 is the spike-triggered covariance matrix, defined as

$$\Lambda_1 = \frac{1}{n} \sum_{i=1}^{n} \vec{s}(t_i) \vec{s}(t_i)^T \tag{1.6}$$

where n is the number of spike-triggered stimuli. It is worth noting that Λ_1 is also an estimate of the second-order kernel in a Volterra series expansion, though we do not intend to use it for this purpose. Rather, we intend to use the STC matrix only as a tool for dimensionality reduction of the neural coding distribution $P[r|\vec{s}]$. To achieve this, we want to find a set of directions $\{\vec{u}_i\}$ along which the variance of the spike-triggered stimuli differs maximally from that of the raw stimuli. We can find these directions by finding the maxima and minima of the objective function

$$g(\vec{u}) = \frac{\vec{u}^T \Lambda_1 \vec{u}}{\vec{u}^T \Lambda_0 \vec{u}} \tag{1.7}$$

(often called a Rayleigh quotient), where Λ_0 is the covariance matrix of the raw stimuli. The numerator is the variance of the spike-triggered stimuli along $\vec{u}$, and the denominator is the variance of the raw stimuli along $\vec{u}$, so the function computes the ratio of the two variances along a given direction in stimulus space.

Fortunately, eigenvector decomposition provides a standard solution to this problem. Let $\{\vec{v}_i\}$ be the set of eigenvectors of the special matrix

$$M = \Lambda_0^{-\frac{1}{2}^T} \Lambda_1 \Lambda_0^{-\frac{1}{2}}. \tag{1.8}$$

Then

$$\{\vec{u}_i\} = \{\Lambda_0^{\frac{1}{2}} \vec{v}_i\} \tag{1.9}$$

form a complete set of vectors producing the local extrema of the function g (eq. 1.7). The associated eigenvalues give the ratio between the variances of

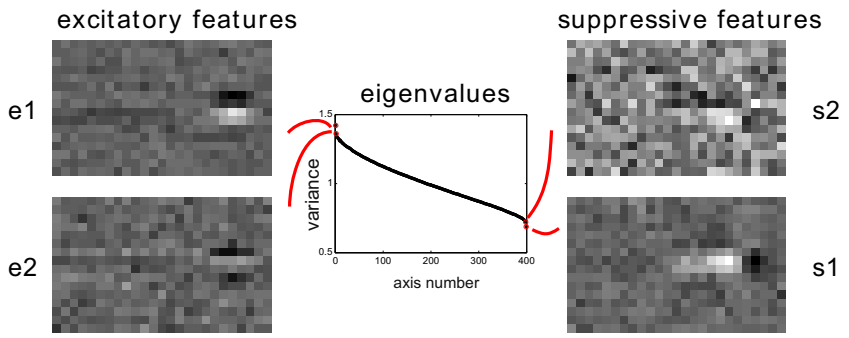

Figure 1.3: Excitatory and suppressive STC features for an ON retinal ganglion cell. Center graph shows the sorted eigenvalues of the spike-triggered stimulus covariance matrix. The largest two and smallest two eigenvalues (circled in red) were statistically determined to be significantly different than expected from random sampling of the raw stimuli. Left plots show the eigenvectors associated with the two largest eigenvalues. These stimulus patterns correspond to an increase in variance of the spike-triggered stimulus, meaning that the cell was more likely to spike if the projection of the stimulus onto these "excitatory" features had large magnitude (positive or negative). Right plots show the eigenvectors associated with the smallest eigenvalues, corresponding to a decrease in variance of the spike-triggered stimuli. The cell was less likely to spike if the stimulus had a large projection onto these "suppressive" features.

the spike-triggered and raw stimuli along each of these directions, which are guaranteed to be local maxima or local minima of g. Eigenvalues greater than 1 indicate that the spike-triggered stimuli have larger variance than the raw stimuli, whereas eigenvalues less than 1 indicate smaller variance.

Figure 1.3 shows the results of STC analysis applied to the ON cell shown in fig. 1.2. The central plot shows the sorted eigenvalues of M. These eigenvalues have some amount of scatter due to finite sampling of the covariance matrices Λ_0 and Λ_1. (i.e. the eigenvalues would disperse over some finite range even if the spike-triggered stimuli were a randomly selected subset

of the raw stimuli, so there was no statistical difference between the two sets). A statistical test is therefore necessary to determine whether these eigenvalues are higher (or lower) than expected given the number of samples. For this cell, the two largest eigenvalues and two smallest eigenvalues (circled in red) exceeded a statistical criterion, and were therefore taken to indicate stimulus directions along which the variance of the spike triggered stimuli differed significantly from that of the raw stimuli.

Plots on the left and right of fig. 1.3 show the eigenvectors associated with the significant high and low eigenvalues, respectively. The organized spatiotemporal structure apparent in these vectors provides an additional heuristic for the physiological relevance of the features. (Conversely, if we examine the vectors associated with non-significant eigenvalues, they appear random and exhibit no spatiotemporal organization).

For this cell, we can therefore conclude that response is calculated in a subspace of dimensionality at least 5 (the STA plus four STC axes). All 22 cells in the experiment exhibited at least one significant STC eigenvalue, meaning that no cell could be completely characterized using only the STA.

Technical aside: the equations 1.8 and 1.9 above are slightly more complicated than necessary if we are using Gaussian white noise stimuli and the raw stimulus covariance Λ_0 is well sampled. This derives from the fact that the raw stimulus covariance Λ_0 converges to the identity matrix as the number of raw stimuli grows relative to the stimulus dimensionality. In this limit, $\Lambda_0^{-\frac{1}{2}}$, which is sometimes called the *whitening matrix*, also converges

to the identity matrix, so we are left with stimulus features $\vec{u}_i$ that are the eigenvectors of the Λ_1. For this reason, we will occasionally slip into the lazy habit of referring to the $\vec{u}_i$ as eigenvectors of the STC.

Another important point to note is that because we are fundamentally interested in achieving a dimensionality reduction, we began this analysis by projecting all the stimuli $\vec{s}$ onto the subspace orthogonal to the STA. The rationale for taking this step is that the STA has already been identified as an axis which influences the neural response. We turn to STC analysis to identify *additional* dimensions, so we consider the STC only in the orthogonal subspace. The STC features $\vec{u}_i$ that we obtain are therefore all orthogonal to the STA.

Taken together, STA and STC analysis provide a linear basis for the subspace in which a neuron computes its response. We will label this basis B; it is composed of the STA plus the significant STC axes. We have therefore reduced the problem of modeling $P[r|\vec{s}]$ to that of modeling $P[r|\vec{y}]$, where $\vec{y} = B\vec{s}$ (i.e. the projection of $\vec{s}$ onto the basis B). Given that $\vec{y}$ is of much lower dimension than $\vec{s}$, this still represents an enormous reduction in the complexity of finding a model of the neural response.

However, we have not yet obtained a full model of the response because we have not yet characterized the nonlinear rule that converts $\vec{y}$ into a probability of spiking. In general, if the dimensionality of $\vec{y}$ is greater than 3, then we still have no hope of characterizing the nonlinearity with an empirical density estimate. Moreover, any model expressed in terms of

multidimensional empirical densities has limited interpretability in terms of the underlying neurobiology. Although we might perhaps obtain satisfactory functional models using such an approach, it is desirable to find models which have at least a loose connection to mechanisms commonly understood to exist in neurons.

For this reason, we view the basis B as a starting point, or a set of constraints for fitting a more biologically plausible model of RGC responses. The next step in our analysis is therefore to examine the structure of the STA and STC components in hopes of gaining insight into the kind of model which might give rise to their structure.

1.4 Separability and Subspace STC

We inspected the space-time structure of the STA and STC components obtained for each cell in a population of 22 RGCs. We observed that in all cells these components were approximately space-time separable, meaning that they could be decomposed as the outer product of a spatial profile with a temporal profile. The optimal such decomposition in a least-squares sense can be obtained using singular value decomposition of each filter. Figure 1.4 shows a space-time decomposition of the STA and STC features for the cell shown in fig. 1.3; each component was fit as the outer product of a 32-element time vector and an 18-element space vector.

Having noticed this regularity, we were motivated to apply a subsequent

STC analysis in order to dissociate spatial and temporal effects in the spike-triggered stimulus distribution. This analysis, which can be termed "subspace STC", relies on the simple idea of projecting the raw stimuli into a reduced-dimensional subspace and performing STC analysis there. We noted that if we filtered each temporal frame of the stimulus with a spatial filter derived from the spatial profile of the STA, we could obtain a stimulus which strongly predicted the neural response, but which was a function of time only. Conversely, if we filtered the time-history of each spatial bar of the stimulus with a temporal filter derived from the temporal profile of the STA, we could obtain a purely spatial stimulus.

One advantage of this approach is that it severely reduces the dimensionality of the stimulus, and therefore provides much more statistical power for detecting changes in variance. The raw stimulus had 576 dimensions (18 spatial bars x 32 stimulus frames), but the purely temporal and purely spatial stimulus had only 32 and 18 dimensions, respectively.

Figure 1.5 shows the result of this analysis applied to the ON cell examined previously. First, note that for the spatial analysis (above), the number of large eigenvalues (circled in red) increased to three. This can be explained by the increased statistical power due to the smaller dimensionality of the space. On the other hand, we found no small eigenvalues, indicating that there are no axes of reduced variance if we apply the STA temporal filter to each spatial location. The temporal analysis (fig. 1.5B) reveals a single small eigenvalue, which is more clearly separated from the

remaining eigenvalues than the two eigenvalues detected in the original analysis (fig. 1.3).

Now, we note first of all that the STC features obtained in the full STC analysis can be well approximated by combinations of the spatial and temporal STA and STC features obtained with subspace STC. For example, the feature associated with the smallest eigenvalue resembles the outer product of the spatial STA with the suppressive temporal eigenvector. The second suppressive feature resembles the outer product of the first excitatory STC feature and the suppressive temporal STC feature. Note also that although we can produce the original STC features using outer products of the subspace STC features, the subspace STC analysis has revealed something not readily apparent in the original analysis: that the neural response exhibits excitation over multiple spatial dimensions and suppression along a single temporal dimension. This observation served as motivation for a particular model of the RGC response consisting of shifted excitatory subunits and divisive temporal feedback.

1.5 Subunit model

Figure 1.6 shows a diagram of the subunit model we fit to the response of each RGC in our population. The motivation for this particular model arose from an analysis of the excitatory and suppressive features observed using subspace STC analysis. As shown in fig. 1.5, spatial features tended to have

more high-frequency spatial structure than the STA. This effect can arise functionally from the presence of multiple shifted subunits whose spatial structure is finer than that of the STA, and subunits have previously been hypothesized to explain the nonlinear characteristics of Y retinal ganglion cell responses in cat (J. D. Victor & Shapley, 1979a). Additionally, a single suppressive temporal feature whose structure resembled the derivative of the temporal STA suggested a form of temporal feedback suppression, as suppressive features of this type are observed in models such as integrate-and-fire that possess a refractory period (Arcas & Fairhall, 2003). These considerations led us to consider a model consisting of shifted subunits with a nonlinear combination rule and temporal feedback suppression.

In order to fit the parameters of this model, we began by searching for a subunit spatial profile such that shifted copies of this profile spanned the same linear subspace as the spatial STA and STC features (fig 1.5). We optimized the spatial profile and the spacing of subunits for each cell by minimizing the angle between the subspace spanned by the STA and STC features and that spanned by the shifted subunits. The resulting subunit profile and spacing obtained for a single ON cell is plotted in the diagram in figure 1.6.

We constrained the (common) temporal profile of the model subunits to lie in the subspace spanned by the temporal STA and suppressive STC feature, and fit its actual shape using maximum likelihood (a one-parameter fit, since we only need the angle in this 2-dimensional subspace). The re-

maining parameters, governing the two sigmoidal nonlinearities, combination weights, and kernel for the divisive temporal feedback signal (a 5-tap filter) were also fit using maximum likelihood, using 30 minutes of RGC response to a white noise stimulus. A statistical comparison between the performance of this model and that of the one-dimensional LNP model is presented in figures 1.7 and 1.8.

1.6 Model Validation

Figure 1.7 shows the results of a subspace-STC analysis (like in fig. 1.5) performed on the simulated response of the subunit model to a novel stimulus. Red lines indicate the spatial and temporal features obtained from the subunit model spike responses and grey lines indicate those of the original RGC. Note that although the subunit model contains no filters corresponding directly to any of the features revealed with spatial or temporal STC analysis (the only filters it contains are the shifted spatial profile of the subunits and a temporal filter not precisely matched to either temporal feature), the model nevertheless exhibits the same STC components as those observed in RGC responses. This illustrates consistency in the characterization procedure: that if we analyze the model output with the same statistical tools used to analyze the RGC responses, we obtain similar results.

It is worth noting that the simulated output of the one-dimensional LNP model (not shown) exhibits no significant eigenvalues if examined with STC

analysis, as we would expect with any one-dimensional model.

Figure 1.8 shows a more direct statistical comparison of the performance of the two models. Panel 1.8A shows a 2-second portion of a novel stimulus (not used for fitting the model parameters) and the associated RGC response. Panel 1.8B shows the rate predictions of both the subunit and LNP models, which helps to provide an intuition for how to compare the performance of the two models. Better prediction consists in having higher predicted firing rates during the times when spikes occurred, and lower firing rates during the periods of silence. Note that while the models show considerable agreement, there are obvious differences in the predictions.

The likelihood of the RGC response under each model prediction gives us a quantitative measure of the prediction accuracy of each model. Panel 1.8C shows a histogram of the ratio of mean likelihood-per-spike between the subunit and LNP model predictions across all cells in the population. A value of 1.2 means that the average interspike interval was 1.2 times more likely under the subunit model than under the LNP model. The ratio lies above 1 for all cells, indicating that the subunit model gave improved predictions for the responses of all 22 cells in the population.

Finally, figure 1.8D shows the difference in the spike-triggered rate predictions of the two models. This allows us to examine the qualitative differences in the rate predictions of the two models. The early peak indicates that subunit model gives a slightly higher rate prediction prior to the occurrence of a spike (and a higher rate prediction during the spike time

itself). Following a spike, the subunit model predicts a much greater decline in the spike rate than that predicted by the LNP model, a difference attributable to the nonlinear temporal feedback component included in the subunit model.

Discussion

We have explored the use of STA and STC analysis as tools for dimensionality reduction, in order to constrain models of the neural code. We performed STA and STC analysis on a population of macaque retinal ganglion cells, all of which exhibited at least a single STC axis in addition to the STA. This implies that a one-dimensional LNP model fails to capture the statistical features observed to be significant in RGC responses.

Subsequently, we introduced a space-time separable STC analysis of RGC responses by filtering the raw stimulus with either the spatial component or the temporal component of the STA. This analysis revealed excitatory spatial features with structure finer than the STA and suppressive temporal features with derivative-like structure. This motivated us to fit a model consisting of identical, shifted spatial subunits with a common temporal filter and nonlinearity, followed by an output nonlinearity and divisive temporal feedback. The spatial profile of the subunits was fit to span the same subspace as the spatial STA and STC vectors, while the temporal filter was fit in the subspace spanned by the temporal STA and STC vector.

The remaining parameters of the model were fit using maximum likelihood.

Finally, the performance of this model was compared to that of the one-dimensional LNP model, which provides a baseline point of comparison. Subspace STC analysis applied to simulated spike trains of the subunit model revealed features like those observed in RGC responses, whereas no STC features were observed in simulated LNP responses. In all cells, novel spike trains were better predicted by the subunit model than the one-dimensional LNP model.

1.7 Methods

Experimental Measurements & Stimuli

Multi-electrode extracellular recordings were obtained *in vitro* from a small piece of retina in a macaque monkey, with retinal pigment epithelium attached, maintained at 32-36 degrees C, pH 7.4. The retina was stimulated with a photopic, achromatic, spatially varying, optically reduced image of a cathode ray tube display refreshing at 120 Hz. The stimulus was a spatiotemporal sequence consisting of vertical bars, each of whose intensity was drawn i.i.d. from a Gaussian distribution of fixed variance on every refresh of the monitor. The contrast (standard deviation divided by mean) of the sequence was 48%. Model characterization was performed on one pseudo-random sequence (30 min), and model validation was performed on a different sequence (10 min). Analysis was restricted to two physiologically-

defined classes of cells that very likely correspond to ON and OFF parasol cells based on several lines of evidence (Chichilnisky & Kalmar, 2002).

Choosing the dimensionality of the spike-triggered stimulus ensemble

STC analysis requires selecting a particular space-time portion of the stimulus to be regarded as causally responsible for the generation of each spike. Clearly it is reasonable to assume some such finite window exists, since receptive fields are limited in both their temporal and spatial extent. Choosing the size of this window involves a tradeoff between statistical power and completeness in the representation of spatiotemporal features. Choosing a large window entails a high-dimensional spike-triggered stimulus ensemble. (The dimensionality of this space is determined by the number of "pixels" in the space-time stimulus: the number of bars times the number of time samples included). High dimensionality, in turn, leads to low statistical power for detecting significant STC effects, since the covariance of the raw and spike-triggered stimuli (which has $O(n^2)$ terms, where n is dimensionality of the stimulus) is much harder to sample in high dimensions; the number and dispersion of eigenvalues goes up with dimension, making it harder to detect significant changes in variance.

On the other hand, choosing a smaller window means that we might miss functional dependencies which lie outside the window selected. In practice, we applied an ad hoc procedure for determining the size of the spike-

triggered stimulus. We inspected the STA and STC features that emerged as window size was varied. In general, we did not find STC features exhibiting obvious dependency on spatiotemporal regions outside those present in the STA. We therefore selected a window of 32 time-samples and a variable number of spatial samples depending on the STA of each cell, but which generally extended one bar on either side beyond the visible surround of the STA. Generally we sought a window size that maximized the number of significant STC eigenvalues, though in most cells this was a relatively smooth function.

When performing subspace STC analysis, we included all 18 spatial dimensions of the stimulus, since the dimensionality was much lower and the covariance therefore much easier to estimate.

Other technical details

Mean likelihood-per-spike for each model (figure 1.8) was computed using

$$\exp(\frac{1}{n}\sum_{j=0}^{T}\log P[r_j|\lambda_j(\Delta t)], \tag{1.10}$$

where n is the number of spikes T is the number of time bins, indexed by j, and $P[r_j|\lambda_j(\Delta t)]$ is the probability under Poisson statistics of observing r_j spikes (the number of spikes actually observed in the jth time bin) in a bin of width Δt with a predicted rate of λ_j for that bin. The ratio of the subunit model's to the LNP model's mean likelihood per spike is shown in figure 1.8C.

Spike-triggered rate prediction: the cross-correlation between the actual spike train $r(t)$ and the predicted rate r_{pred}, or $\langle r(t) \cdot r_{pred}(t+\tau) \rangle$.

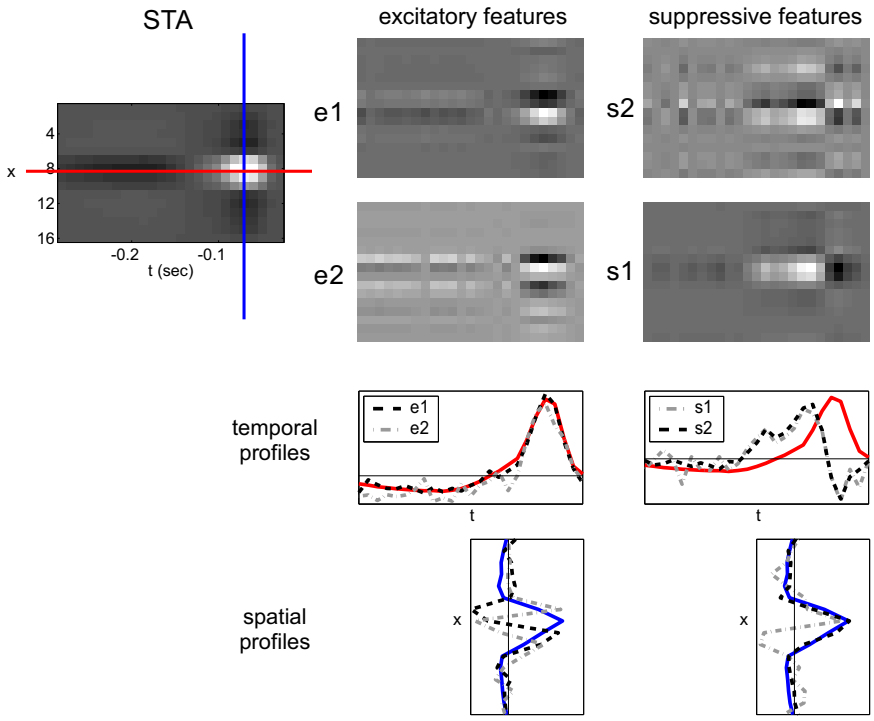

Figure 1.4: Space-time separable analysis of the STA and STC features. **Upper-left:** Spike-Triggered Average of the cell shown in figures 1.2 and 1.3, fit as a the outer product of a single spatial and a single temporal vector. This space-time separable filter captured 95% of the variance of the full STA. **Upper-right:** Space-time separable fits to the excitatory and suppressive stimulus features discovered for this cell using STC. These fits capture the dominant space-time structure in each filter (shown in fig. 1.3). **Below:** Plots of spatial and temporal structure of the STA and STC features. Red lines indicate the temporal and blue lines indicate spatial profiles of the STA. $e1$ and $e2$ denote the vectors ("excitatory) features") associated with the two largest eigenvalues (middle). $s1$ and $s2$ denote the vectors ("suppressive features") associated with the smallest and second-smallest eigenvalues, respectively (right).

