REVIEW **Communicated by Matthias Bethge**

Learning Quadratic Receptive Fields from Neural Responses to Natural Stimuli

Kanaka Rajan
krajan@princeton.edu
Joseph Henry Laboratories of Physics and Lewis–Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544, U.S.A.

Olivier Marre
oliver.marre@gmail.com
Institut de la Vision, UPMC UMRS 968, INSERM, F-75012 Paris, France

Gašper Tkačik
gtkacik@ist.ac.at
Institute of Science and Technology Austria, A-3400 Klosterneuburg, Austria

Models of neural responses to stimuli with complex spatiotemporal correlation structure often assume that neurons are selective for only a small number of linear projections of a potentially high-dimensional input. In this review, we explore recent modeling approaches where the neural response depends on the quadratic form of the input rather than on its linear projection, that is, the neuron is sensitive to the local covariance structure of the signal preceding the spike. To infer this quadratic dependence in the presence of arbitrary (e.g., naturalistic) stimulus distribution, we review several inference methods, focusing in particular on two information theory–based approaches (maximization of stimulus energy and of noise entropy) and two likelihood-based approaches (Bayesian spike-triggered covariance and extensions of generalized linear models). We analyze the formal relationship between the likelihood-based and information-based approaches to demonstrate how they lead to consistent inference. We demonstrate the practical feasibility of these procedures by using model neurons responding to a flickering variance stimulus.

1 Introduction

A basic challenge in sensory neuroscience has been to develop concise descriptions of how neurons encode and transmit information about the stimulus. Models that attempt to capture the essence of this neural computation—the transformation of stimuli into spiking responses—without necessarily being derived from an underlying dynamical or

biophysical model of neural function are called functional models (see Wu, David, & Gallant, 2006 for an in-depth review; also Agüera y Arcas, Fairhall, & Bialek, 2003; Agüera y Arcas & Fairhall, 2003; Hong, Agüera y Arcas, & Fairhall, 2007; Lundstrom, Hong, & Fairhall, 2008; Ostojic & Brunel, 2011, for papers that link functional to dynamical models). As a consequence, functional models are usually fully learned from data, and their success depends critically on two factors: whether typical electrophysiological recordings can provide adequate data for successful inference of the model's parameters and whether effective inference algorithms exist for these parameters. Within these limitations, functional models have dramatically influenced our view of early sensory processing by mathematically summarizing the notions of receptive field, linear, and high-order stimulus sensitivity (captured by the filtering operations of matching order performed on the stimulus), as well as subsequent neural computations leading to the generation of a spike (captured by the nonlinearities operating on filter outputs).

In the functional modeling framework, the responses of many sensory neurons can be well characterized by assuming that the initial transformation of the stimulus is a linear filtering operation, that is, that the response of the neuron depends on only a single projection $\mathbf{k}\cdot\mathbf{s}(t)$ of the (possibly high-dimensional) stimulus $\mathbf{s}(t)$ onto the neuron's linear filter $\mathbf{k}$ (see Figure 1a). This view has been so successful that we tend to use the terms *filter* and *receptive field* interchangeably. For some neurons, however, their description in terms of a single linear filter is insufficient. One of the best-known examples is that of a complex cell in the primary visual cortex. Complex cells are characterized by the invariance of their responses to changes in the phase of the stimulus: their response remains constant as $\mathbf{s}(t)$ is changed into $-\mathbf{s}(t)$ by flipping dark regions of the stimulus into bright ones and vice versa. The simplest way in which such an invariance could be captured mathematically would be to assume that the stimulus $\mathbf{s}$ enters the neural response squared rather than at linear order. In other words, the stimulus sensitivity of complex cells is quadratic; it depends on the term $\mathbf{s}^{\mathrm{T}}(t)\mathbf{Q}\mathbf{s}(t)$, where $\mathbf{Q}$ is a quadratic filter specific to each cell (see Figure 1c).

Even neurons that are adequately characterized by a single linear filter when probed by relatively simple stimuli could modulate their responses strongly when high-order features of the stimulus change. For example, while retinal ganglion cells exhibit strong center-surround filters, they also change the gain of their responses with changing contrast, a second-order stimulus feature, within their receptive fields. Similar cases, where the stimulus sensitivity is purely quadratic (e.g., non-phase-locked auditory models or some motion-sensitive neurons), or where quadratic features like the shape of the signal envelope have a strong modulatory effect, are abundant in the sensory periphery. Additionally, response phenomena beyond phase invariance in the visual cortex, grouped together as relating to the nonclassical receptive field, could also be manifestations of quadratic or high-order sensitivity (Zetzsche & Nuding, 2005). This has recently sparked a lot of