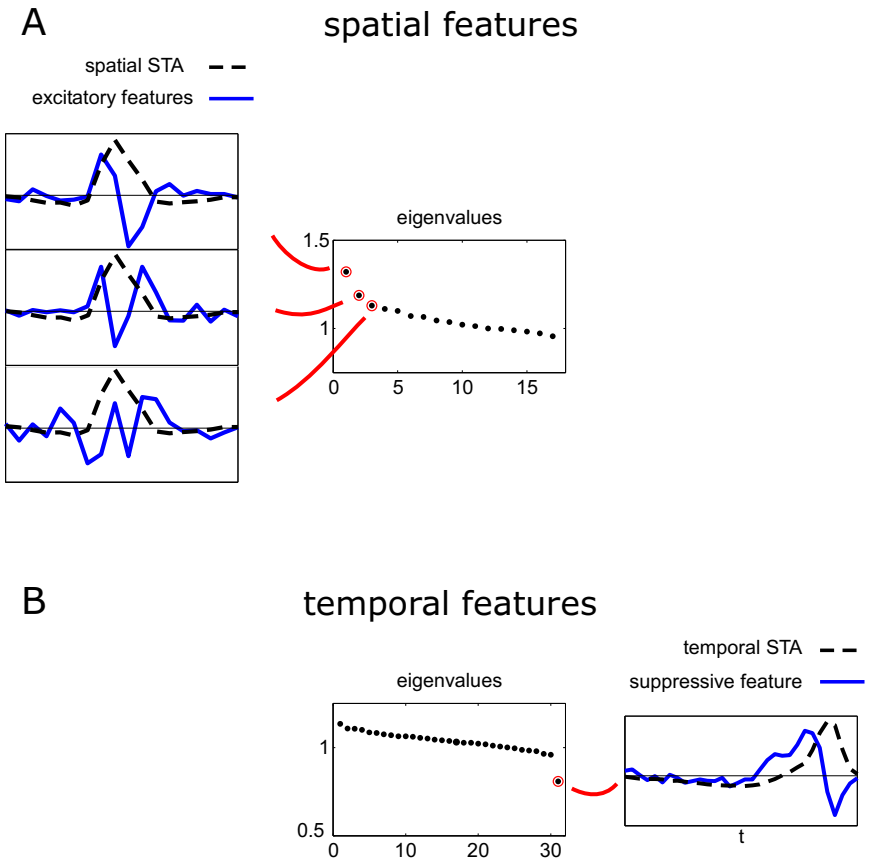

Figure 1.5: Results of space-time separable (subspace) STC analysis. **(A)** Spatial features: each bar of the stimulus was convolved with a temporal filter corresponding to the temporal profile of the STA (shown in fig. 1.4). This amounts to a projection of the original stimulus into a subspace where response is a function of space only. We performed STC analysis in this subspace, which resulted in three eigenvalues significantly larger than those of the raw stimuli (middle), and none smaller. The associated eigenvectors are shown in blue (left), and exhibit finer structure than the spatial STA (dashed trace). **(B)** Temporal features: each temporal frame of the stimulus was filtered with the STA spatial profile, yielding a purely temporal stimulus. STC analysis in this subspace yields a single suppressive feature, whose shape (right) resembles the derivative of the temporal STA.

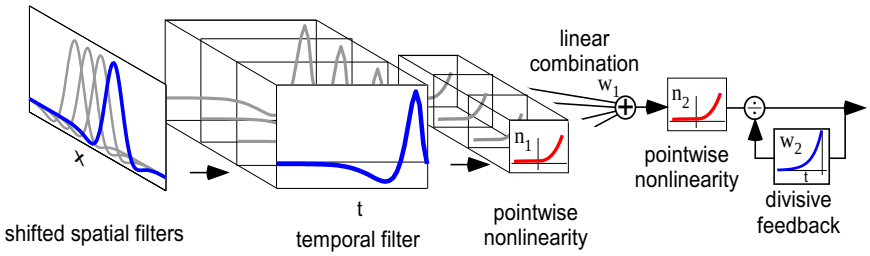

Figure 1.6: Schematic of subunit model. It consists of a set of identical, evenly spaced linear subunits, each of whose output undergoes an identical sigmoidal nonlinearity. The subunit outputs are then combined via a set of linear weights and the value obtained is put through an output nonlinearity, which is then modified divisively by a linearly filtered version of the model output.

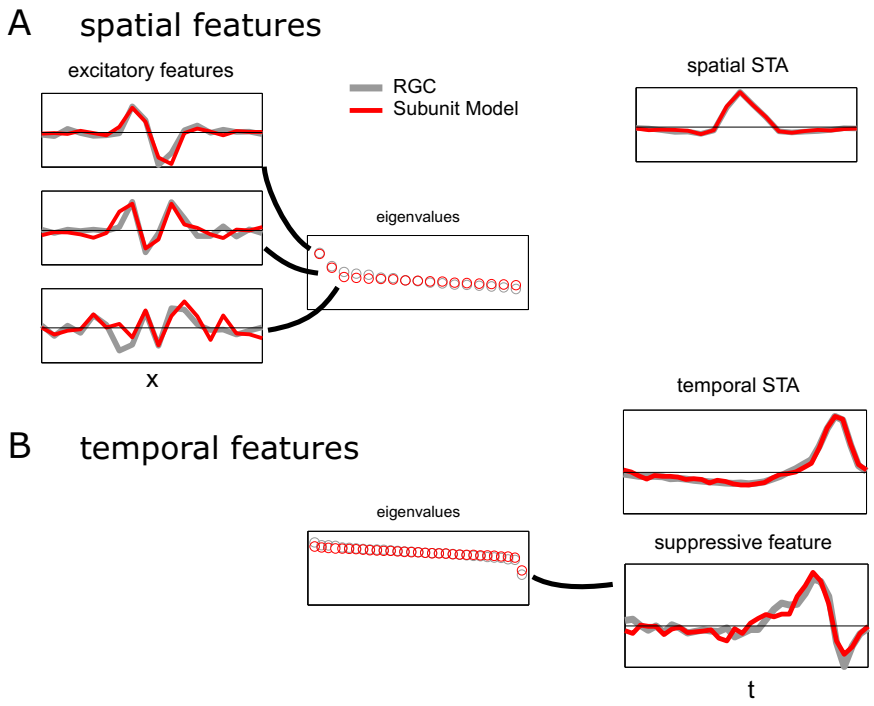

Figure 1.7: STA and STC features of the subunit model compared with those of the real RGC cell. The subunit model was simulated with a stimulus like that used to characterize the RGC (30 min spatio-temporal white noise), and an identical subspace STC analysis was applied to the model's output. **(A)** Analysis of spatial features. Red lines refer to features of the subunit model response and gray lines denote those obtained previously in the RGC response. Spatial STA is shown on the right. The sorted eigenvalues of the spatial stimuli (middle, obtained by filtering with the temporal STA as in fig. 1.5) exhibit three excitatory features, shown at left to be in good agreement with the corresponding features of the RGC. **(B)** The temporal STA of the subunit model (above right) closely matches that of the real cell. STC analysis of the temporal stimuli (obtained by filtering with the spatial STA) reveals a single suppressive feature, which agrees closely with feature from the RGC response.

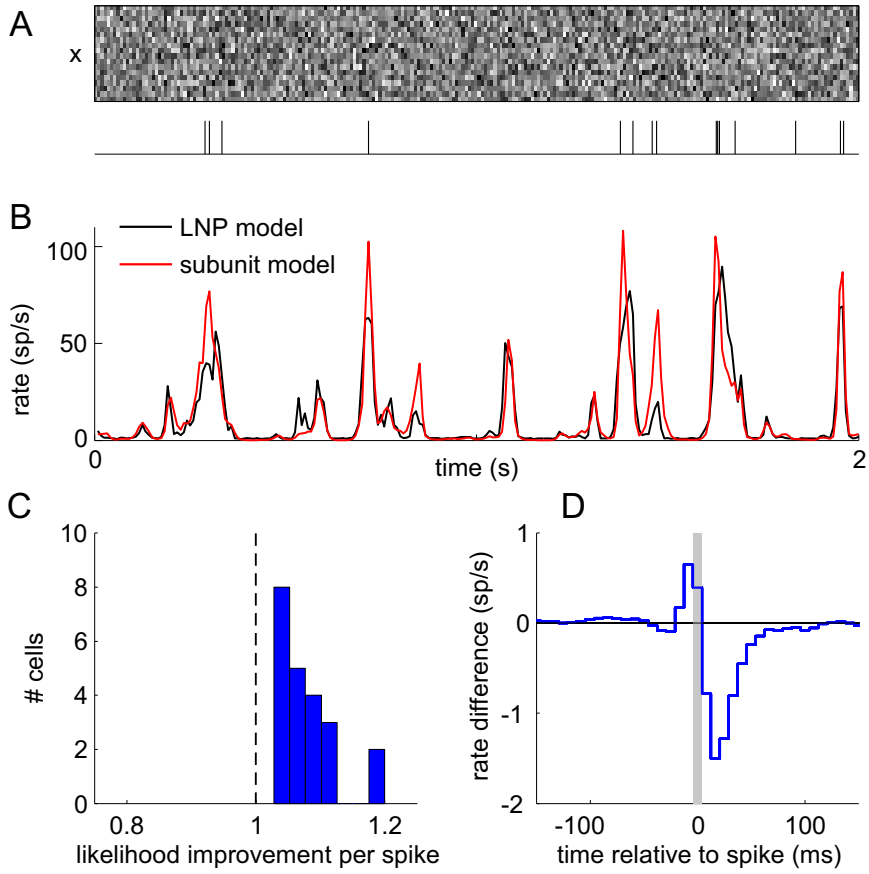

Figure 1.8: Subunit model comparison and validation. **(A)** Two-second portion of a novel stimulus (above) and associated RGC spike response (below). **(B)** Associated rate predictions of the one-dimensional LNP model (black) and the subunit model (red). The likelihood of the actual response under each model was computed by binning the response, computing the probability of the spike count observed in each bin (zero or one) given the predicted rate, and taking the product over all bins. For this two-second segment, which contained 15 spikes, the subunit model predicted the observed spike train with an average likelihood per spike of 0.074, vs. 0.062 for the LNP model. **(C)** Improvement in the likelihood per spike of the subunit model relative to LNP. Histogram shows the ratio of the likelihood per spike under the subunit model to that under the LNP model, for each cell. Values above one indicate that the subunit model predicted RGC spike trains with higher probability than the LNP model. **(D)** The average difference in the spike-triggered rate prediction for the subunit and LNP model across all cells, which illustrates qualitatively how the predictions of the two models differ. The gray box highlights the timeshift t=0, which shows that the subunit model elicited a higher spike rate prediction for the time bins in which spikes actually occurred.

CHAPTER 2

Estimation of a Deterministic IF model

White noise analysis methods for characterizing neurons typically ignore the dynamics of neural spike generation, assuming that spikes arise from an inhomogeneous Poisson process. We show that when spikes arise from a leaky integrate-and fire mechanism, a classical white-noise estimate of a neuron's temporal receptive field is significantly biased. We develop a modified estimator for linear characterization of such neurons, and demonstrate its effectiveness in simulation. Finally, we apply it to physiological data and show that spiking dynamics may account for changes observed in the receptive fields measured at different contrasts.

White noise analysis has become a widely used technique for characterizing response properties of spiking neurons in sensory systems. A sequence of stimuli are drawn randomly from an ensemble and presented in rapid

succession, and one examines the stimuli that elicit action potentials. In the most widely used form of this analysis, one estimates a linear approximation to the receptive field (i.e. first-order Wiener kernel) by computing the spike-triggered average (STA); that is, the average stimulus preceding a spike (deBoer & Kuyper, 1968; Jones & Palmer, 1987). Under the assumption that spikes are generated by a Poisson process with instantaneous rate determined by linear projection onto a kernel followed by a static nonlinearity, the STA provides an unbiased estimate of the underlying kernel (Chichilnisky, 2001).

The white noise approach is considered to have several advantages over traditional characterization approaches, including the the ability to explore a large portion of the input space and receptive field estimation that is robust to drift or fluctuation in the responsiveness of a neuron. Despite these advantages, it has also become clear that there are drawbacks to the characterizations obtained with white noise methods. One such shortcoming is the well-known phenomenon that the shape of the STA varies with the amplitude (e.g. contrast) of the white noise stimuli. (Smirnakis et al., 1997; Chander & Chichilnisky, 2001; Kim & Rieke, 2001, e.g.). This type of change cannot be explained by a linear model followed by a static nonlinearity and Poisson spike generation (the 'Linear-Nonlinear-Poisson', or L-N-P model), since it implies a change in the linear front end. We have previously shown that nonlinear suppressive interactions such as those found in cortical neurons can explain biases in the STA, that a spike-triggered co-

variance analysis can be used to characterize these suppressive interactions, and that the resulting corrected model can account for the changes of STA with contrast (Schwartz et al., 2002).

Here, we explore another potential source of failure in white noise characterization: the assumption of Poisson spike generation. The significance of temporal dynamic (i.e. non-Poisson) properties of biological spike generation for white noise characterization of neurons has not been thoroughly analyzed.

However, we show that in simulated white noise experiments, a linear model which drives an integrate-and-fire spiking mechanism is inaccurately characterized by the STA. Furthermore, we show that the integrative behavior of this model can account for some of the changes in STA estimated at different stimulus amplitudes in real neurons. Finally, we propose a new method for recovering the linear temporal filter governing neural response. We demonstrate through simulation that this approach can correctly estimate the linear kernel of a model neuron, and we also apply our method to real neural data, demonstrating that the recovered linear kernel is fairly stable with changes in stimulus contrast. We thus conclude that the recovered linear kernel may provide a more fundamental functional description of neural behavior, and might well be more directly related to the mechanisms underlying neural response.

2.1 Leaky integrate-and-fire model

Our analysis is based on a leaky integrate-and-fire (LIF) model. The input is convolved with a linear filter K, and this response drives a leaky integrator. When the level of this integrator reaches a threshold value, the neuron fires a spike and the integrator is reset to zero. The time evolution of the model membrane potential $V(t)$ is characterized by a single differential equation:

$$\frac{dV}{dt} = -\frac{1}{\tau}V(t) + I(t), \tag{2.1}$$

where τ is the time constant governing decay of the membrane potential, and $I(t)$ is the input current, generated by convolving the input signal $S(t)$ with the fixed kernel K:

$$I(t) \;=\; K * S(t) \;=\; \int_{-\infty}^{0} K(u)S(t-u)du. \tag{2.2}$$

This model has an analytical solution relative to the time of the most recent spike:

$$V(t) = \int_{t^-}^{t} I(u)e^{(u-t)/\tau}du, \tag{2.3}$$

where t^- is the time of occurrence of the last spike before t. This dependence on the time of the previous spike (and past input to the integrator) represents a fundamental departure from L-N-P model described earlier, where the probability of firing a spike is an instantaneous function of the projection of the stimulus onto K.

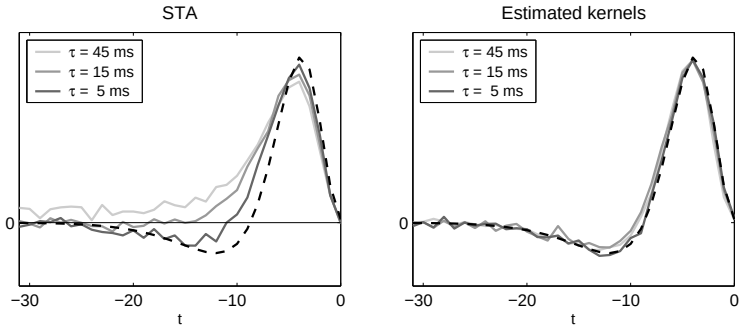

Figure 2.1: Simulation of integrate-and-fire neuron. **Left:** STA kernels retrieved for three different contrast levels (solid lines), plotted along with the true model kernel (dashed curve). **Right:** Kernels recovered using our algorithm.

2.2 Simulation results and comparison

We simulated a white noise analysis experiment with the model described above. In our simulations, the kernel K was chosen to be a 32-sample function whose shape loosely resembles temporal kernels measured in retinal ganglion cells. As in classical white noise experiments, we generated a random discrete stimulus $S(t)$ that was temporally white, drawing the stimulus intensity as an independent Gaussian random variable in each time step. We computed the STA as the average stimulus in the 32 time bins preceding each spike.

Figure 2.2(left) shows a plot of the actual kernel K superimposed on the STA for three different values of the membrane time constant τ. First, note that in all three cases, the STA differs significantly from K. This bias reflects the integrative spiking mechanism of the LIF model, as

the STA is quite close to K if the same input were given to an L-N-P model (Chichilnisky, 2001). Furthermore, the discrepancy between K and the STA depends on τ. For small τ (i.e. rapid decay of V), the STA is more closely resembles K, whereas larger τ (slower decay) gives rise to an STA which is smoother and more biased away from the true K. Note that although this basic effect is unsurprising, it is *not* the case that the STA shape arises simply from a low-pass filtering of K with an exponential filter. Specifically, the STA measured for a stand-alone LIF spike generator is decidedly non-exponential.

Physiological evidence indicates that at higher firing rates, the membrane conductance of neurons increases, which corresponds to a decrease in membrane time constant τ (Borg-Graham, Monier, & Fregnac, 1998; Hirsch, Alonso, Reid, & Martinez, 1998; Anderson, Lampl, Gillespie, & Ferster, 2000). Moreover, STAs measured in real neurons at high contrast tend to be narrower than those measured at low contrasts. This suggests that an integrative spiking mechanism with time constant that depends on firing rate is at least consistent with contrast-dependent changes in the STA of real neurons.

2.3 Recovering the linear kernel

Assuming that the input to an integrate-and-fire spiking model is determined by projection onto a linear kernel, how can the kernel be recovered

from the response to white noise stimuli? Equation (2.3) provides a deterministic expression for the voltage at any time since the most recent spike. The voltage at any spike time is therefore given by:

$$V(t^+) = V_{th} = \int_{t^-}^{t^+} \Big[K * S(t)\Big] e^{(t-t^+)/\tau} dt, \tag{2.4}$$

where V_{th} is threshold, t^- is the time index of the previous spike and t^+ that of the current spike. Using equation 2.2, we can rewrite this (by switching the order of integration):

$$V_{th} = \int_{-\infty}^{0} K(u) \left[\int_{t^-}^{t^+} S(t-u) e^{(t-t^+)/\tau} dt\right] du. \tag{2.5}$$

Note that, for fixed τ, this equation provides a linear constraint on K, since it expresses V_{th} as the inner product of K with the exponentially weighted S (back to the time of the previous spike). Every spike in the spike train provides one such constraint, so a discretized K can be overconstrained so long as its dimensionality is smaller than the number of spikes collected. K can easily be estimated by finding the least squares solution to this overconstrained linear system.

In practice, one would like to estimate both τ and K simultaneously, since both are unknown for data collected in real neurons. This can be achieved simply using a nested optimization (a line search algorithm) to find the τ which minimizes the squared error in the least squares solution for K. This algorithm is guaranteed to converge, and although the solution may not be only a local minimum, in simulations it was well-behaved for a

wide variety of kernel shapes and a large range of τ values. Figure 2.2(right) shows the kernels estimated for simulations conducted with three different values of τ. (Close estimates of the true values of τ were also obtained.) For both graphs in this figure, the stimulus contained 40,000 time samples and approximately 2,000 spikes were collected for each τ.

It should be noted, finally, that this estimator for K and τ ignores a huge set of additional constraints– namely, that $V(t)$ be less than threshold at all other times. However, because the problem is already overconstrained by the constraint on $V(t)$ at spike times, and because the additional constraints are much harder to implement, they can be ignored. A significant improvement to the estimator may nevertheless be obtained by considering additional constraints only on the time steps immediately preceding a spike. (This can be implemented by allowing a contribution to the squared error for any pre-spike time bin where V exceeds threshold). Montecarlo simulations exhibit rapid convergence to the true values of K and τ for this revised estimator.

2.4 Recovering a kernel from neural data

Our procedure for linear kernel estimation is based on an overly simplistic integrate-and-fire model for neural spike generation. We thus cannot be sure it will be applicable to real neural data. But we note that STA techniques have been used for decades to estimate linear kernels under the assumption

of a Poisson spike generator. The integrate-and-fire model incorporates a dependence on the time of the previous spike and is likely to provide a more accurate description of spiking in real neurons.

We have applied our procedure directly to data drawn from a monkey retinal ganglion cell (Chander & Chichilnisky, 2001). The data were recorded *in vitro*, using a stimulus consisting of 80,000 time samples of full-field 120 Hz flickering binary white noise The stimulus vectors $\vec{s}$ of this sequence are defined over a 25-segment (0.21 sec) time window. Two data sets were recorded, at contrasts of 32% and 64%.

Figure 2 (left) shows example STA estimates for both contrast levels. The kernels are quite different; the low-contrast STA is smoother and its peak that is shifted earlier in time than the high contrast STA. Figure 2 (right) shows the kernels resulting from our estimation procedure. Note that the estimated kernel is now quite stable across different contrasts, a desirable property for a functional description of neural behavior. The recovered time constants of 19.1 msec and 6.5 msec are within ranges considered biologically plausible, although their ratio indicates a greater change with amplitude than is commonly reported for cortical neurons (Borg-Graham et al., 1998; Hirsch et al., 1998; Anderson et al., 2000, e.g.).

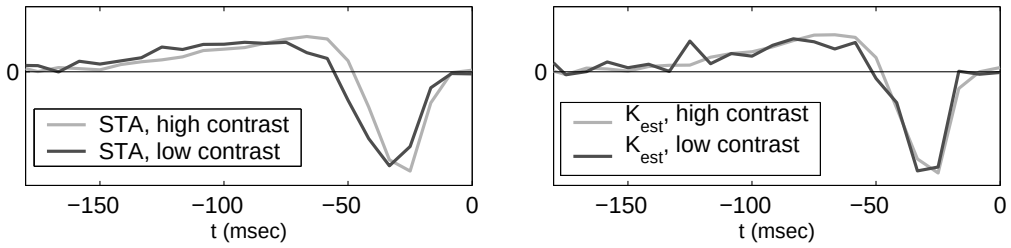

Figure 2.2: Characterization of macaque retinal ganglion cell responses **Left:** STA estimates based on responses recorded at two different input contrast levels. **Right:** Kernels recovered using our procedure. The associated time constant estimates are 19.1 and 6.5 msec.

2.5 Discussion

Our results show that spike generation mechanisms can affect the interpretation of results obtained with white noise analysis. In particular, we have shown that even for a simple integrate-and-fire model, the temporal STA does not accurately recover the temporal linear input kernel. For this model, the magnitude of bias in the STA is influenced by the membrane conductance, which is believed to vary with stimulus strength. This amplitude-dependence of the STA mirrors changes in the STA of real neurons measured at different contrasts, and cannot be captured by an L-N-P model.

Based on this simple LIF model, we have developed a new method for the recovery of the linear kernel integration time constant from responses to white noise stimluli. To our surprise, this kernel estimation procedure recovers a stable linear kernel when applied to data recorded from monkey

retinal ganglion cells, and the associated estimates of membrane conductance are within a biologically plausible range. Finally, while not discussed here, our technique also appears to be quite robust to the presence of noise in the membrane potential

We are currently exploring the generalization of these results to more realistic models. In particular, we have have found that the incorporation of a voltage floor in the model (corresponding to an ionic reversal potential) produces an STA which is sharper and closer to the true input kernel at high contrast, independent of any changes in membrane conductance. The significance of this phenomenon, along with that of other nonlinearities associated with spike generation, remains to to be analyzed.

Our results suggest a mechanistic explanation of the behaviors captured by current functional models of retinal ganglion cells (Shapley & Victor, 1981, e.g.), in which a nonlinear feedback signal is used to adjust the gain of the neuron. We have also previously shown that nonlinear gain control operations might account for a variety of apparent changes in receptive field properties at different contrast levels (Schwartz & Simoncelli, 2001). The results presented in this paper suggest that some such changes might be due to intracellular mechanisms of spike generation. It would be interesting to test such hypotheses against intracellular measurements.

CHAPTER 3

Estimation of a Stochastic, Recurrent IF model

Recent work has examined the estimation of models of stimulus-driven neural activity in which some linear filtering process is followed by a nonlinear, probabilistic spiking stage. We analyze the estimation of one such model for which this nonlinear step is implemented by a noisy, leaky, integrate-and-fire mechanism with a spike-dependent after-current. This model is a biophysically plausible alternative to models with Poisson (memory-less) spiking, and has been shown to effectively reproduce various spiking statistics of neurons in vivo. However, the problem of estimating the model from extracellular spike train data has not been examined in depth. We formulate the problem in terms of maximum likelihood estimation, and show that the computational problem

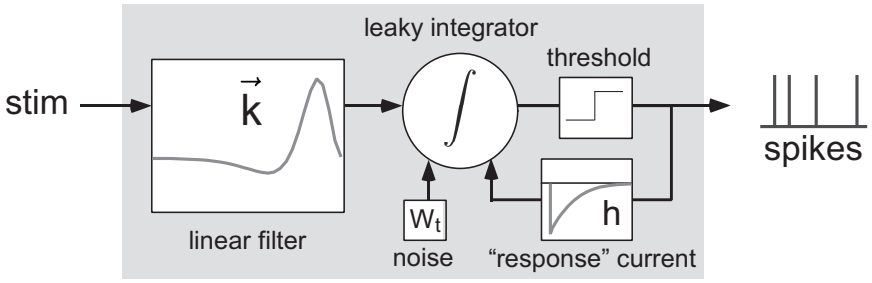

Figure 3.1: Schematic of the L-NLIF model.

of maximizing the likelihood is tractable. Our main contribution is an algorithm and a proof that this algorithm is guaranteed to find the global optimum with reasonable speed. We demonstrate the effectiveness of our estimator with numerical simulations.

A central issue in computational neuroscience is the characterization of the functional relationship between sensory stimuli and neural spike trains. A common model for this relationship consists of linear filtering of the stimulus, followed by a nonlinear, probabilistic spike generation process. The linear filter is typically interpreted as the neuron's "receptive field," while the spiking mechanism accounts for simple nonlinearities like rectification and response saturation. Given a set of stimuli and (extracellularly) recorded spike times, the characterization problem consists of estimating both the linear filter and the parameters governing the spiking mechanism.

One widely used model of this type is the Linear-Nonlinear-Poisson

(LNP) cascade model, in which spikes are generated according to an inhomogeneous Poisson process, with rate determined by an instantaneous ("memoryless") nonlinear function of the filtered input. This model has a number of desirable features, including conceptual simplicity and computational tractability. Additionally, reverse correlation analysis provides a simple unbiased estimator for the linear filter (Chichilnisky, 2001), and the properties of estimators (for both the linear filter and static nonlinearity) have been thoroughly analyzed, even for the case of highly non-symmetric or "naturalistic" stimuli (Paninski, 2003). One important drawback of the LNP model, however, is that Poisson processes do not accurately capture the statistics of neural spike trains (Berry & Meister, 1998; Keat, Reinagel, Reid, & Meister, 2001; Reich, Victor, & Knight, 1998; Arcas & Fairhall, 2003; Sakai, Fnahashi, & Shinomoto, 1999). In particular, the probability of observing a spike is not a functional of the stimulus only; it is also strongly affected by the recent history of spiking.

The leaky integrate-and-fire (LIF) model provides a biophysically more realistic spike mechanism with a simple form of spike-history dependence. This model is simple, well-understood, and has dynamics that are entirely linear except for a nonlinear "reset" of the membrane potential following a spike. Although this model's overriding linearity is often emphasized (due to the approximately linear relationship between input current and firing rate, and lack of active conductances), the nonlinear reset has significant functional importance for the model's response properties. In previous

work, we have shown that standard reverse correlation analysis fails when applied to a neuron with deterministic (noise-free) LIF spike generation; we developed a new estimator for this model, and demonstrated that a change in leakiness of such a mechanism might underlie nonlinear effects of contrast adaptation in macaque retinal ganglion cells (Pillow & Simoncelli, 2003). We and others have explored other "adaptive" properties of the LIF model (Rudd & Brown, 1997; Paninski, Lau, & Reyes, 2003; Yu & Lee, 2003).

In this paper, we consider a model consisting of a linear filter followed by noisy LIF spike generation with a spike-dependent after-current; this is essentially the standard LIF model driven by a noisy, filtered version of the stimulus, with an additional current waveform injected following each spike. This as the the "L-NLIF" model (Linear - Noisy Leaky Integrate-and-Fire) is illustrated in figure 3.1. The probabilistic nature of this model provides several important advantages over the deterministic version we have considered previously. First, an explicit noise model allows us to couch the problem in the terms of classical estimation theory. This, in turn, provides a natural "cost function" (likelihood) for model assessment and leads to more efficient estimation of the model parameters. Second, noise allows us to explicitly model neural firing statistics, and could provide a rigorous basis for a metric distance between spike trains, useful in other contexts (J. Victor, 2000). Finally, noise influences the behavior of the model itself, giving rise to phenomena not observed in the purely deterministic model (Levin & Miller, 1996).

Our main contribution here is to show that the maximum likelihood estimator (MLE) for the L-NLIF model is computationally tractable. Specifically, we describe an algorithm for computing the likelihood function, and prove that this likelihood function contains no non-global maxima, implying that the MLE can be computed efficiently using standard ascent techniques. The desirable statistical properties of this estimator (e.g. consistency, efficiency) are all inherited "for free" from classical estimation theory. Thus, we have a compact and powerful model for the neural code, and a well-motivated, efficient way to estimate the parameters of this model from extracellular data.

3.1 The Model

We consider a model for which the (dimensionless) subthreshold voltage variable V evolves according to

$$dV = \left(-gV(t) \; + \; \vec{k} \cdot \vec{x}(t) + \sum_{j=0}^{i-1} h(t - t_j) \right) dt + \sigma N_t, \qquad (3.1)$$

and resets to V_r whenever $V = 1$. Here, g denotes the leak conductance, $\vec{k} \cdot \vec{x}(t)$ the projection of the input signal $\vec{x}(t)$ onto the linear kernel $\vec{k}$, h is an "afterpotential," a current waveform of fixed amplitude and shape whose value depends only on the time since the last spike t_{i-1}, and N_t is an unobserved (hidden) noise process with scale parameter σ. Without loss of generality, the "leak" and "threshold" potential are set at 0 and 1,

respectively, so the cell spikes whenever $V = 1$, and V decays back to 0 with time constant $1/g$ in the absence of input. Note that the nonlinear behavior of the model is completely determined by only a few parameters, namely $\{g, \sigma, V_r\}$, and h (where the function h is allowed to take values in some low-dimensional vector space). The dynamical properties of this type of "spike response model" have been extensively studied (Gerstner & Kistler, 2002); for example, it is known that this class of models can effectively capture much of the behavior of apparently more biophysically realistic models (e.g. Hodgkin-Huxley).

Figures 3 and 3.4 show several simple comparisons of the L-NLIF and LNP models. In 1, note the fine structure of spike timing in the responses of the L-NLIF model, which is qualitatively similar to *in vivo* experimental observations (Berry & Meister, 1998; Reich, Victor, Knight, Ozaki, & Kaplan, 1997; Reich et al., 1998; Keat et al., 2001; Sakai et al., 1999)). The LNP model fails to capture this fine temporal reproducibility. At the same time, the L-NLIF model is much more flexible and representationally powerful, as demonstrated in Fig. 3.4: by varying V_r or h, for example, we can match a wide variety of dynamical behaviors (e.g. adaptation, bursting, bistability) known to exist in biological neurons.

3.2 The Estimation Problem

Our problem now is to estimate the model parameters $\{\vec{k}, \sigma, g, V_r, h\}$ from a sufficiently rich, dynamic input sequence $\vec{x}(t)$ together with spike times $\{t_i\}$. A natural choice is the maximum likelihood estimator (MLE), which is easily proven to be consistent and statistically efficient here. To compute the MLE, we need to compute the likelihood and develop an algorithm for maximizing it.

The tractability of the likelihood function for this model arises directly from the linearity of the subthreshold dynamics of voltage $V(t)$ during an interspike interval. In the noiseless case (Pillow & Simoncelli, 2003), the voltage trace during an interspike interval $t \in [t_{i-1}, t_i]$ is given by the solution to equation (3.1) with $\sigma = 0$:

$$V_0(t) = V_r e^{-gt} + \int_{t_{i-1}}^{t} \left(\vec{k} \cdot \vec{x}(s) + \sum_{j=0}^{i-1} h(s - t_j) \right) e^{-g(t-s)} ds, \qquad (3.2)$$

which is simply a linear convolution of the input current with a negative exponential. It is easy to see that adding Gaussian noise to the voltage during each time step induces a Gaussian density over $V(t)$, since linear dynamics preserve Gaussianity (Karlin & Taylor, 1981). This density is uniquely characterized by its first two moments; the mean is given by (3.2), and its covariance is $\sigma^2 E_g E_g^T$, where E_g is the convolution operator corresponding to e^{-gt}. Note that this density is highly correlated for nearby points in time, since noise is integrated by the linear dynamics. Intuitively, smaller leak

conductance g leads to stronger correlation in $V(t)$ at nearby time points. We denote this Gaussian density $G(\vec{x}_i, \vec{k}, \sigma, g, V_r, h)$, where index i indicates the ith spike and the corresponding stimulus chunk $\vec{x}_i$ (i.e. the stimuli that influence $V(t)$ during the ith interspike interval).

Now, on any interspike interval $t \in [t_{i-1}, t_i]$, the only information we have is that $V(t)$ is less than threshold for all times before t_i, and exceeds threshold during the time bin containing t_i. This translates to a set of linear constraints on $V(t)$, expressed in terms of the set

$$C_i = \bigcap_{t_{i-1} \le t < t_i} \Big\{ V(t) < 1 \Big\} \cap \big\{ V(t_i) \ge 1 \big\}.$$

Therefore, the likelihood that the neuron first spikes at time t_i, given a spike at time t_{i-1}, is the probability of the event $V(t) \in C_i$, which is given by

$$L_{\vec{x}_i, t_i}(\vec{k}, \sigma, g, V_r, h) = \int_{C_i} G(\vec{x}_i, \vec{k}, \sigma, g, V_r, h),$$

the integral of the Gaussian density $G(\vec{x}_i, \vec{k}, \sigma, g, V_r, h)$ over the set C_i.

Spiking resets V to V_r, meaning that the noise contribution to V in different interspike intervals is independent. This "renewal" property, in turn, implies that the density over $V(t)$ for an entire experiment factorizes into a product of conditionally independent terms, where each of these terms is one of the Gaussian integrals derived above for a single interspike interval. The likelihood for the entire spike train is therefore the product of these terms over all observed spikes. Putting all the pieces together, then,

the full likelihood is

$$L_{\{\vec{x}_i,t_i\}}(\vec{k},\sigma,g,V_r,h) = \prod_i \int_{C_i} G(\vec{x}_i,\vec{k},\sigma,g,V_r,h), \tag{3.3}$$

where the product, again, is over all observed spike times $\{t_i\}$ and corresponding stimulus chunks $\{\vec{x}_i\}$.

Now that we have an expression for the likelihood, we need to be able to maximize it. Our main result now states, basically, that we can use simple ascent algorithms to compute the MLE without getting stuck in local maxima.

Theorem 1. *The likelihood $L_{\{\vec{x}_i,t_i\}}(\vec{k},\sigma,g,V_r,h)$ has no non-global extrema in the parameters $(\vec{k},\sigma,g,V_r,h)$, for any data $\{\vec{x}_i,t_i\}$.*

The proof (see Appendix A) is based on the log-concavity of $L_{\{\vec{x}_i,t_i\}}(\vec{k},\sigma,g,V_r,h)$ under a certain parametrization of $(\vec{k},\sigma,g,V_r,h)$. The classical approach for establishing the nonexistence of non-global maxima of a given function uses concavity, which corresponds roughly to the function having everywhere non-positive second derivatives. However, the basic idea can be extended with the use of any invertible function: if f has no non-global extrema, neither will $g(f)$, for any strictly increasing real function g. The logarithm is a natural choice for g in any probabilistic context in which independence plays a role, since sums are easier to work with than products. Moreover, concavity of a function f is strictly stronger than logconcavity, so logconcavity can be a powerful tool even in situations for which concavity is useless

(the Gaussian density is logconcave but not concave, for example). Our proof relies on a particular theorem (Bogachev, 1998) establishing the logconcavity of integrals of logconcave functions, and proceeds by making a correspondence between this type of integral and the integrals that appear in the definition of the L-NLIF likelihood above.

We should also note that the proof extends without difficulty to some other noise processes which generate logconcave densities (where white noise has the standard Gaussian density); for example, the proof is nearly identical if N_t is allowed to be colored or non-Gaussian noise, with possibly nonzero drift.

3.3 Computational Methods and Numerical Results

Theorem 1 tells us that we can ascend the likelihood surface without fear of getting stuck in local maxima. Now how do we actually compute the likelihood? This is a nontrivial problem: we need to be able to quickly compute (or at least approximate, in a rational way) integrals of multivariate Gaussian densities G over simple but high-dimensional orthants C_i. We discuss two ways to compute these integrals; each has its own advantages.

The first technique can be termed "density evolution" (Knight, Omurtag, & Sirovich, 2000; Paninski, Lau, & Reyes, 2003). The method is based on the following well-known fact from the theory of stochastic differential equations (Karlin & Taylor, 1981): given the data $(\vec{x}_i, t_{i-1})$, the probabil-

ity density of the voltage process $V(t)$ up to the next spike t_i satisfies the following partial differential (Fokker-Planck) equation:

$$\frac{\partial P(V,t)}{\partial t} = \frac{\sigma^2}{2}\frac{\partial^2 P}{\partial V^2} + g\frac{\partial[(V - V_{eq}(t))P]}{\partial V}, \tag{3.4}$$

under the boundary conditions

$$\begin{aligned} P(V, t_{i-1}) &= \delta(V - V_r), \\ P(V_{th}, t) &= 0; \end{aligned} \tag{3.5}$$

where $V_{eq}(t)$ is the instantaneous equilibrium potential:

$$V_{eq}(t) = \frac{1}{g}\left(\vec{k}\cdot\vec{x}(t) + \sum_{j=0}^{i-1} h(t - t_j)\right). \tag{3.6}$$

Moreover, the conditional firing rate $f(t)$ satisfies

$$\int_{t_{i-1}}^{t} f(s)ds = 1 - \int P(V,t)dV. \tag{3.7}$$

Thus standard techniques for solving the drift-diffusion evolution equation (3.4) lead to a fast method for computing $f(t)$ (as illustrated in Fig. 2). Finally, the likelihood $L_{\vec{x}_i,t_i}(\vec{k},\sigma,g,V_r,h)$ is simply $f(t_i)$.

While elegant and efficient, this density evolution technique turns out to be slightly more powerful than what we need for the MLE: recall that we do not need to compute the conditional rate function f at all times t, but rather just at the set of spike times $\{t_i\}$, and thus we can turn to more specialized techniques for faster performance. We employ a rapid technique for computing the likelihood using an algorithm due to Genz

(Genz, 1992), designed to compute exactly the kinds of multidimensional Gaussian probability integrals considered here. This algorithm works well when the orthants C_i are defined by fewer than ≈ 10 linear constraints on $V(t)$. The number of actual constraints on $V(t)$ during an interspike interval $(t_{i+1} - t_i)$ grows linearly in the length of the interval: thus, to use this algorithm in typical data situations, we adopt a strategy proposed in our work on the deterministic form of the model (Pillow & Simoncelli, 2003), in which we discard all but a small subset of the constraints. The key point is that, due to strong correlations in the noise and the fact that the constraints only figure significantly when the $V(t)$ is driven close to threshold, a small number of constraints often suffice to approximate the true likelihood to a high degree of precision.

The accuracy of this approach improves with the number of constraints considered, but performance is fastest with fewer constraints. Therefore, because ascending the likelihood function requires evaluating the likelihood at many different points, we can make this ascent process much quicker by applying a version of the coarse-to-fine idea. Let L_k denote the approximation to the likelihood given by allowing only k constraints in the above algorithm. Then we know, by a proof identical to that of Theorem 1, that L_k has no local maxima; in addition, by the above logic, $L_k \to L$ as k grows. It takes little additional effort to prove that

$$\text{argmax } L_k \to \text{argmax } L;$$

thus, we can efficiently ascend the true likelihood surface by ascending the "coarse" approximants L_k, then gradually "refining" our approximation by letting k increase.

An application of this algorithm to simulated data is shown in Fig. 4. Further applications to both simulated and real data will be presented elsewhere.

3.4 Time Rescaling

Once we have obtained our estimate of the parameters $(\vec{k}, \sigma, g, V_r, h)$, how do we verify that the resulting model provides a self-consistent description of the data? This important "model validation" question has been the focus of recent elegant research, under the rubric of "time rescaling" techniques (Brown, Barbieri, Ventura, Kass, & Frank, 2002). While we lack the room here to review these methods in detail, we can note that they depend essentially on knowledge of the conditional probability of spiking $f(t)$. Recall that we showed how to efficiently compute this function in the last section and examined some of its qualitative properties in the L-NLIF context in Fig. 3.4.

The basic idea is that the conditional probability of observing a spike at time t, given the past history of all relevant variables (including the stimulus and spike history), can be very generally modeled as a standard (homogeneous) Poisson process, under a suitable transformation of the time

axis. The correct such "time change" is fairly intuitive: we want to speed up the clock exactly at those times for which the conditional probability of spiking is high (since the probability of observing a Poisson process spike in any given time bin is directly proportional to the length of time in the bin). This effectively "flattens" the probability of spiking.

To return to our specific context, if a given spike train was generated by an L-NLIF cell with parameters θ, then the following variables should constitute an i.i.d. sequence from a standard uniform density:

$$q_i \equiv \int_{t_i}^{t_{i+1}} f(s)ds,$$

where $f(t) = f_{\vec{x}_i,t_i,\theta}(t)$ is the conditional probability (as defined in the preceding section) of a spike at time t given the data $(\vec{x}_i, t_i)$ and parameters θ. The statement follows directly from the time-rescaling theorem (Brown et al., 2002), the inverse cumulative integral transform, and the fact that the L-NLIF model generates a conditional renewal process. This uniform representation, in turn, can be tested via standard techniques such as the Kolmogorov-Smirnov test and tests for serial correlation.

3.5 Extensions

It is worth noting that the methods discussed above can be extended in various ways, enhancing the representational power of the model significantly.

3.5.1 Interneuronal interactions

First, we should emphasize that the input signal $\vec{x}(t)$ is not required to be a strictly "external" observable; if we have access to internal variables such as local field potentials or multiple single-unit activity, then the influences of this network activity can be easily included in the basic model. For example, say we have observed multiple (single-unit) spike trains simultaneously, via multielectrode array or tetrode. Then one effective model might be

$$dV = \Big(-g(V(t) - V_l) + I_{stim}(t) + I_{hist}(t) + I_{interneuronal}(t) \Big) dt + W_t,$$

with the interneuronal current defined as a linearly filtered version of the other cells' activity:

$$I_{interneuronal}(t) = \sum_l \vec{k}_l^n \cdot n_l(t);$$

here $n_l(t)$ denotes the spike train of the l-th simultaneously recorded cell, and the additional filters k_l^n model the effect of spike train l on the cell of interest. Similar models have proven useful in a variety of contexts (Tsodyks, Kenet, Grinvald, & Arieli, 1999; Harris, Csicsvari, Hirase, Dragoi, & Buzsaki, 2003; Paninski, Fellows, Shoham, Hatsopoulos, & Donoghue, 2003); the main point is that none of the results mentioned above are at all dependent on the identity of $\vec{x}(t)$, and therefore can be applied unchanged in this new, more general setting.

3.5.2 Nonlinear input

Next, we can use a trick from the machine learning and regression literature (Duda & Hart, 1972; Cristianini & Shawe-Taylor, 2000; Sahani, 2000) to relax our requirement that the input be a strictly linear function of $\vec{x}(t)$; instead, we can write

$$I_{stim} = \sum_k a_k F_k[\vec{x}(t)]$$

where k indexes some finite set of functionals $F_k[.]$ and a_k are the parameters we are trying to learn. This reduces exactly to our original model when F_k are defined to be time-translates, that is, $F_k[\vec{x}(t)] = \vec{x}(t-k)$. We are essentially unrestricted in our choice of the nonlinear functionals F_k, since, as above, all we are doing is redefining the input $\vec{x}(t)$ in our basic model to be $\vec{x}^*(t) \equiv \{F_k(\vec{x}(t))\}$; under the obvious linear independence restrictions on $\{F_k(\vec{x}(t))\}$, then, the model remains identifiable (and in particular the MLE remains consistent and efficient under smoothness assumptions on $\{F_k(\vec{x}(t))\}$). Clearly the post-spike and interneuronal currents $I_{hist}(t)$ and $I_{interneuronal}(t)$, which are each linear functionals of the network spike history, may also be replaced by nonlinear functionals; for example, $I_{hist}(t)$ might include current contributions just from the preceding spike (Gerstner & Kistler, 2002), not the sum over all previous spikes.

Some obvious candidates for $\{F_k\}$ are the Volterra operators formed by taking products of time-shifted copies of the input $\vec{x}(t)$ (Dayan & Abbott,

2001; Dodd & Harris, 2002):

$$F[\vec{x}(t)] = \vec{x}(t-\tau_1) \cdot \vec{x}(t-\tau_2),$$

for example, with τ_i ranging over some compact support. Of course, it is well-known that the Volterra expansion (essentially a high-dimensional Taylor series) can converge slowly when applied to neural data; other more sophisticated choices for F_k might include, e.g., a set of basis functions (Zhang, Ginzburg, McNaughton, & Sejnowski, 1998) that span a reasonable space of possible nonlinearities, such as the principal components of previously observed nonlinear tuning functions (see also (Sahani & Linden, 2003) for a similar idea, but in a purely linear setting).

3.6 Discussion

We have shown here that the L-NLIF model, which couples a linear filtering stage to a biophysically plausible and flexible model of neuronal spiking, can be efficiently estimated from extracellular physiological data using maximum likelihood. Moreover, this model lends itself directly to analysis via tools from the modern theory of point processes. For example, once we have obtained our estimate of the parameters $(\vec{k}, \sigma, g, V_r, h)$, how do we verify that the resulting model provides an adequate description of the data? This important "model validation" question has been the focus of some recent elegant research, under the rubric of "time rescaling" techniques (Brown et al., 2002). While we lack the room here to review these

methods in detail, we can note that they depend essentially on knowledge of the conditional firing rate function $f(t)$. Recall that we showed how to efficiently compute this function in the last section and examined some of its qualitative properties in the L-NLIF context in Figs. 3.4 and 33.3.

We are currently in the process of applying the model to physiological data recorded both *in vivo* and *in vitro*, in order to assess whether it accurately accounts for the stimulus preferences and spiking statistics of real neurons. One long-term goal of this research is to elucidate the different roles of stimulus-driven and stimulus-independent activity on the spiking patterns of both single cells and multineuronal ensembles.

Appendix A: Proof of Log-Concavity of Model Likelihood

The following is a proof that the likelihood function for the L-NLIF model (a linear filter followed by noisy leaky integrate-and-fire spike generation) is logconcave (i.e., is the logarithm of a concave function), under a certain smooth, invertible reparametrization of the model parameters $\{\vec{k}, g, V_r, \sigma\}$. Here, $\vec{k}$ is a linear kernel which filters the incoming stimulus $x(t)$, g is the membrane leak conductance, V_r is the voltage reset, and σ is a scale parameter for the membrane noise. Since diffeomorphisms can neither create nor destroy local minima of smooth functions, this concavity result is a sufficient condition for establishing that the likelihood function has no non-

global local maxima, meaning that gradient ascent methods are guaranteed to find the global maximum of the likelihood function. See the full paper for a discussion of the properties and behavior of the L-NLIF model.

We recall that the model is governed by the stochastic differential equation

$$dV(t) = \Big[-gV(t) + \vec{k} \cdot \vec{x}(t) \Big] dt + \sigma n(t), \tag{3.8}$$

with $V(t)$ reset to V_r whenever $V(t) = 1$. The noise process $n(t)$ induces a Gaussian density on $V(t)$ during an interspike interval, which we denote $G_{X_i,\vec{k},g,V_r,\sigma}$, where subscripts indicate the density's dependence on the stimulus X_i (the stimuli that influence $V(t)$ during the ith interspike interval) and the model parameters. Note that in equation (3.8), $\vec{x}(t)$ denotes the stimulus vector that influences $dV(t)$ directly (via the linear filter $\vec{k}$) at time t, so X_i consists of all vectors $\big\{\vec{x}(t) : t \in [0, t_i]\big\}$.

Now, if the neuron resets at time $t = 0$, the probability of having a spike at time t_i is the probability that $V(t)$ lies in the constraint set C_i:

$$C_i = \bigcap_{0 \le t < t_i} \Big\{ V(t) < 1 \Big\} \cap \big\{ V(t_i) \ge 1 \big\},$$

the set of all voltage traces which do not exceed threshold until time t_i. The probability that $V(t) \in C_i$ is given by the integral of the density $G_{X_i,\vec{k},g,V_r,\sigma}$ over the set C_i. Because the noise in different interspike intervals is independent, the likelihood function for an entire spike train factorizes

as a product of such integrals, one for each observed interspike interval:

$$L_{\{X_i,t_i\}}(\vec{k},g,\sigma,V_r) = \prod_i \int_{C_i} G_{X_i,\vec{k},g,V_r,\sigma}(V)dV. \tag{3.9}$$

Our proof consists of demonstrating that integrals of the form

$$\int_{C_i} G_{X_i,\vec{k},g,V_r,\sigma}(V)dV$$

are logconcave for a particular reparametrization of $\{\vec{k},g,V_r,\sigma\}$. Because logconcavity is preserved under multiplication, this suffices to prove logconcavity for the entire likelihood function.

Proof

The proof is built on the following basic result (see, e.g., (Bogachev, 1998)).

Theorem ((Rinott, 1976)). *For any Borel sets A_1 and A_2, and t in $[0,1]$, define the set*

$$tA_1 + (1-t)A_2 \equiv \{ta_1 + (1-t)a_2 \mid a_i \in A_i, i = 1,2\}$$

If p is a logconcave probability density function on Euclidean space, then

$$\log p(tA_1 + (1-t)A_2) \geq t \log p(A_1) + (1-t)\log p(A_2) \ \forall t \in [0,1].$$

That is, the corresponding measure is logconcave.

(For strong uniqueness of the global maximum, we would need the simple extension of this result that if p is strictly logconcave and A_1 and A_2 have positive p-measure, then the inequality is strict for all t in the open unit

interval. To prove the nonexistence of local extrema, however, this is not necessary.)

As a corollary, we have that for any measurable convex set $R \subset \Re^n$, the integral of a logconcave density $p(x)$ over invertible affine mappings of R is a logconcave function of the map. More precisely,

$$f(M, B) = \int_{MR+B} p(x)dx$$

is logconcave in (M, B), where the set $\{MR + B\}$ is defined as all points $\{x : \ x = My + B, y \in R\}$, B is allowed to take values in the full Euclidean space, and M lives in some convex set of invertible matrices.

Our proof basically consists of translating this result into the terminology of our problem. We show specifically that the likelihood function can be written as the integral of a Gaussian density over a linearly parametrized family of sets. We begin by making a change of variables $Y = MV(t) + B$ so that $Y(t)$ has the standard normal density. We then show that the constraint set under this mapping, $MC_i + B$, is linear in a reparametrization of $\{\vec{k}, g, V_r, \sigma\}$. In other words, we find a set of parameters θ such that θ maps in a smooth, invertible way to $\{\vec{k}, g, V_r, \sigma\}$, and the matrices M and B are linear in θ. This satisfies the conditions of the corollary, since the standard normal density is logconcave and the constraint sets C_i are convex.

First, let us examine the density $G_{X_i,\vec{k},g,V_r,\sigma}(V)$. The mean, μ, of this Gaussian is equal to the solution of the noiseless version of the integrate-

and-fire dynamics, on the interval $[0, t_i]$:

$$\frac{dV}{dt} = -gV(t) \ + \ \vec{k} \cdot \vec{x}(t), \tag{3.10}$$

with initial data

$$V(0) = V_r.$$

Thus,

$$\mu(t) = V_r e^{-gt} + \int_0^t \Big(\vec{k} \cdot \vec{x}(s)\Big) \, e^{-gs} ds;$$

we rewrite this in operator form:

$$\mu(t) = E_g\big[V_r \delta(0) \ + \ k \cdot \vec{x}(t)\big], \tag{3.11}$$

where E_g is the convolution operator corresponding to e^{-gt}. The covariance of G, in turn, is given by

$$\sigma^2 E_g E_g^T. \tag{3.12}$$

As usual, G is completely characterized by its mean and covariance.

Now, if we reparametrize by

$$\begin{aligned} Y &= \frac{1}{\sigma} E_g^{-1}\Big(V(t) - \mu(t)\Big) \\ &= \frac{1}{\sigma} E_g^{-1}\Big(V(t) - E_g\big[V_r \delta(0) \ + \ k \cdot \vec{x}(t)\big]\Big) \\ &= \frac{1}{\sigma}\Big(E_g^{-1} V(t) - V_r \delta(0) - k \cdot \vec{x}(t)\Big), \end{aligned} \tag{3.13}$$

it is clear that Y has a standard normal distribution $N(0, I)$. Then set

$$M = \frac{1}{\sigma} E_g^{-1}, \tag{3.14}$$

$$B_i = -\frac{1}{\sigma}\big[V_r \delta(0) + k \cdot \vec{x}(t)\big], \tag{3.15}$$

allowing us to write $Y = MV + B_i$. Our constraint set under this change of variables becomes $D_i = MC_i + B_i$. It remains only to describe a diffeomorphism between $\theta = (M, B_i)$ and $\{\vec{k}, g, V_r, \sigma\}$, given the data X_i.

We begin this by noting that E_g^{-1} can be written as a bi-diagonal matrix

$$E_g^{-1} = \begin{bmatrix} 1 & & & & \\ -\alpha & 1 & & & \\ & -\alpha & 1 & & \\ & & \ddots & \ddots & \\ & & & -\alpha & 1 \end{bmatrix}, \tag{3.16}$$

where $\alpha = e^{-g\Delta t}$, and Δt is the time bin width. (This can be shown by direct computation). We can therefore write M as

$$M = \begin{bmatrix} a & & & & \\ -b & a & & & \\ & -b & a & & \\ & & \ddots & \ddots & \\ & & & -b & a \end{bmatrix}, \tag{3.17}$$

where $a = 1/\sigma$ and $b = \alpha/\sigma$. M is therefore linear in the parameters (a, b), and the original parameters can be recovered via $\sigma = 1/a$ and $g = -(\log b\sigma)/\Delta t$.

If we now turn to B_i, note first of all that $[\vec{x}(t) \cdot \vec{k} + V_r\delta(0)]$ can be

written as a matrix, allowing us to rewrite B_i:

$$B_i = -\frac{1}{\sigma} \begin{bmatrix} 1 & \vec{x}(t_0)^T \\ 0 & \vec{x}(t_1)^T \\ \vdots & \vdots \\ 0 & \vec{x}(t_i)^T \end{bmatrix} \begin{bmatrix} V_r \\ \vec{k} \end{bmatrix}. \tag{3.18}$$

Therefore, if we let

$$\theta = -\frac{1}{\sigma} \begin{bmatrix} V_r \\ \vec{k} \end{bmatrix},$$

then B_i is clearly linear in θ, since the matrix term depends only on the stimulus X_i. Moreover, we can recover $\vec{k}$ and V_r uniquely from θ since we already know the value of σ. We have therefore shown that the likelihood function for this model can be written as the integral of a Gaussian density over a linearly parametrized family of sets $MC_i + B_i$, which completes the proof.

Appendix B: Computing the Likelihood Gradient

The ascent of the likelihood surface is greatly accelerated by the computation of the gradient. This gradient can always be computed by finite differencing schemes, of course; however, in the case of a large number of parameters, it is much more efficient to compute gradients with respect to a few auxiliary parameters, then arrive at the gradient with respect to the full parameter set via the chain rule for derivatives. The following applies to

computing gradients of the likelihood computed using the Genz algorithm.

We focus on the discretized case for clarity. Thus, we take the derivatives with respect to the mean function $V_0(t)$, evaluated at the constraint times $\{t_k\}_{1\leq k\leq j}$. These derivatives turn out to be Gaussian integrals themselves, albeit over a $(j-1)$- instead of j-dimensional box, and can be easily translated into derivatives with respect to the parameters.

In order to derive the gradient, note that the discretized approximation to the likelihood can be written

$$L_j = \int_{-\infty}^{z_1} \cdots \int_{z_j}^{\infty} p(y_1, \ldots, y_j) dy_1 \cdots dy_j,$$

where y_k represent the transformed variables $y_k = V(t_k) - V_0(t_k)$, $z_k = 1 - V_0(t_k)$, and p denotes the corresponding Gaussian density, with 0 mean and covariance we'll call Λ. Now, the partial derivatives of L with respect to the z_k are:

$$\begin{aligned}
\frac{\partial}{\partial z_k} L &= \int_{-\infty}^{z_1} \cdots \int_{-\infty}^{z_{k-1}} \int_{-\infty}^{z_{k+1}} \cdots \int_{z_j}^{\infty} p(y_1, \ldots, y_k = z_k, \ldots, y_j) dy_1 \cdots dy_j \\
&= \left(\int_{C_{i\neq k}} p(\vec{y}_{i\neq k} | y_k = z_k) d\vec{y}_{i\neq k} \right) p(y_k = z_k),
\end{aligned}$$

with a sign change to account for the upward integral corresponding to the final, above-threshold constraint.

We can compute the marginal and conditional densities $p(y_k = z_k)$ and $p(\vec{y}_{i\neq k} | y_k = z_k)$ using standard Gaussian identities:

$$\begin{aligned}
p(y_k = z_k) &= \mathcal{N}(0, \Lambda_{k,k})(z_k), \\
p(\vec{y}_{i\neq k} | y_k = z_k) &= \mathcal{N}(\mu^*, \Lambda^*)(\vec{1}),
\end{aligned}$$

where

$$\begin{aligned}\mu^* &= \vec{V_0}(t_{i\neq k}) + \frac{z_k}{\Lambda_{k,k}}\vec{\Lambda}_{i\neq k,k} \\ \Lambda^* &= \Lambda_{i\neq k,h\neq k} - \frac{\vec{\Lambda}_{i\neq k,k}\vec{\Lambda}_{k,i\neq k}}{\Lambda_{k,k}}\end{aligned}$$

Thus, the gradient $\nabla_z L$ requires computing one Gaussian integral for each constraint z_k. From the vector $\nabla_z L$, we can use simple linear operations to obtain the gradient with respect to any of the parameters which enter only via $V_0(t)$, namely $h, \vec{k}$, and V_l.

Appendix C: Numerical Methods for Fokker-Planck Equation

In order to compute the likelihood function $L_{\{\vec{x}_i,t_i\}}(\vec{k}, \sigma, g, V_r, h)$, we used a second-order numerical method for solving the Fokker-Planck (FP), or "drift-diffusion" equation (eq. 3.4). This equation describes the time evolution of $P(V,t)$, the probability density over sub-threshold voltage V at time t, as a function of the input and model parameters.

Our general approach involves discretizing V so that we can represent $P(V,t^*)$ at a fixed time t^* by a set of discrete values. We then propagate this density forward in time using the FP equation to obtain $P(V,t^*+\Delta t)$, the probability over V at the next time step. Intuitively, the likelihood of a spike occurring during the interval $[t^*, t^*+\Delta t]$ is given by the amount of probability mass which leaks over the (absorbing) boundary at threshold

$(V = 1)$ during this time step.

We now describe the density propagation algorithm in detail. Let $\{v_i\}_{i=1}^n$ denote the discretization over V, consisting of n evenly spaced bins with a separation of Δv. We let $v_n = 1$ (threshold) and set v_1 to some voltage sufficiently low that we can represent $P(V)$ accurately at all time points. We will use i to index voltage and j to index time, so p_i^j denotes the probability mass associated with the ith bin of the voltage discretization and time bin j. And, in a slight abuse of notation, we will use p^j to refer to the entire density over voltage at the jth time step.

We initialize the algorithm with a density p^0, computed a short time after the most recent spike, when the subthreshold probability density over V is still well-approximated by a Gaussian. This initial density is given by

$$p_i^0 = N(v_i;\ V_{eq}^0, \tfrac{1}{2g}(1 - e^{-2gT})\sigma^2), \tag{3.19}$$

the standard Gaussian density with mean V_{eq}^0 and variance $\frac{1}{2g}(1 - e^{-2gT})\sigma^2$, evaluated at at each grid point v_i, where V_{eq}^0 is the instantaneous voltage reversal potential (eq. 3.6) and T is the time since the most recent spike.

Recall that the FP equation (eq. 3.4) for the model is given by

$$\begin{aligned} \frac{\partial P(V,t)}{\partial t} &= \frac{\sigma^2}{2}\frac{\partial^2 P}{\partial V^2} + g\frac{\partial[(V - V_{eq}(t))P]}{\partial V}, \\ &= \frac{\sigma^2}{2}\frac{\partial^2 P}{\partial V^2} + g(V - V_{eq}(t))\frac{\partial P}{\partial V} + gP \end{aligned} \tag{3.20}$$

We solve this equation using a scheme related to the Crank-Nicolson method for solving diffusive PDEs (Press, Teukolsky, Vetterling, & Flannery, 1992).

This involves substituting discrete approximations for the partial derivatives as follows:

$$\frac{p_i^{j+1} - p_i^j}{\Delta t} = \frac{\sigma^2}{2}\left[\frac{(p_{i+1}^{j+1} - 2p_i^{j+1} + p_{i-1}^{j+1}) + (p_{i+1}^j - 2p_i^j + p_{i-1}^j)}{2(\Delta v)^2}\right] + g(v_i - V_{eq}(t))\left[\frac{(p_{i+1}^{j+1} - p_{i-1}^{j+1}) + (p_{i+1}^j - p_{i-1}^j)}{4(\Delta v)}\right] + g\frac{p_i^{j+1} + p_i^j}{2}. \quad (3.21)$$

Note that the right-hand-side derivatives are evaluated by averaging over partial derivatives at the jth and $j+1$st time steps, leading to a method which is second-order accurate in V and t.

For the sake of clarity, we can rewrite (3.21) as a sparse matrix equation, which can be solved efficiently in $o(n)$ operations. We have:

$$\tfrac{1}{\Delta t}(p^{j+1} - p^j) = \tfrac{\sigma^2}{4(\Delta v)^2} D''\Big(p^{j+1} + p^j\Big) + \tfrac{g}{4(\Delta v)} D'(V - V_{eq}(t))\Big(p^{j+1} + p^j\Big), \quad (3.22)$$

where D' and D'' are tri-diagonal matrices corresponding to derivative and second-derivative operators (with the values $[\ -1 \quad 0 \quad 1\ ]$ and $[\ 1 \quad -2 \quad 1\ ]$ along the main diagonals, respectively), V is a diagonal matrix filled with the grid points $\{v_i\}$ along the main diagonal, and $V_{eq}(t_j)$ is a scalar which depends on the input during the current time step. By collecting like terms, this equation can be simplified to have the form

$$\Big(A + V_{eq}(t_j)\Big)p^{j+1} = \Big(B - V_{eq}(t_j)\Big)p^j, \quad (3.23)$$

where A and B are both tri-diagonal. We used a special routine written in C to solve this equation for p^{j+1}, which effects the density propagation.

Of course, we must also specify the correct boundary conditions (eq. 3.5) to ensure that probability mass leaks only one way across the spike threshold, which we use to compute p(spike) during each time step. We enforce the upper (absorbing) boundary condition by replacing the nth columns of the D' and D'' matrices with the nth column of the identity matrix (i.e. zero except for 1 in the nth position), which conserves probability mass in the last bin and drift or diffusion from p_n to p_{n-1}. We enforce the lower (reflecting) boundary condition by adding to the first entry of D' and D'' so that first column sums to 1, which ensures that probability mass is conserved at the lower boundary (i.e. it doesn't leak out of the range of $\{v_i\}$).

After having initialized the density at p_0, we perform density propagation (computing $V_{eq}(t)$ at each time step and solving equation 3.23) until we reach the next spike time t_k. Here, p_n^k gives the cumulative probability of a spike having occurred by time t_k, and the likelihood of a spike occurring at t_k is $\frac{p_n^k - p_n^{k-1}}{\Delta t}$.

Appendix D: The Gaussian process $V(t)$

Here we derive discrete and continuous solutions for the mean and variance of $V(t)$, the membrane potential of the IF model, in the absence of spiking. $V(t)$ is a Gaussian (Ornstein-Uhlenbeck) process, and therefore completely characterized by its mean and variance.

Mean:

The mean, $\mu(t)$, of the Gaussian (governing the evolution of $P(V)$, the density over membrane potential) is equal to the solution of the noiseless version of the integrate-and-fire dynamics, on the interval $[0, t_i]$ (eq. 3.10):

$$\frac{dV}{dt} = -gV(t) \; + \; \vec{k} \cdot \vec{x}(t)$$

with initial data

$$V(0) = V_r.$$

Thus,

$$\mu(t) = V_r e^{-gt} + \int_0^t \left(\vec{k} \cdot \vec{x}(s) \right) e^{-gs} ds.$$

To simplify notation, we can rewrite $\mu(t)$ in operator form:

$$\mu(t) = E_g \big[V_r \delta(0) \; + \; k \cdot \vec{x}(t) \big],$$

where E_g is the convolution operator corresponding to e^{-gt}.

If we consider the problem discretized in time bins of width Δt and set $\alpha = e^{-g\Delta t}$, the operator E_g can be written as a matrix:

$$E_g = \begin{bmatrix} 1 & & & & \\ \alpha & 1 & & & \\ \alpha^2 & \alpha & 1 & & \\ \vdots & & & \ddots & \\ \alpha^n & \dots & \alpha^2 & \alpha & 1 \end{bmatrix}. \tag{3.24}$$

The first row corresponds to the filtering during the first time bin and the nth row corresponds to filtering for the nth time bin, or $t = n(\Delta t)$ of the solution.

Analytically, for continuous time, we can express the mean as:

$$\mu(t) = \tfrac{1}{g}(1 - e^{-gt})I,$$

if I is a (constant) injected current.

Covariance:

The covariance matrix Λ for $V(t)$ is given by the outer product of E_g with itself (this is true for any linear operator applied to a Gaussian random variable): $\Lambda = E_g E_g^T$. Written as a matrix, this gives:

$$\Lambda = \begin{bmatrix} 1 & \alpha & \alpha^2 & \dots & \alpha^n \\ \alpha & 1+\alpha^2 & \alpha(1+\alpha^2) & \dots & \alpha^{n-1}(1+\alpha^2) \\ \alpha^2 & \alpha(1+\alpha^2) & 1+\alpha^2+\alpha^4 & & \\ \vdots & \vdots & & \ddots & \\ \alpha^n & \alpha^{n-1}(1+\alpha^2) & & & 1+\ldots+\alpha^{2n} \end{bmatrix}. \tag{3.25}$$

The nth term along the diagonal is

$$\Lambda(n,n) = \sum_{j=0}^{n} \alpha^j \;\; = \;\; \frac{1-\alpha^{2n}}{1-\alpha^2}$$

and off-diagonal terms $\Lambda(i,j) = \Lambda(i,i)\alpha^{j-i}$, for $i < j$. We can also express Λ analytically in continuous time. Diagonal terms are given by:

$$\Lambda(t,t) = \int_0^t e^{-2gs} ds = \tfrac{1}{2g}(1 - e^{-2gt})$$

and off-diagonal terms by $\Lambda(t,t') = e^{-g(t'-t)}\Lambda(t,t)$, for $t < t'$.

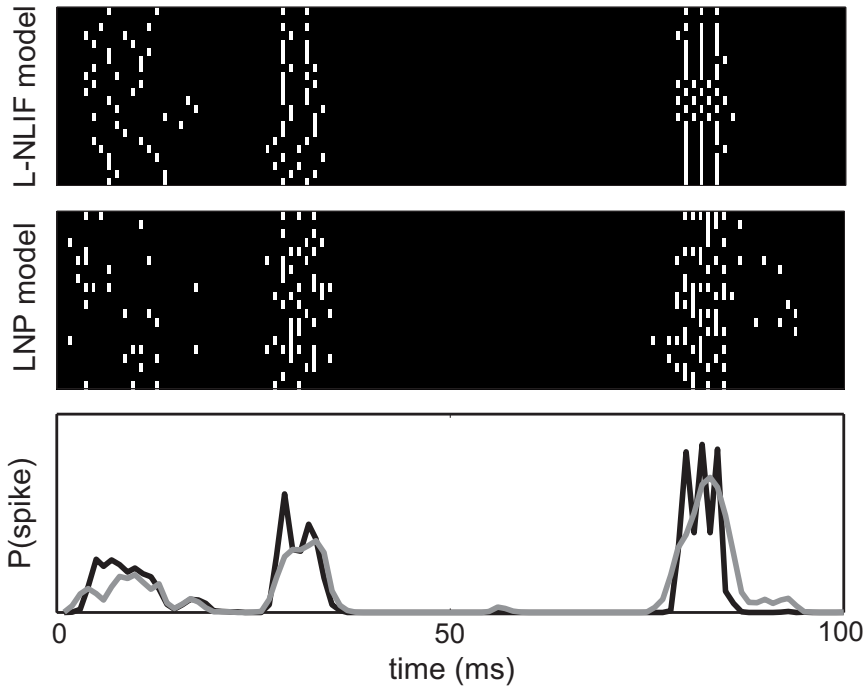

Figure 3.2: Simulated responses of L-NLIF and LNP models to 20 repetitions of a fixed 100-ms stimulus segment of temporal white noise. **Top:** Raster of responses of L-NLIF model, where $\sigma_{noise}/\sigma_{signal} = 0.5$ and g gives a membrane time constant of 15 ms. The top row shows the fixed (deterministic) response of the model with σ_{noise} set to zero. **Middle:** Raster of responses of LNP model, with parameters fit with standard methods from a long run of the L-NLIF model responses to non-repeating stimuli. **Bottom:** (Black line) Post-stimulus time histogram (PSTH) of the simulated L-NLIF response. (Gray line) PSTH of the LNP model. Note that the LNP model fails to preserve the fine temporal structure of the spike trains, relative to the L-NLIF model.

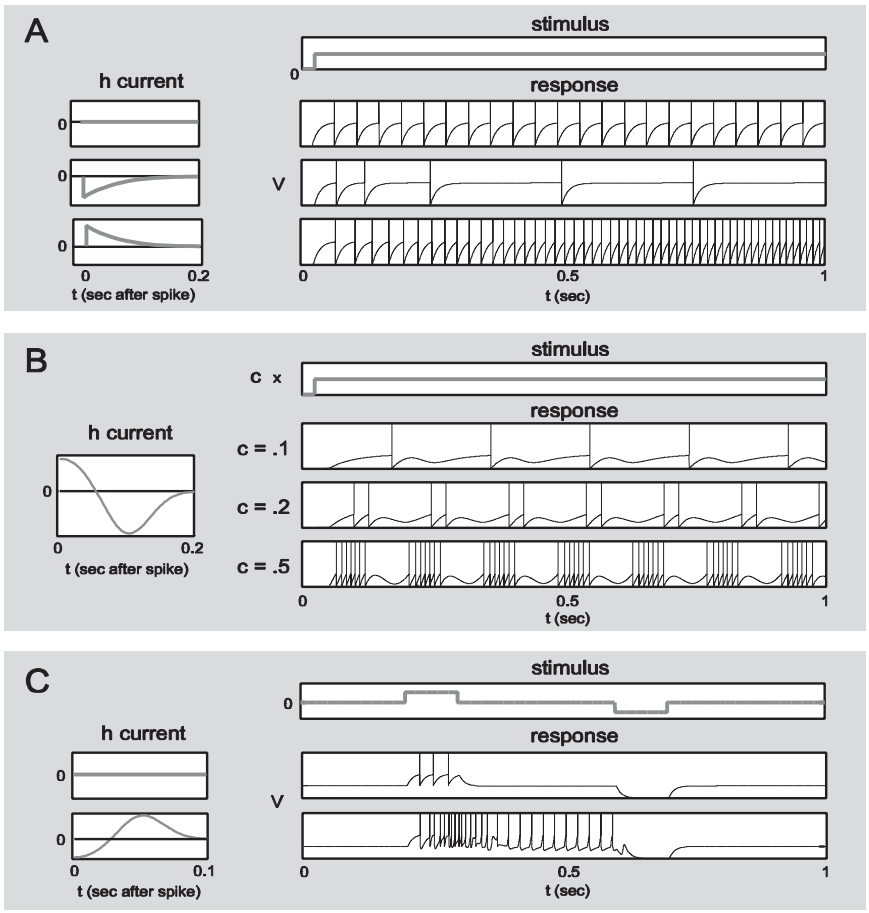

Figure 3.3: Illustration of various dynamic behaviors of L-NLIF model. **A:** Firing rate adaptation. A positive DC current (top) was injected into three model cells differing only in their h currents (shown on left: top, $h = 0$; middle, h depolarizing; bottom, h hyperpolarizing). Voltage traces of each cell's response (right, with spikes superimposed) exhibit rate facilitation for depolarizing h (middle), and rate adaptation for hyperpolarizing h (bottom). **B:** Bursting. The response of a model cell with a biphasic h current (left) is shown as a function of the three different levels of DC current. For small current levels (top), the cell responds rhythmically. For larger currents (middle and bottom), the cell responds with regular bursts of spikes. **C:** Bistability. The stimulus (top) is a positive followed by a negative current pulse. Although a cell with no h current (middle) responds transiently to the positive pulse, a cell with biphasic h (bottom) exhibits a bistable response: the positive pulse puts it into a stable firing regime which persists until the arrival of a negative pulse.

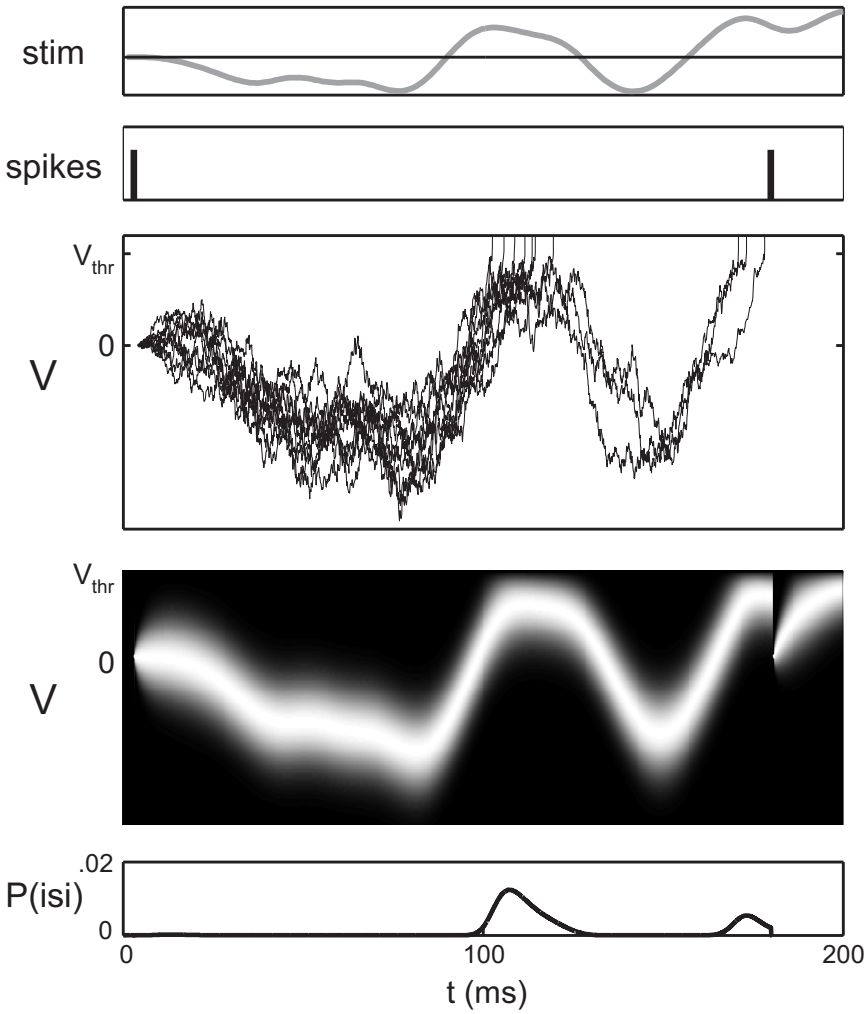

Figure 3.4: Analysis of a single interspike interval under the L-NLIF model, for a single (repeated) input current (top). **Top:** Ten simulated voltage traces $V(t)$, evaluated up to the first threshold crossing, conditional on a spike at time zero ($V_r = 0$). Note the strong correlation between neighboring time points, and the sparsening of the plot as traces are eliminated by spiking. **Middle:** Time evolution of $P(V)$. Each column represents the conditional distribution of V at the corresponding time (i.e. for all traces that have not yet crossed threshold). **Bottom:** Probability density of the interspike interval (isi) corresponding to this particular input. Note that probability mass is concentrated at the points where input drives $V_0(t)$ close to threshold.

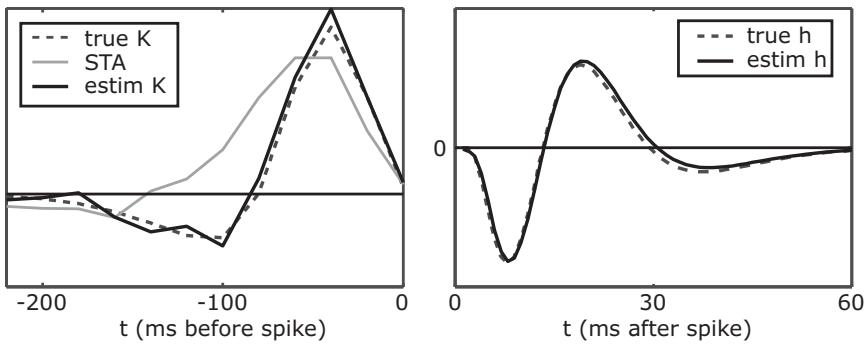

Figure 3.5: Demonstration of the estimator's performance on simulated data. Dashed lines show the true kernel $\vec{k}$ and aftercurrent h; $\vec{k}$ is a 12-sample function chosen to resemble the biphasic temporal impulse response of a macaque retinal ganglion cell, while h is function specified in a five-dimensional vector space, whose shape induces a slight degree of burstiness in the model's spike responses. The L-NLIF model was stimulated with parameters $g = 0.05$ (corresponding to a membrane time constant of 20 time-samples), $\sigma_{noise} = 0.5$, and $V_r = 0$. The stimulus was 30,000 time samples of white Gaussian noise with a standard deviation of 0.5. With only 600 spikes of output, the estimator is able to retrieve an estimate of $\vec{k}$ (gray curve) which closely matches the true kernel. Note that the spike-triggered average (black curve), which is an unbiased estimator for the kernel of an LNP neuron (Chichilnisky, 2001), differs significantly from this true kernel (see also (Pillow & Simoncelli, 2003)).

CHAPTER 4

Prediction and Decoding of Retinal Spike Responses with a Probabilistic Spiking Model

Sensory encoding in spiking neurons depends on both the spatiotemporal integration of sensory inputs and the intrinsic mechanisms governing the dynamics and variability of neural spike responses. Here we show that a generalized integrate-and-fire model can account for the stimulus selectivity, reliability, and timing precision of primate retinal ganglion cell (RGC) spike responses. The model consists of a leaky integrate-and-fire spike generator driven by the linearly filtered stimulus, a spike after-current, and a Gaussian noise current. We used maximum likelihood to fit this model to the extracellular responses of individual RGCs, stimulated with a non-repeating stochastic visual stimulus. We show that the model can

predict the detailed time structure and variability of RGC responses to repeated presentations of novel stimuli, and provides a mechanistic explanation of RGC spike timing precision in terms of stimulus selectivity, cellular noise, and the dynamics of spike generation. Moreover, the tractability of the model for computing likelihoods means that it can be used to perform an explicit decoding of neural spike trains, and provides a tool for assessing the limitations imposed by spike timing variability on sensory performance.

Introduction

Sensory experience depends on the encoding of external events in the spiking activity of neurons. In the visual system, a large number of experimental studies of neural encoding have motivated the formulation of a variety of models. The simplest and most widely known models are receptive field descriptions, which provide a summary of stimulus selectivity (e.g. (Kuffler, 1953; Hubel & Wiesel, 1968; Movshon & Newsome, 1996)). Other approaches have emphasized the intrinsic variability of spike responses and have sought to model the statistical structure of neuronal spike responses (e.g. (Reich et al., 1997; Troy & Lee, 1994; Berry et al., 1997)). Still others have been constructed from detailed descriptions of the cellular and biophysical mechanisms that underlie neuronal function (e.g. (Sterling, 1983; Smith & Sterling, 1990)).

However, many problems in sensory coding can only be addressed in a framework that integrates stimulus selectivity, response variability, and mechanistic interpretation. For example, studies of timing precision in retinal spike trains have emphasized the importance of spike timing for neural coding and information transmission, but have failed to explain the observed precision with a mechanistic model of stimulus selectivity and spike generation. Similarly, studies of visual sensitivity and performance have emphasized stimulus selectivity and response variability, but a plausible mechanistic interpretation is also required in order to guide future experiments at the level of circuits, cells, and channels.

Recent work has revealed that integrate-and-fire (IF) models, which provide a parsimonious mechanistic description of the conversion of continuous membrane currents into discrete spike trains, are capable of exhibiting some of the important statistical spiking behaviors of real neurons (Reich et al., 1998; Shadlen & Newsome, 1998; Jolivet, Lewis, & Gerstner, 2003; Keat et al., 2001). Here we show that a generalized IF model, driven by a linear filtering of the stimulus and a spike-dependent aftercurrent plus noise, can account for both the stimulus dependence and variability of light responses in retinal ganglion cells (RGCs) of the macaque monkey. The resulting model faithfully reproduces the detailed structure of spike trains elicited by novel stimuli. It also captures the trial-to-trial variability of responses, despite the fact that its parameters were fit with responses to a single non-repeating stimulus.

Although the model bears similarity to several recently proposed models of neural function ((Keat et al., 2001; Jolivet, Lewis, & Gerstner, 2004)), the prob-

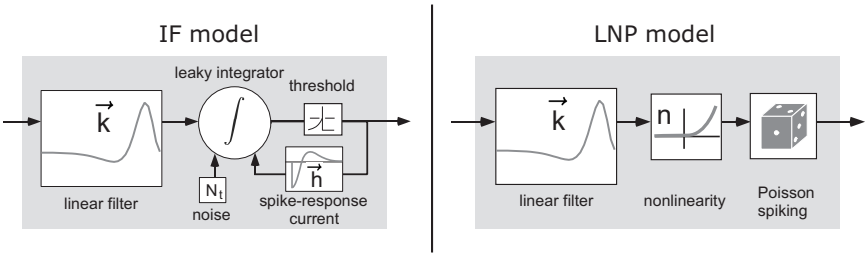

Figure 4.1: Schematic diagrams of the generalized integrate-and-fire model (left), and the standard linear-nonlinear-Poisson model (right).

abilistic formulation, along with an algorithm for computing response likelihood, provides two new insights into neural coding. First, the model provides a simple mechanistic interpretation of the origin of spike timing precision recently observed in RGCs, as well as a more principled and complete framework for describing precision than has been previously available. Second, the model provides an optimal (maximum likelihood) decoding rule for extracting stimulus information from spike trains, and our results demonstrate that this method is capable of extracting information from spike trains more faithfully than a generic linear filtering model. These findings provide a powerful tool for probing signaling by sensory neurons, place bounds on the fidelity of spike train decoding in the brain, and yield specific predictions about the limits on visual performance imposed by cellular noise and spike generation.

Results

Figure 4.1a illustrates the components of the generalized IF model. It is a standard leaky integrate-and-fire model driven by three time-varying input currents:

a stimulus-dependent current I_{stim}, a spike-history dependent current I_{sp}, and a noise current I_{nse}. I_{stim} is the linear convolution of the stimulus with input filter $\vec{k}$, which represents the neuron's spatio-temporal receptive field. I_{sp} results from a current waveform $\vec{h}$ injected following each spike, and captures the influence of spike train history. This is equivalent to convolving $\vec{h}$ with the spike train, so $\vec{h}$ can also be thought of as a linear filter operating on the spike train. Note that $\vec{h}$ can assume an arbitrary shape, and can therefore generate a diverse array of behaviors observed in real neurons, including refractoriness, spike rate adaptation, spike rate facilitation, bursting and bistability (Jolivet et al., 2004; Paninski et al., 2004). This flexibility endows the model with more biological realism than the classic integrate-and-fire model or Poisson models. I_{nse} consists of Gaussian white noise with standard deviation σ_n and represents the net contribution of all noise sources to the membrane potential. The last two parameters of the model are τ, the membrane time constant, and V_l, the reversal potential of the leak current. Without loss of generality, the spike threshold voltage is set to 1 and the reset voltage to 0. The model dynamics are then given by

$$\frac{dV}{dt} = -\frac{1}{\tau}(V - V_l) + I_{stim} + I_{sp} + I_{nse}, \tag{4.1}$$

When $V = 1$, a spike occurs and V is reset instantaneously to 0. The full model is specified by the parameters $\{\vec{k}, g, V_r, \sigma\}$. Characterizing a neuron's response requires the estimation of these parameters from a sequence of stimuli and the times of elicited spikes. For this, we use a recently developed algorithm for computing the maximum likelihood estimator of these parameters, which has guaranteed global convergence for any stimulus and spike train data (Paninski

et al., 2004) (see Methods).

Model validation

Responses of parasol (magnocellular-projecting) RGCs were collected using multi-electrode extracellular recordings from isolated macaque monkey retinas. Stimuli were spatially uniform, achromatic binary temporal white noise sequences (random flicker). For these stimuli, $\vec{k}$ represents the temporal receptive field, while in general $\vec{k}$ can represent the neuron's spatio-temporal-chromatic receptive field. The IF model parameters were fit to spike responses from a single (non-repeating) stimulus train. Subsequently, RGC responses to multiple repeats of a novel stimulus were recorded, and predictions generated by the IF model were assessed using several quantitative measures.

Figure 4.2a-c shows model parameters fit to data from a collection of ON and OFF RGCs recorded simultaneously. Note that the filters, $\vec{k}$, are consistent in waveform and timescale within each cell type. The same is true of the spike current waveforms, $\vec{h}$, which operate on a faster timescale. The biphasic shape of the $\vec{h}$ currents allows the model to reproduce burstiness in RGC responses—the initial positive component drives voltage up close to threshold following voltage reset, and the later negative component (possibly accumulated over several spikes) exerts a hyperpolarizing effect to end a burst. The larger amplitude $\vec{h}$ in ON cells matches the burstier responses observed in ON compared to OFF cells (see Figs. 4.3-4.4). Figure 4.2c also shows histograms of the three scalar parameters $\{\sigma_n, \tau, V_l\}$ for all cells.

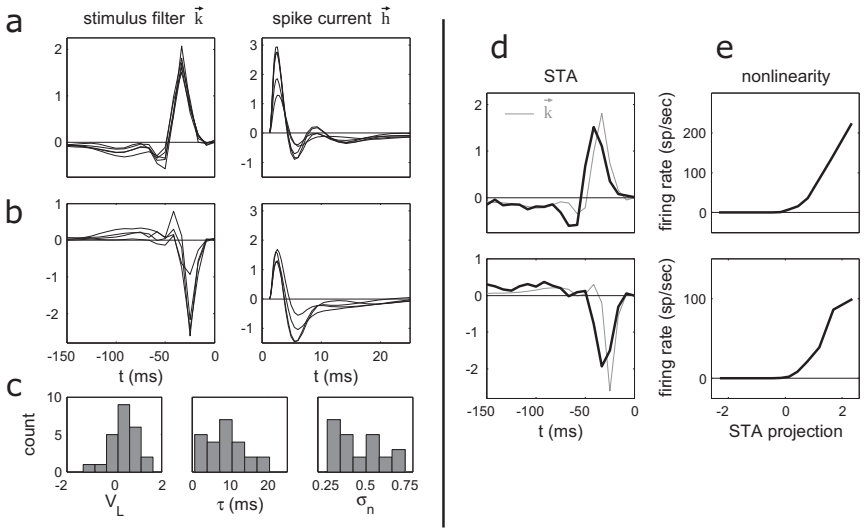

Figure 4.2: Parameters obtained from fits to RGC data for IF model (left) and LNP model (right). **(a)** Filters $\vec{k}$ and spike-response currents $\vec{h}$ obtained for five ON cells in one retina. **(b)** Corresponding filters for four OFF cells. **(c)** Histograms of model parameters V_l, τ and σ_n for all 24 cells in 3 retinas. **(d)** Comparison of linear filters for IF model (gray) and LNP models (black), for one ON cell (above) and one OFF cell (below). **(e)** Measured LNP point nonlinearities for converting filter output to instantaneous spike rate.

To provide a baseline for comparison, the same data were fit using the simplest and most widely-used cascade model of visual responses: a Linear-Nonlinear-Poisson (LNP) model, shown schematically in Fig. 4.1. In the LNP model, a single linear filter captures the neuron's stimulus dependence. The output of this filter passes through an instantaneous nonlinearity, which determines the rate of an inhomogeneous Poisson process that generates spikes. Although Poisson processes cannot exhibit refractoriness, bursting, or other known statistical features of spike trains, the LNP model is widely used because of its computational simplicity and the ease with which its parameters can be estimated using reverse correlation with white noise stimuli(Chichilnisky, 2001). Figure 4.2d shows linear filters obtained for both the LNP model and the IF model, for one ON cell and one OFF cell. Note that the linear filter recovered for the LNP model (i.e. the spike-triggered average) is noticeably different than that obtained for the IF model (Figure 4.2d). Given that the IF model incorporates more realistic spike generation and provides more accurate predictions of real spike trains (see below), this suggests that the LNP model provides an inaccurate description of how neurons integrate visual inputs over time (Berry & Meister, 1998; Pillow & Simoncelli, 2003; Arcas & Fairhall, 2003). Also, note that the nonlinear function of the LNP model (Figure 4.2e) has no direct counterpart in the IF model, in which firing rate is determined implicitly by the dynamics of the leaky integrator and spike threshold.

Both models provide predictions about the mean and variability of responses to any stimulus. We examined the accuracy of these predictions using repeated

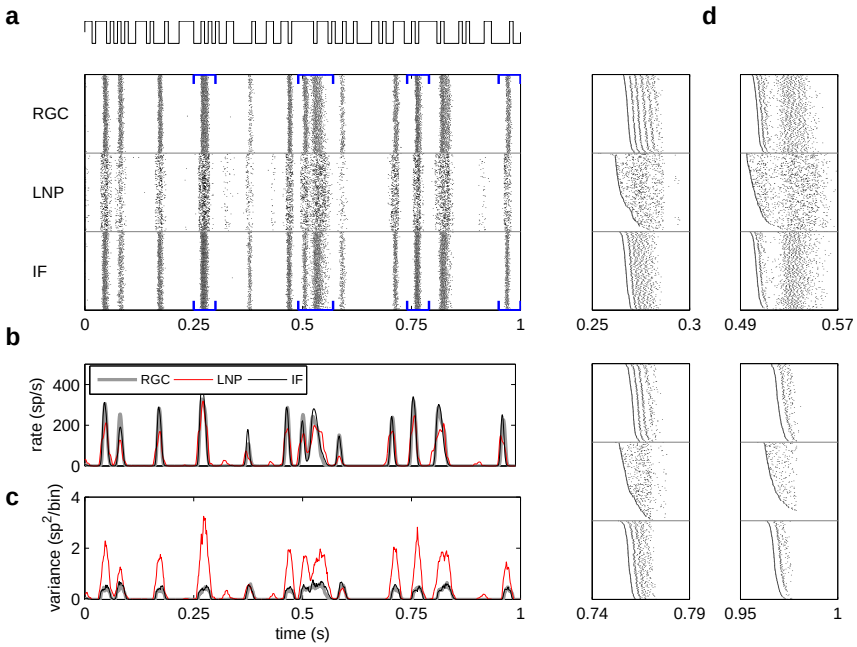

Figure 4.3: Responses of an ON cell to a repeated stimulus. **(a)** Recorded responses to repeated one-second stimulus (top), simulated LNP (middle) and IF model (bottom) spike trains. Each row corresponds to the response during a single stimulus repeat; 167 repeats are shown. **(b)** Peristimulus time histogram (PSTH), or mean spike rate, for the RGC, LNP model and IF model. For this cell, the IF model accounts for 91% of the variance of the PSTH, while the LNP model accounts for 75%. **(c)**: spike count variance computed in a sliding 10-ms window. **(d)**: Magnified sections of rasters, with rows sorted in order of first spike time within the window. The four sections shown are indicated with blue brackets in **a**.

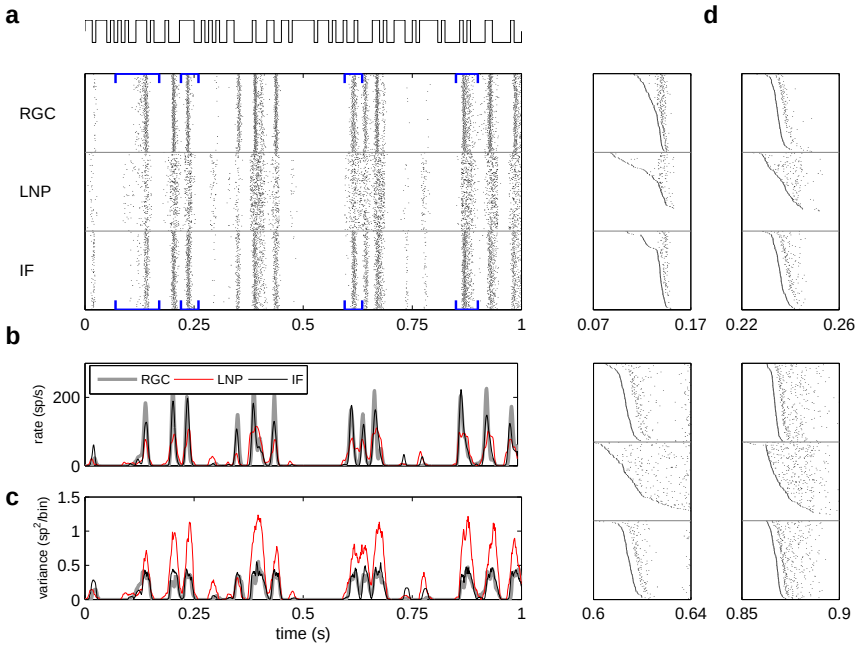

Figure 4.4: Responses of an OFF cell to a repeated stimulus. Details same as Fig. 4.3.

presentations of a novel stimulus sequence. Figures 4.3 and 4.4 illustrate recorded responses and predictions of the IF model and the LNP model, for an ON and an OFF cell, respectively. For both models, parameters were fit using responses to a single 50-second white noise stimulus sequence. Rasters of RGC responses and corresponding simulated responses from both models illustrate that the IF model (bottom rows) captures the structure of the RGC spike trains (top rows) more faithfully than the LNP model (middle rows).

The peristimulus time histogram (PSTH, Figs. 4.3b & 4.4b) summarizes the time-varying firing rate exhibited by the data and both models. The IF model (black trace) matches the sharp peaks in the PSTH more accurately than the LNP model (red trace). Trial-to-trial variability of the responses is reflected by the peristimulus time variance (PSTV, Figs. 4.3c & 4.4c), computed by sliding a 10-ms window along the response raster and computing the variance across trials of the number of spikes in that window. Because RGC spike trains have history-dependence which makes them much less variable than a Poisson process (Uzzell & Chichilnisky, 2004), it is unsurprising that the LNP model fails to match the PSTV of the data. The IF model provides a more accurate prediction. Although integrate-and-fire models have been shown previously to reproduce the spike count variability in neurons (Reich et al., 1998), it is notable that the IF model does so despite a fitting procedure (maximum likelihood) that does not include a measure of variability and does not require repeated stimuli.

A more detailed view of spike train structure and model performance was obtained by sorting the rows of a response raster in order of the first spike time

in a given window (Figs. 4.3d and 4.4d). Sorting reveals considerable structure in RGC interspike intervals, which is largely captured in the sorted responses of the IF model (bottom) but is completely absent in the sorted responses of the LNP model (middle).

Summary statistics of IF model performance and comparison to the LNP model are shown in Fig. 4.5, for all RGCs examined. Figure 4.5a shows a comparison of the likelihood of responses to novel stimuli for the IF and LNP models, obtained by using the fitted parameters for each model to compute the probability of the observed responses. This provides the most direct and powerful statistical test of performance, because it measures, in a probabilistic setting, the ability of each model to predict the response to novel stimuli. Using this metric, the IF model provided a significantly higher likelihood per spike for all cells, in many cases nearly two-fold. Figures 4.5b and c show comparisons of the similarity of PSTH and PSTV obtained from RGC spike trains and model simulations. In both cases, the IF model outperforms the LNP model for all cells.

To compare the accuracy of model predictions to the intrinsic variability in RGC spike trains, we also applied a previously-used summary measure of distances between spike trains (J. D. Victor & Purpura, 1997). This distance is the minimum cost of transforming one spike train into another using the elementary operations of adding, deleting, and shifting spikes. A timescale parameter expresses the cost of shifting per unit time relative to that of adding or deleting. Although this is an imperfect measure of model performance because it neglects any effect of stimulus or spike train history on the probabilistic cost of shifting

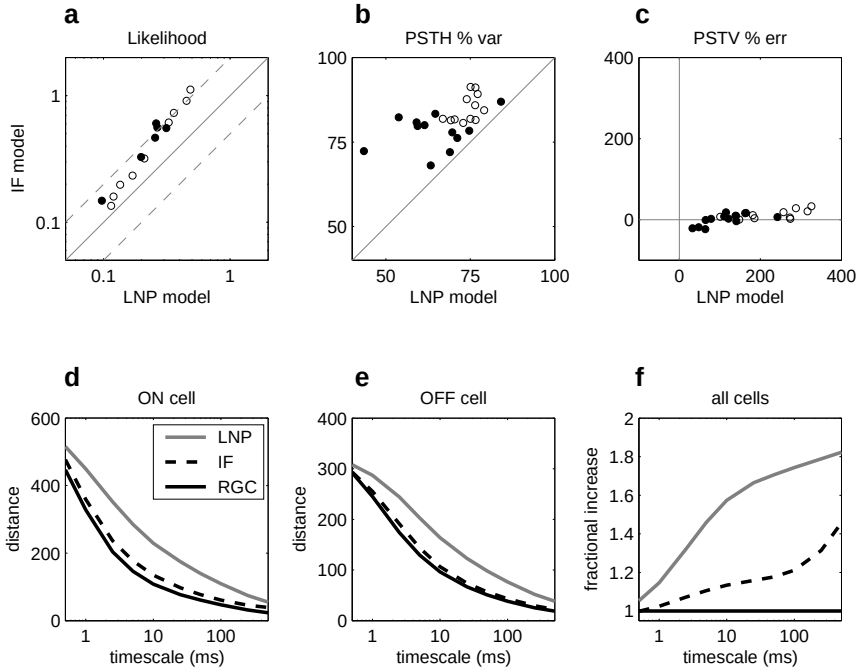

Figure 4.5: Performance comparison across cells. Empty and filled circles represent ON and OFF cells, respectively. **(a)** likelihood per spike of novel RGC responses under the fitted IF model and LNP model. Value plotted is the geometric mean of the likelihood of over all spikes under each model. Gray dashed lines represent a factor of 2 above and below identity. Data from 24 cells (3 retinas) are shown. **(b)** percent of the variance in the PSTH accounted for by both models, for each cell. Points above the diagonal represent superior performance by IF model. **(c)**: percent error in the peristimulus time variance for the IF and LNP models, across cells. **(d)** Average pairwise distance between spike trains, as a function of timescale of analysis (see Results). Black trace shows the median distance between responses of an ON RGC to repeated presentations of the same stimulus. Dashed trace shows the median distance between IF model response and data. Gray trace shows median distance between LNP model response and data. **(e)** Same as D, for an OFF RGC. **(f)** Fractional increase in spike-time distance for the IF and LNP models averaged over all cells. Dashed curve: ratio of IF model distance to the average pairwise distance between RGC responses, normalized by the number of spikes from each cell and averaged across cells. Gray curve: same, for LNP model.

spikes (e.g. it would not penalize a model for ignoring the refractory period), it is easy to compute and provides a direct benchmark for comparing the performance of present and future models.

The solid curves in Fig. 4.5d-f show the average distance between pairs of RGC responses to repeated presentations of the same stimulus. This provides a measure of intrinsic variability in spike trains. The distance falls monotonically as a function of timescale of analysis (Fig. 4.5d,e). Dashed and gray curves indicate distances between simulated IF and LNP model spike trains and recorded RGC spike trains. This provides a measure of the discrepancy between model predictions and data. The IF model distances were systematically higher than the intrinsic variability, indicating that IF model responses differ noticeably from RGC spike trains. However, across a wide range of timescales (1-100 ms) and in all cells, IF model distances were smaller than LNP model distances. Figure 4.5f shows a summary of the IF and LNP model distances expressed as a fraction of the intrinsic variability, averaged over all cells. IF model error exceeds intrinsic variability by up to $\sim 20\%$; LNP model error is roughly three-fold higher for most time scales.

Precision of spike times

The accurate descriptions of RGC spike trains provided by the IF model present an opportunity to examine the origins of spike timing precision, which has been widely discussed in recent studies. RGCs are capable of firing spikes precisely time-locked to the onset of a stimulus (Reich et al., 1998; Berry & Meister, 1998;

Keat et al., 2001; Uzzell & Chichilnisky, 2004). In some cases, the variation in the onset time of spiking across repeated stimulus presentations is as low as ~ 1 ms. Although precise timing during periods of rapid firing may be explained by action potential refractoriness (Berry & Meister, 1998; Uzzell & Chichilnisky, 2004), the origin and significance of the precision in firing onset time is unknown.

A simple hypothesis (see, e.g., (Bryant & Segundo, 1976; Cecchi, Sigman, Alonso, & Martinez, 2000; Uzzell & Chichilnisky, 2004)) is that more precise firing onset results from more rapid threshold crossing by the membrane voltage. If voltage crosses threshold with a steep slope, then the noise current exerts little influence on the time of the spike; if voltage crosses threshold with a shallow slope, noise has a greater influence. Fig. 4.6 illustrates this intuition graphically with an example from an RGC response raster. The raster shows two adjacent periods of rapid firing. The histograms below show the distributions of the time of the first spike in each period of firing, and indicate that the first period of firing exhibited more precise timing. The trace below shows the (noiseless) voltage response of the IF model, obtained using the stimulus and the parameters fit for this cell. The slope of V at threshold crossing is indicated by red lines. The period of firing that begins with a steeper voltage slope exhibits much more precise timing. Qualitatively, this example supports the simple hypothesis.

As a more thorough and quantitative test, an analysis of precision was performed for all identified firing onsets during the 7-second response raster for this cell. Firing onsets were defined after periods of silence at least 8 ms long across all trials, followed by a spike on at least 80% of trials within a window of 40 ms

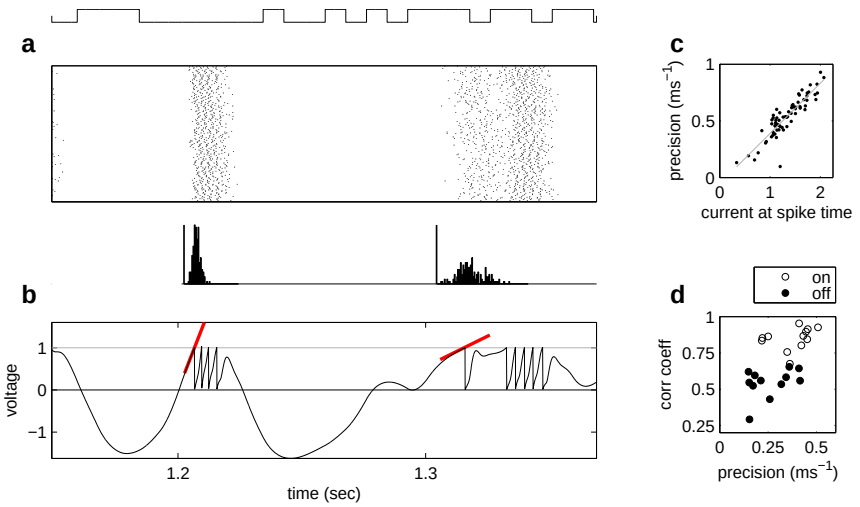

Figure 4.6: Precision of firing onset times. **(a)** 200 ms portion of stimulus and a response raster showing two periods of firing onset. Below are histograms of the time of the first spike in each event. Standard deviations: 1.3 ms (left) and 4.7 ms (right). **(b)** simulated voltage response from the fitted IF model with noise set to zero. Tangent at time of firing onset is shown in red. **(c)** Precision of first spike times (inverse of standard deviation) as a function of the mean current at the time of the first spike, over 70 isolated firing onsets. Correlation coefficient: 0.89. **(d)** Correlation coefficient between precision and IF model current prediction, as a function of the inverse of the average SD of the first spike time. Open circles denote ON cells and filled circles denote OFF cells.

(see (Uzzell & Chichilnisky, 2004)). The standard deviation of the first spike time during each onset was computed, and compared to the average current produced by the IF model at those times. Figure 4.6c shows a scatter plot of precision (inverse of standard deviation) as a function of model input current (which is proportional to the slope dV/dt crossing threshold) across 70 such firing onsets. The model accounted for 89% of the variability in precision for this cell. Note that total membrane current (plotted on the abscissa) exactly determines the membrane voltage slope at spike times (equation 4.1).

Figure 4.6d shows correlation coefficients between IF model current and first spike precision for all cells tested. A significant proportion of the variability in precision across time for each cell was explained by the membrane voltage slope at the onset of firing. Note that the IF model captures more of the variability in precision for ON cells (open circles) than for OFF cells (filled circles). However, for both ON and OFF cells, spike timing precision is not an intrinsic property of the cell: it varies substantially as a function of the stimulus history, and in a manner predicted by elicited currents and the intrinsic noise in the IF model.

One shortcoming of this analysis is that the notion of precision relies on an arbitrary criterion for identifying firing onsets, assumes a univariate measure of precision (standard deviation), and ignores all spikes beyond the first. One indication that this approach is incomplete is that subsequent spikes sometimes exhibit more precise timing than the first spike. Figure 4.7a shows a period of firing in which the second spike is more precise than the first, a behavior seen in more than one third of the firing onsets recorded from this cell. Figure 4.7b

shows a more unusual example in which the last spike is more precise than the first. Thus, restricting the analysis to the first spike fails to capture important aspects of spike train precision. Furthermore, measures of spike precision are bound up inextricably with the notion of spiking reliability. Attempts to separate precision (jitter in spike time) from reliability (probability of spike occurrence in a particular time interval) require the use of *ad hoc* criteria for defining firing events: the length of silence preceding an event and the fraction of repeats containing a spike necessary to constitute an event. In our data set, changes in these criteria result in the identification of quite different numbers of firing onsets. Given these difficulties, it is natural to ask whether the IF model can provide a mechanistic explanation which accounts for the precision and reliability of all recorded spikes.

Such an account emerges naturally from the likelihood function used in fitting the IF model. The machinery for computing the likelihood of spike trains can be used to compute the probability density for a particular spike time, conditional on the stimulus and spike train history, as shown in Fig. 4.7c-d. Using the linear kernel $\vec{k}$ and after-current waveform $\vec{h}$, we can compute the predicted intracellular current ($I_{stim} + I_{sp}$) during the interval. This input, combined with the noise current I_{nse}, determines a probability density over subthreshold voltage $P(V)$ as a function of time, and can be used to compute the probability density of the next spike $P(\text{next spike})$ as a function of time.

Of course, the probability density of the next spike time depends on the particular spike history for that trial, and therefore differs slightly on each trial. For direct comparison to experiment, the next-spike density may be averaged across

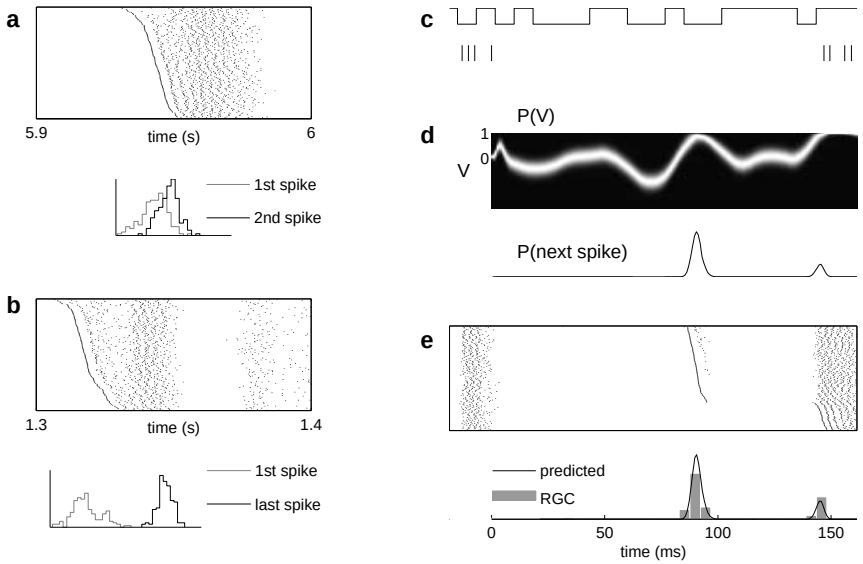

Figure 4.7: Generalized analysis of timing precision. **(a)** RGC response raster sorted in order of first spike time. Below: histogram of first spike (gray) and second spike (black) in the event, illustrating higher precision in the time of the second spike than the first. **(b)** Sorted RGC response raster, with histograms below showing the distribution of the first (gray) and *last* (black) spike in this event. Precision of the last spike is higher than the first. **Right:** Use of the IF model formalism to analyze RGC spike timing precision and reliability. **(c)** 170 msec stimulus fragment and corresponding RGC spike response during one trial. **(d)** Probability distribution over subthreshold V for central interspike interval, on a single trial. The likelihood of the next spike time (below) is given by the probability mass crossing threshold at each moment in time. Note that the probability distribution of the next spike time is bimodal. **(e)** Raster of repeated RGC responses to this stimulus fragment, with rows sorted in order of first spike time. Below: probability density of the next spike, averaged across 25 trials. Black trace shows model prediction, Gray bars show actual distribution.

trials and compared to the observed next spike time distribution. Figure 4.7e shows an example where the distribution of the next spike time following a period of silence is bimodal, a condition where standard measures of spike time precision are problematic. This bimodality is accurately reflected in the density of the next spike time computed using the IF model. Thus the next spike time density provides a more complete description of timing variability than *ad hoc* summary measures of precision and reliability.

Decoding of spike responses

Perhaps the most important role for a model of RGC responses is to provide a precise description of the visual information transmitted to the brain. The generalized IF model makes it possible to assess the degree to which variability in spike trains imposes limitations on the fidelity of information transmission. Specifically, the model can be used to compute the probability that an observed spike train was elicited by any given stimulus. This provides a powerful method for decoding the information contained in neural spike responses.

One method to illustrate decoding would be to simply use the model to perform stimulus reconstruction from measured spike responses. Given a particular spike response x, we can obtain the probability that it was elicited by a stimulus y via Bayes' rule:

$$P(y|x) = P(x|y)P(y)/P(x), \tag{4.2}$$

where $P(x|y)$ is obtained from the model likelihood calculation, $P(y)$ is the prior over the stimulus, and $P(x)$ is a normalizing term, the marginal probability

of observing response x. We can perform stimulus reconstruction by choosing the y which is the maximum or the mean of $P(y|x)$. The maximum provides the maximum *a posteriori* (MAP) estimate for the stimulus, while the mean provides the Bayes estimate under a squared loss function.

Rather than performing a full stimulus reconstruction, which requires choosing a prior $P(y)$ and a search for the y which achieves the maximum or mean of eq. 4.2, we performed a simple illustration of decoding by using the model to discriminate stimuli in a 2-alternative forced-choice experiment. In this experiment, an observer is presented with two spike trains and two different stimuli, and must decide on the correct pairing of stimuli with elicited responses (Fig. 4.8a). Clearly, the optimal decision rule given a particular model of the response is to use the pairing with the higher likelihood under that model. (Green & Swets, 1966). We used the likelihood of spike responses under the IF model to discriminate pairs of stimuli, and by applying this decision rule to each pair of responses obtained using multiple repeats of the stimuli, we obtained a percent correct for the performance of the model in discriminating a particular pair of stimuli. As a benchmark, we compared this procedure against one in which likelihood was computed (and discrimination performed) using the LNP model. Figure 4.8b shows a comparison of the performance of the two procedures. The predominance of data above the diagonal indicates that, on average, the IF model provided significantly more accurate discrimination than the LNP model. Thus, stimulus decoding based on the IF model exploits information in temporal patterns of spikes that is not captured by decoding based only on firing rate.

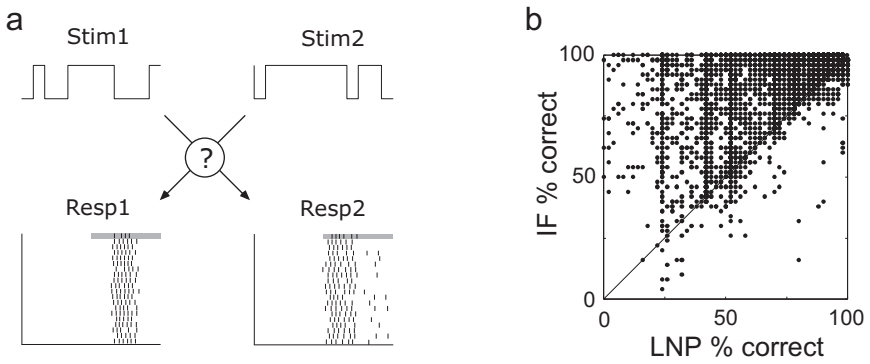

Figure 4.8: Decoding responses using model-derived likelihoods. **(a)** Two stimulus fragments and corresponding fragments of RGC response raster. Gray boxes highlight a 50-ms interval of the first row of each response raster. A two-alternative forced-choice (2AFC) discrimination task was performed on these response fragments, where the task was to determine which stimulus gave rise to each response. The IF and LNP models were used to compute the likelihood of these responses given the "correct" and "incorrect" pairing of stimuli and responses, and the pairing with higher likelihood was selected. This discrimination procedure was applied to each row of the response raster and used to obtain a percent correct for the discrimination performance of each model. **(b)** Discrimination performance of the IF and LNP models. Each point corresponds to the percent correct of a 2AFC discrimination task using two randomly-selected 50-ms windows of the response. Although both models obtain perfect performance for a majority of such randomly-selected response windows, the scatter of points above the diagonal shows that when discrimination performance is imperfect, the IF model is far better at decoding the neural spike responses.

Discussion

An integrate-and-fire model, generalized to include a linear receptive field, a spike after-current and a noise source, provides an accurate description of the detailed structure of RGC spike trains elicited by visual stimuli. The model relies only on measurements of spike times, but provides a full description of the hypothesized underlying currents, along with detailed, accurate predictions of spiking behavior. The model can be fit reliably to responses from arbitrary stimuli using a straightforward procedure, and does not require long measurements of responses to repeated or specialized stimuli. Most importantly, the model provides mechanistic insights into the origins of spike timing precision, as well as an optimal decoding procedure that exploits temporal patterns in spike trains.

Variability, structure, and fidelity of retinal signals

In previous work, the timing precision of firing onsets and the large gaps between periods of firing have been suggested as the basis for a novel interpretation of retinal coding (Berry et al., 1997; Keat et al., 2001). In this view, the retina encodes visual information in discrete firing events whose timing and spike count encode the timing and features of the stimulus, respectively. The IF model provides a more mechanistic interpretation of spike train structure. Spiking precision at firing onset results primarily from stimulation that causes the membrane potential to cross threshold rapidly, leaving little opportunity for current fluctuations to influence the time of the spike (Fig 4.6) (Bryant & Segundo, 1976; Cecchi et al., 2000). Similarly, long gaps between periods of firing result from stimuli that

effectively suppress spiking by providing strong currents of opposite polarity. Periods of more gradual firing rate modulation, which do not conform easily to the theory of "firing events", arise from stimuli poorly matched to the linear filter, or following periods of maintained firing. Thus, the encoding process is fundamentally linear, and the precision and structure of spike trains simply reflect the interaction of the stimulus with intrinsic filtering, noise, and spike generation.

In the model, a single current noise parameter accounts for all the variability in neural response, and thus summarizes the effects of noise in transduction, synaptic transfer, and cellular integration. This summary measure of response variability could serve as a parsimonious replacement for previously proposed measures such as spike time precision and reliability (Keat et al., 2001; Berry et al., 1997; Uzzell & Chichilnisky, 2004; Reinagel & Reid, 2000). Measures of spike time precision rely *ad hoc* criteria for identifying of firing onsets, and may exclude many spikes from analysis. Measures of response reliability require restrictions on the time window of spike occurrence. These measures do not reveal intrinsic properties of the cell because they vary greatly with the stimulus (Berry et al., 1997; Uzzell & Chichilnisky, 2004) and have no mechanistic interpretation. In contrast, the noise parameter of the IF model, measured with a single stimulus sequence, explains many different measures of response variability (Fig 4.7), has a rough physiological interpretation in terms of membrane current (Fig 4.1), and can be used to predict the variability to novel stimuli.

The IF model also provides a potentially significant technique for assessing the fidelity of sensory signals (Fig 4.8). Because the probability of an observed

spike train given any stimulus can be computed directly, the optimal stimulus discrimination procedure is known. Such an optimal procedure is essential for a meaningful investigation of the factors that limit sensory performance. Stimulus discrimination based on the IF model was significantly more accurate than discrimination based on the LNP model, indicating that temporal patterns of spikes convey information not captured in the time-varying firing rate. Stimulus decoding based on the IF model also provides a bound on the accuracy with which the brain could decode stimulus information from RGCs. This approach to describing the fidelity of neural coding in the context of a mechanistic model could provide a valuable complement to more abstract approaches based on information theory (Bialek et al., 1991).

Limitations of the IF model

Although the IF model offers many practical and theoretical advantages, it has some shortcomings. First, although it is clearly outperforms the LNP model by a variety of measures, the IF model still fails to account for ∼10-25% of the variance in the PSTH of RGCs (Fig. 4.5b), and deviations from the observed data exceed the variability of repeated responses by up to ∼20% (Fig. 4.5). One likely source of error is the existence of nonlinear stimulus dependencies (Hochstein & Shapley, 1976; J. D. Victor & Purpura, 1997; Benardete et al., 1992) not captured by the linear front end. Nonlinearities may be incorporated into the model without compromising its stable fitting properties by expanding the filter $\vec{k}$ to operate on nonlinear functions of the input. In preliminary studies, the inclusion of

terms with quadratic dependence on the input led to a 5-10% improvement in the percentage of explained variance. A drawback is that the elaborated model has more parameters to describe the nonlinear stimulus dependence, and thus requires more data for estimation of those parameters.

Second, the fitting procedure tends to set the noise parameter somewhat higher than is necessary to account for the timing variability observed during periods of rapid firing. This is evident in the sorted rasters of Fig. 4.3, where the repeated interval structure of the RGC response is more regular than that of IF model responses. The discrepancy may result from the fact that the noise parameter accounts for both true variability in neural responses and any systematic errors in the model. Specifically, nonlinear mechanisms not captured by the model require an increase in the noise parameter to keep the observed spike train likelihood from becoming prohibitively small. This shortcoming could be addressed by separately optimizing the noise parameter using an objective function that isolates the stochastic behaviors of the neural response, such as the repeat interval structure (Fig. 4.3) or the spike-time distance (Fig. 4.5). A complete solution, however, requires the incorporation of nonlinear stimulus selectivity in the model (see above). This may be particularly important for applying the IF characterization to neurons in visual cortex.

Third, although the IF model is clearly more realistic than models with Poisson firing, it provides only a simplified description of known mechanisms of neural response. The model is based on currents rather than conductances, has linear subthreshold dynamics, and includes only a single Gaussian white noise source

to account for a variety of real noise sources (e.g. photon noise, synaptic failure, and channel noise). These simplifying assumptions make the fitting procedure and interpretation more tractable.

Extensions

The present findings suggest a variety of extensions. First, several studies (Smirnakis et al., 1997; Chander & Chichilnisky, 2001; Kim & Rieke, 2001; Baccus & Meister, 2002) have revealed slow contrast adaptation in RGCs, and integrate-and-fire models can likewise exhibit a form of spike rate adaptation to stimulus variance (Rudd & Brown, 1997). Our preliminary observations indicate this is true of the IF model as well. Second, in cases where responses are driven by nonlinear stimulus transformations of a known form, the model may be augmented as described above. The model could also be used, as the LNP model has, to characterize the steady-state results of slow nonlinearities such as adaptation, perhaps providing more accurate information about its effects on stimulus selectivity, noise, and spike generation (Chander & Chichilnisky, 2001; Kim & Rieke, 2001; Baccus & Meister, 2002). Third, the fitting method is based on gradient ascent of the likelihood function, and so may be applied to data collected with any sufficiently rich set of stimuli. Thus, unlike reverse correlation approaches to estimating LNP model parameters, IF model characterization does not require white noise stimuli. This raises the possibility of characterizing light responses using stimuli that drive cortical neurons more strongly than white noise, or that more closely approximate the environment in which the visual system normally

operates (Reinagel, 2001).

Methods

Experimental Measurements & Stimuli

The data presented in this paper are a subset of the data in (Uzzell & Chichilnisky, 2004); experimental methods are described in detail there. Briefly, multi-electrode extracellular recordings were obtained *in vitro* from small pieces of retina from 3 macaque monkeys, with retinal pigment epithelium attached, maintained at 32-36 degrees C, pH 7.4. The retina was stimulated with a photopic, achromatic, spatially uniform, optically reduced image of a cathode ray tube display refreshing at 120 Hz. The stimulus was a temporal sequence consisting of two intensity values, pseudo-randomly selected on every refresh. The contrast (standard deviation divided by mean) of the sequence was 96%. Model characterization was performed on one pseudo-random sequence (50 sec), and model validation was performed on a different sequence (7-30 sec) repeated 37-176 times. Analysis was restricted to two physiologically-defined classes of cells that very likely correspond to ON and OFF parasol cells based on several lines of evidence (Chichilnisky & Kalmar, 2002).

Model Parameter Fitting

In previous work, we developed an efficient and computationally tractable algorithm for estimating the parameters of the generalized IF model used in this

paper (Paninski et al., 2004). Specifically, we showed that the likelihood function can be computed efficiently, and is log-concave for any stimulus and spike train data. This guarantees that the IF model, unlike many other models, can be fit reliably using simple gradient ascent techniques, without risk of converging on local maxima. The likelihood function itself, described in detail in (Paninski et al., 2004), was computed by numerically solving the Fokker-Plank equation for subthreshold voltage during each interspike interval. This amounts to finding the probability under the Gaussian noise model that voltage crossed threshold at precisely the observed spike times.

Model parameters $\{\vec{k}, g, V_r, \sigma\}$ were fit by performing gradient ascent of the likelihood function $P(\text{spikes}|\text{stim}, \{\vec{k}, g, V_r, \sigma\})$. $\vec{k}$ and $\vec{h}$ were taken to be 15 and 10 dimensional vectors, respectively, in a vector space with basis vectors of the form $\sin(\log(a * t + b))$, where a and b are scalars. These basis vectors (similar to those in (Keat et al., 2001)) have fine temporal structure near the time of a spike and are smooth at longer timescales, allowing $\vec{k}$ and $\vec{h}$ to be represented with relatively few parameters. The basis for $\vec{k}$ accurately reproduced the shape of the spike-triggered average, taken to be 40 stimulus frames (333 ms) long. Example comparisons with $\vec{k}$ fit in the full 40-dimensional space of the spike-triggered average indicated that the choice of basis did not affect model performance. The basis for $\vec{h}$ spanned a time interval of 200 ms, though (as shown in fig. 4.2a and b), the actual fits of $\vec{h}$ were essentially zero outside of the first 20 ms following a spike). Finally, a robust version of maximum likelihood was used to obtain parameter estimates that were less sensitive to statistical outlier

spikes. The likelihood of the bottom 5% of the interspike intervals was ignored when optimizing the parameters. Empirically, this resulted in lower estimates of noise parameter σ_n, in keeping with the intuition that the model compensates for exceedingly low-probability spikes by artificially increasing the noise parameter (see Discussion). The robust estimate of σ provided an improved match to the first-spike time precision of spike trains.

The LNP model was fit to spiking data using the method described in (Chichilnisky, 2001). First, the linear filter was obtained through reverse correlation of stimulus with spike train. The point nonlinearity was then recovered by examining the firing rate as a function of the linear filter response to the stimulus. The shape of the nonlinearity was fit using maximum likelihood to estimate a linear combination Gaussian radial basis functions, and a similar robust procedure (ignoring the likelihood of the lowest 5% of interspike intervals) was employed to put likelihood comparisons with the IF model on equal footing. Nearly identical performance was obtained when point nonlinearities were instead fit using least-squares estimation of a cubic spline.

Calculation of Percent Error and Distance Between Spike Trains

Figure 4.5b shows the percent of the variance in the PSTH accounted for by each model, or $100 * \left(1 - \langle(\mathrm{PSTH}_{RGC} - \mathrm{PSTH}_{model})^2/(\mathrm{PSTH}_{RGC} - \langle \mathrm{PSTH}_{RGC}\rangle)^2\rangle\right)$, where $\langle\cdot\rangle$ indicates an average over time. PSTHs were computed by binning each response, summing, and filtering with a Gaussian with a standard deviation 1 ms.

Figure 4.5b shows the percent error in the PSTV for each model, which is given by $100 * \langle \text{PSTV}_{RGC} - \text{PSTV}_{model} \rangle / \langle \text{PSTV}_{RGC} \rangle$. PSTV was computed by sliding a 10-ms window across the response raster and calculating the variance (across trials) of the number of spikes in that window.

Figure 4.5d-f examines the similarity between pairs of spike trains to repeated presentations of an identical stimulus, using a specific spike train distance measure (J. D. Victor & Purpura, 1997). The measure relies on a timescale parameter λ and is defined as the minimum cost for bringing one spike train into alignment with another by shifting and adding/deleting spikes, where adding or deleting has a cost of 1, and shifting a spike by t ms entails a cost of t/λ. For any two spike trains, this distance measure is bounded above by the sum of the number spikes in both spike trains ($\lambda = 0$) and bounded below by their spike count difference ($\lambda = \infty$). Software for computing the spike time distance measure was obtained from the author's website(J. D. Victor & Purpura, 1997).

Calculation of Likelihoods for Discrimination Task

Suppose an observer is given spike trains $\{s_A, s_B\}$ in response to presentation of stimulus sequences $\{A, B\}$, and must decide which of the two stimuli is associated with which spike train. The optimal decision rule comes from comparing the likelihood of each stimulus under each observed spike train (Green & Swets, 1966):

$$R(A, B) = \frac{P(s_A|A) \cdot P(s_B|B)}{P(s_A|B) \cdot P(s_b|A)}$$

If $R(A, B)$ is greater than unity, the correct choice is made; otherwise the stimuli are paired with the wrong spike trains. The conditional probabilities in this expression may be determined directly from any stochastic model of neural response, such as the IF or LNP models.

Discussion

We have examined two models of retinal ganglion cell responses, estimated using spike responses to white noise stimuli and evaluated using prediction accuracy of novel responses. Both models have an explicitly probabilistic formulation and can be used to compute the encoding probability distribution $P(r|s)$ for any stimulus. The first model was derived and fit using spike-triggered covariance analysis, to data consisting of recorded spike responses to 1-dimensional spatiotemporal white noise (flickering bars). The model consisted of a bank of shifted linear subunit filters, a nonlinear combination rule, divisive temporal feedback, and Poisson spike generation. The second, a generalized IF model, was fit with maximum likelihood, using data consisting of spike responses to temporal binary white noise. This model contained a single linear stimulus filter, leaky integrate-and-fire spike generation, a linear spike after-current and additive Gaussian voltage noise.

While we have not yet performed a detailed side-by-side comparison of the performance of these two models, it is useful to review their most salient theoretical differences. The subunit model captures nonlinearities in the spatial pooling of information; these nonlinearities are not captured by the IF model, which uses

a single linear stimulus filter and was fit using purely temporal stimuli. Although we can locally maximize the likelihood for the parameters of an IF model augmented to include multiple filters and a nonlinear combination rule, we have no guarantee that the estimator for the resulting model will achieve the global maximum of the likelihood function. However, nonlinear stimulus dependence and log-concavity of the likelihood function *can* both be achieved if we reframe the model by using a linear kernel that operates on the stimulus after some (fixed) mapping to a nonlinear feature space. Unfortunately, it seems hard to know *a priori* how to choose the optimal nonlinear feature space, and there is no obvious sequential learning procedure for first estimating a subspace and then estimating a nonlinear combination rule, as we did in STC analysis.

On the other hand, one of the principal advantages of the IF model is that, unlike STC analysis, it can be reliably estimated using any stimulus, not just Gaussian white noise. It is therefore quite simple to estimate the IF model using responses to naturalistic stimuli, or any other set of stimuli which might be of interest. The IF model also provides a more biophysically realistic description of spike generation, gives a more accurate prediction of RGC spike train statistics (recall that the subunit model's output is Poisson), and can exhibit a more diverse repertoire of dynamical behaviors, due to the history-dependence introduced by the spike aftercurrent. Incidentally, this feature of the IF model also entails that the likelihood calculation directly elicits the probability $P(r|\vec{s}, \vec{r}_{hist})$, the probability of the response conditioned on the stimulus *and* the spike history $\vec{r}_{hist}$, rather than just $P(r|\vec{s})$. Explicitly computing $P(r|\vec{s})$ is now more difficult,

since we must average over all possible spike histories at each moment in time. Thus, although the IF model may describe the detailed time structure of the neural response much more accurately than a model with inhomogeneous Poisson spiking, it is more difficult with such a model to obtain a clear, intuitive picture of the neuron's computational transformation of the raw stimulus.

Recently, we have begun to examine a third model that occupies a middle ground between integrate-and-fire and Poisson spiking models. We call this a "generalized linear model"(GLM), which we can describe formally as

$$P(r|\vec{s},\vec{r}_{hist}) = f(\vec{k}\cdot\vec{s} + \vec{h}\cdot\vec{r}_{hist}), \tag{4.3}$$

where $\vec{h}$ is a linear filter that operates on the spike train history. Spike generation is determined probabilistically by an instantaneous function f applied to the filtered input, but the resulting process is clearly no longer Poisson, due to the induced dependence on spike train history. Recent work has shown that log-likelihood for this model is also concave (implying guaranteed convergence to the global maximum) if we impose certain constraints on the nonlinear function f[1] The likelihood calculation is much simpler and faster to implement than that for the IF model, and preliminary analyses indicate that its performance is be comparable to the IF model in accounting for RGC responses. Although the GLM is less well-known than the IF or LNP models as a description of neural spike trains, it has a simple interpretation which is not altogether biophysically implausible: we interpret the linear signal as $V = \vec{k}\cdot\vec{s} + \vec{h}\cdot\vec{r}_{hist}$, the internal

[1]Namely, that $f(x)$ be convex and log-concave, a condition satisfied by such functions as $exp(x)$ (Paninski, 2004)

voltage of the neuron, and let the accelerating nonlinearity $f(V)$ represents a "fuzzy threshold", such that the instantaneous probability of spiking increases exponentially as a function of the depolarization of V. We are currently pursuing a more thorough analysis of this model's performance on RGC data.

Undoubtedly, the work of applying and extending these ideas to provide more general and powerful models of the neural code is still in its infancy. In this thesis, we have applied statistical characterization procedures to data from retinal ganglion cells, a class whose functional properties have already been extensively explored and modeled using classical analyses. Nevertheless, important questions about the coding properties of such neurons remain unanswered. Two phenomena which demand immediate attention include the characterization of adaptive behavior in the processing of natural scenes, and the modeling of joint encoding by groups of neurons with common input and correlated responses; we have current plans to address both of these issues using the IF and GLM models. Our ultimate goal, nevertheless, is to push ahead with models that will provide insight into neural responses much deeper in the visual processing pathway. As the development of these techniques matures, we are hopeful that statistical characterization methods will eventually score major advances in revealing the coding properties of neurons in brain areas not currently understood with classical methods.

References

Adrian, E. D. (1926). The impulses produced by sensory nerve endings. *Journal of Physiology*, *61*, 49–72.

Anderson, J., Lampl, I., Gillespie, D., & Ferster, D. (2000). The contribution of noise to contrast invariance of orientation tuning in cat visual cortex. *Science*, *290*, 1968–1972.

Arcas, B. Aguera y, & Fairhall, A. L. (2003). What causes a neuron to spike? *Neural Computation*, *15*(8), 1789–1807.

Baccus, S. A., & Meister, M. (2002). Fast and slow contrast adaptation in retinal circuitry. *Neuron*, *36*, 909–919.

Balboa, R. M., & Grzywacz, N. M. (2000). The role of early lateral inhibition: More than maximizing luminance information. *Visual Neuroscience*, *17*, 77–89.

Benardete, E. A., Kaplan, E., & Knight, B. W. (1992). Contrast gain control in the primate retina: P cells are not x-like, some m cells are. *Vis. Neurosci.*, *8*(5), 483–486.

Berry, M., & Meister, M. (1998). Refractoriness and neural precision. *Journal of Neuroscience*, *18*, 2200–2211.

Berry, M., Warland, D. K., , & Meister, M. (1997). The structure and precision of retinal spike trains. *PNAS*, *94*, 5411–5416.

Bialek, W., Rieke, F., Steveninck, R. R. de Ruyter van, & Warland, D. (1991). Reading a neural code. *Science*, *252*, 1854–1857.

Bogachev, V. (1998). *Gaussian measures*. New York: AMS.

Borg-Graham, L., Monier, C., & Fregnac, Y. (1998). Visual input evokes transient and strong shunting inhibition in visual cortical neurons. *Nature*, *393*, 369–373.

Brenner, N., Bialek, W., & Steveninck, R. de Ruyter van. (2000). Adaptive rescaling optimizes information transmission. *Neuron*, *26*, 695–702.

Brown, E., Barbieri, R., Ventura, V., Kass, R., & Frank, L. (2002). The time-rescaling theorem and its application to neural spike train data analysis. *Neural Computation*, *14*, 325–346.

Bryant, H., & Segundo, J. (1976). Spike initiation by transmembrane current: a white-noise analysis. *Journal of Physiology*, *260*, 279–314.

Bussgang, J. (1952). Crosscorrelation functions of amplitude-distorted gaussian signals. *RLE Technical Reports*, *216*.

Cecchi, G. A., Sigman, M., Alonso, J. M., & Martinez, L. (2000). Noise in neurons is message dependent. *Proc Natl Acad Sci*, *97*(10), 5557–5561.

Chander, D., & Chichilnisky, E. (2001). Adaptation to temporal contrast in primate and salamander retina. *Journal of Neuroscience*, *21*, 9904–16.

Chichilnisky, E. J. (2001). A simple white noise analysis of neuronal light responses. *Network: Computation in Neural Systems*, *12*, 199–213.

Chichilnisky, E. J., & Kalmar, R. S. (2002). Functional asymmetries in on and off ganglion cells of primate retina. *J Neurosci*, *22*, 2737–2747.

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines*. Cambridge University Press.

Dayan, P., & Abbott, L. (2001). *Theoretical neuroscience*. MIT Press.

deBoer, E., & Kuyper, P. (1968). Triggered correlation. In *Ieee transact. biomed. eng.* (Vol. 15, pp. 169–179).

Dodd, T., & Harris, C. (2002). Identification of nonlinear time series via kernels. *International Journal of Systems Science*, *33*, 737–750.

Duda, R., & Hart, P. (1972). *Pattern classification and scene analysis*. New York: Wiley.

Enroth-Cugell, C., & Robson, J. G. (1966). The contrast sensitivity of retinal ganglion cells of the cat. *J. Physiol.*, *187*, 517–22.

Genz, A. (1992). Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics*, *1*, 141–149.

Gerstner, W., & Kistler, W. (2002). *Spiking neuron models: Single neurons, populations, plasticity*. Cambridge University Press.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Gross, C. G. (1998). *Brain, vision, memory; tales in the history of neuroscience*. Cambridge, MA: MIT Press.

Harris, K., Csicsvari, J., Hirase, H., Dragoi, G., & Buzsaki, G. (2003). Organization of cell assemblies in the hippocampus. *Nature*, *424*, 552–556.

Hirsch, J., Alonso, J. M., Reid, C. R., & Martinez, L. (1998). Synaptic integration in striate cortical simple cells. *J Neuroscience*, *15*, 9517–9528.

Hochstein, S., & Shapley, R. (1976). Linear and nonlinear spatial subunits in y cat retinal ganglion cells. *J. Physiol.*, *262*, 265–284.

Hodgkin, A. L., & Huxley, A. F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physiol.*, *117*(4), 500–544.

Hubel, D. H., & Wiesel, T. (1968). Receptive fields and functional architecture of monkey striate cortex. *J. Physiol (Lond)*, *195*, 215–243.

Jolivet, R., Lewis, T., & Gerstner, W. (2003). The spike response model: a framework to predict neuronal spike trains. *Springer Lecture notes in computer science*, *2714*, 846–853.

Jolivet, R., Lewis, T., & Gerstner, W. (2004). Generalized integrate-and-fire models of neuronal activity approximate spike trains of a detailed model to a high degree of accuracy. *J. Neurophysiology*, *92*, 959–976.

Jones, J. P., & Palmer, L. A. (1987). The two-dimensional spatial structure of simple receptive fields in the cat striate cortex. *J Neurophysiology*, *58*, 1187–11211.

Karlin, S., & Taylor, H. (1981). *A second course in stochastic processes*. New York: Academic Press.

Keat, J., Reinagel, P., Reid, R., & Meister, M. (2001). Predicting every spike: a model for the responses of visual neurons. *Neuron*, *30*, 803–817.

Kim, K. J., & Rieke, F. (2001). Temporal contrast adaptation in the input

and output signals of salamander retinal ganglion cells. *J. Neurosci.*, *21*, 287–299.

Knight, B., Omurtag, A., & Sirovich, L. (2000). The approach of a neuron population firing rate to a new equilibrium: an exact theoretical result. *Neural Computation*, *12*, 1045–1055.

Kuffler, S. W. (1953). Discharge patterns and functional organization of mammalian retina. *J. Neurophysiol.*, *16*, 37–68.

Levin, J., & Miller, J. (1996). Broadband neural encoding in the cricket cercal sensory system enhanced by stochastic resonance. *Nature*, *380*, 165–168.

Marmarelis, P. Z., & Naka, K. (1972). White-noise analysis of a neuron chain: an application of the wiener theory. *Science*, *175*, 1276–1278.

Meister, M., & Berry, M. J. (1999). The neural code of the retina. *Neuron*, *22*, 435–450.

Movshon, J. A., & Newsome, W. T. (1996). Visual response properties of striate cortical neurons projecting to area mt in macaque monkeys. *J. Neurosci.*

Paninski, L. (2003). Convergence properties of some spike-triggered analysis techniques. *Network: Computation in Neural Systems*, *14*, 437–464.

Paninski, L. (2004). Maximum likelihood estimation of cascade point-process neural encoding models. *Submitted, Network: Computation in Neural Systems.*

Paninski, L., Fellows, M., Shoham, S., Hatsopoulos, N., & Donoghue, J. (2003). Nonlinear population models for the encoding of dynamic hand position signals in primary motor cortex. *Annual Computational Neuroscience Meet-*

ing, Alicante, Spain, Poster presentation.

Paninski, L., Lau, B., & Reyes, A. (2003). Noise-driven adaptation: in vitro and mathematical analysis. *Neurocomputing, 52*, 877–883.

Paninski, L., Pillow, J., & Simoncelli, E. (2004). Maximum likelihood estimation of a stochastic integrate-and-fire neural model. *Neural Computation, In press.*

Pillow, J. W., Paninski, L., & Simoncelli, E. (2004). Maximum likelihood estimation of a stochastic integrate-and-fire neural model. In L. S. S. Thrun & B. Scholkopf (Eds.), *Advances in neural information processing systems* (Vol. 16).

Pillow, J. W., Paninski, L., Uzzell, V. J., Simoncelli, E. P., & Chichilnisky, E. J. (2005). Structure, variability and optimal decoding of retinal ganglion cell responses. *Manuscript under review.*

Pillow, J. W., & Simoncelli, E. (2003). Biases in white noise analysis due to non-Poisson spike generation. *Neurocomputing, 52*, 109–115.

Press, W., Teukolsky, S., Vetterling, W., & Flannery, B. (1992). *Numerical recipes in C.* Cambridge University Press.

Reich, D. S., Victor, J. D., & Knight, B. W. (1998). The power ratio and the interval map: Spiking models and extracellular recordings. *The Journal of Neuroscience, 18*, 10090–10104.

Reich, D. S., Victor, J. D., Knight, B. W., Ozaki, T., & Kaplan, E. (1997). Response variability and timing precision of neuronal spike trains in vivo. *J. Neurophysiol., 77*, 2836–41.

Reinagel, P. (2001). How do visual neurons respond in the real world? *Current Opinion in Neurobiology*, *11*, 437–442.

Reinagel, P., & Reid, R. C. (2000). Temporal coding of visual information in the thalamus. *Journal of Neuroscience*, *20*, 5392–5400.

Rinott, Y. (1976). On convexity of measures. *Annals of Probability*, *4*, 1020–1026.

Rudd, M., & Brown, L. (1997). Noise adaptation in integrate-and-fire neurons. *Neural Computation*, *9*, 1047–1069.

Ruderman, D., & Bialek, W. (1994). Statistics of natural images: scaling in the woods. *Physics Review Letters*, *73*, 814–817.

Rust, N. C., Schwartz, O., Movshon, J. A., & Simoncelli, E. P. (2004). Spike-triggered characterization of excitatory and suppressive stimulus dimensions in monkey V1 directionally selective neurons. In *Neurocomputing.* Elsevier. (Presented at Computational NeuroScience (CNS*03), Alicante Spain, July 2003)

Sahani, M. (2000). *Kernel regression for neural systems identification.* Presented at NIPS00 workshop on Information and statistical structure in spike trains; abstract available at http://www-users.med.cornell.edu/∼jdvicto/nips2000speakers.html.

Sahani, M., & Linden, J. (2003). Evidence optimization techniques for estimating stimulus-response functions. *NIPS*, *15*.

Sakai, Y., Fnahashi, S., & Shinomoto, S. (1999). Temporally correlated inputs to leaky integrate-and-fire models can reproduce spiking statistics of cortical neurons. *Neural Networks*, *12*, 1181–1190.

Schwartz, O., Chichilnisky, E. J., & Simoncelli, E. P. (2002). Characterizing neural gain control using spike-triggered covariance. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Adv. neural information processing systems* (Vol. 14). Cambridge, MA: MIT Press.

Schwartz, O., & Simoncelli, E. P. (2001, August). Natural signal statistics and sensory gain control. *Nature Neuroscience*, *4*(8), 819–825.

Shadlen, M., & Newsome, W. (1998). The variable discharge of cortical neurons: implications for connectivity, computation, and information coding. *Journal of Neuroscience*, *18*, 3870–3896.

Shapley, R. M., & Victor, J. D. (1981). How the contrast gain control modifies the frequency response of cat retinal ganglion cells. *J. Physiol. (Lond)*, *318*, 161–171.

Sharpee, T., Rust, N., & Bialek, W. (2004). Analyzing neural responses to natural signals: Maximally informative dimensions. *Neural Computation*, *16*, 223–250.

Simoncelli, E., Paninski, L., Pillow, J. W., & Schwartz, O. (2004). Characterization of neural responses with stochastic stimuli. In M. Gazzaniga (Ed.), *The cognitive neurosciences* (3rd ed.). MIT Press.

Smirnakis, S., Berry, M., Warland, D., Bialek, W., & Meister, M. (1997). Adaptation of retinal processing to image contrast and spatial scale. *Nature*, *386*, 69–73.

Smith, R. G., & Sterling, P. (1990). Cone receptive field in cat retina computed from microcircuitry. *Visual Neuroscience*, *5*, 453–461.

Sterling, P. (1983). Microcircuitry of the cat retina. *Ann. Rev. Neurosci.*, *6*, 149–185.

Touryan, J., Lau, B., & Dan, Y. (2002). Isolation of relevant visual features from random stimuli for cortical complex cells. *Journal of Neuroscience*, *22*, 10811–10818.

Troy, J. B., & Lee, B. B. (1994). Steady discharges of macaque retinal ganglion cells. *Visual Neurosci.*, *11*, 111–118.

Tsodyks, M., Kenet, T., Grinvald, A., & Arieli, A. (1999). Linking spontaneous activity of single cortical neurons and the underlying functional architecture. *Science*, *286*, 1943–1946.

Uzzell, V. J., & Chichilnisky, E. J. (2004). Precision of spike trains in primate retinal ganglion cells. *Journal of Neurophysiology*, *92*, 780–789.

Victor, J. (2000). How the brain uses time to represent and process visual information. *Brain Research*, *886*, 33–46.

Victor, J. D. (1987). The dynamics of the cat retinal x cell centre. *J. Physiol. (Lond)*, *386*, 219–246.

Victor, J. D., & Purpura, K. (1997). Metric-space analysis of spike trains: theory, algorithms, and application. *Network*, *8*, 127–164.

Victor, J. D., & Shapley, R. M. (1979a). The nonlinear pathway of y ganglion cells in the cat retina. *J. Gen. Physiol.*, *74*(6), 671–689.

Victor, J. D., & Shapley, R. M. (1979b). Receptive field mechansims of cat x and y retinal ganglion cells. *J. Gen. Physiol.*, *74*, 275–298.

Yu, Y., & Lee, T. (2003). Dynamical mechanisms underlying contrast gain

control in single neurons. *Physical Review E*, *68*, 011901.

Zhang, K., Ginzburg, I., McNaughton, B., & Sejnowski, T. (1998). Interpreting neuronal population activity by reconstruction: Unified framework with application to hippocampal place cells. *Journal of Neurophysiology*, *79*, 1017–1044.