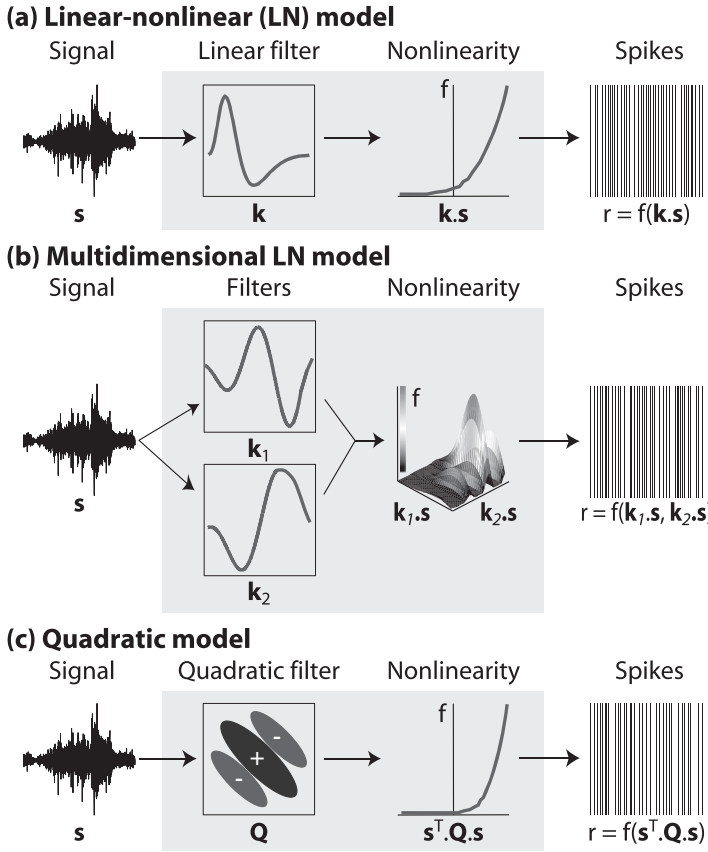

Figure 1: Functional models of neural computation. (a) The instantaneous firing rate r, or probability per unit time of emitting a spike, in a linear-nonlinear model neuron is obtained by passing the input signal through a linear filter $\mathbf{k}$ and mapping the resulting value through a point-wise nonlinearity f. (b) A multidimensional LN model neuron requires the signal to be filtered through K (here, $K = 2$) linear filters. The stimulus projections are mapped into the firing rate through a K-dimensional nonlinear function. (c) The model for quadratic stimulus sensitivity. The initial filtering step is quadratic (determined by the matrix $\mathbf{Q}$), and the nonlinear function f is one-dimensional.

interest in developing generic and tractable methods for the functional characterization of neural responses where the stimulus sensitivity could be as high as second order.

In this review, we focus on neural models with quadratic stimulus sensitivity and the corresponding inference methods, emphasizing recent ones

that can be applied to any stimulus ensemble, including fully naturalistic movies, and also to cases where the quadratic filter **Q** is a full-rank matrix. We start with a brief overview of methods for inferring linear stimulus sensitivity in section 2, which we extend to the discussion of models with multiple linear features in section 3. We show how quadratic stimulus dependence could arise as a special case of neural models with multiple linear features and present biologically motivated examples of quadratic stimulus sensitivity in section 4. We then review several complementary approaches that can be used to learn quadratic stimulus dependence even when neurons are responding to rich, naturalistic stimuli: we discuss the maximally informative stimulus energy (Rajan & Bialek, 2012) and the maximization of noise entropy (Fitzgerald, Rowekamp, Sincich, & Sharpee, 2011; Fitzgerald, Sincich, & Sharpee, 2011; Globerson, Stark, Vaadia, & Tishby, 2009) in sections 5.1 and 5.2 followed by the Bayesian spike-triggered covariance method (Park & Pillow, 2011) and related extensions of generalized linear models to include quadratic stimulus dependence in section 5.3.[1] We show the conditions under which information-theoretic and likelihood-based approaches lead to consistent inference in the appendix.

2 Receptive Fields and Linear Stimulus Dependence

The space of all possible stimuli and the space of all possible neural responses is vast. Consider, for instance, all possible image sequences incident on the retina or sets of output spike trains. Our progress in building functional models must therefore depend on making well-chosen simplifying assumptions. An example of extreme simplification involves varying the stimulus along a single dimension, as in the case of the orientation or wavelength of a drifting grating visual stimulus, and representing the output by a single scalar quantity, like the average firing rate in a chosen time bin. These measurements have traditionally been represented in terms of tuning curves and have provided basic insights into principles of sensory (and population) coding (Dayan & Abbott, 2001). However, the relevance of the tuning curve approach is limited by the choice of the dimension along which the stimulus is manipulated, which may drastically underestimate the complexity in the structure of the stimuli to which the neuron actually responds. Despite these limitations, such studies have helped establish the concept of a receptive field, the region of stimulus space where changes in the stimulus modulate the spiking behavior of the neuron.

Central to the receptive field concept is the notion of locality in the stimulus or feature space. For instance, a ganglion cell in the retina may be sensitive only to specific changes in light intensity that occur within a small visual angle (Hartline, 1940). A productive way of capturing this

[1] We have worked out this problem in parallel with Park and Pillow (2011).

notion of locality has been to think of a receptive field as one or more filters that act on the stimulus; only stimulus variations that result in measurable changes in the filter output have the ability to affect the neural response. In this view, neurons perform dimensionality reduction by projecting the stimulus down into a small number of features. Consequently, the success of data analysis techniques built around this idea must depend on whether a small number of features or filters suffices to fully account for the neuron's sensitivity and its response properties.

Methods based in systems identification theory have provided systematic procedures to infer both the receptive fields of neurons as well as subsequent computations (see Table 1 for an overview of various functional models and related inference techniques). These techniques usually share two key features. One is that they can (and sometimes must) be used with stimuli that sample the stimulus space broadly, making no explicit assumptions about which stimulus features are important. This is in contrast to the restricted stimuli employed for measuring tuning curves. The other is that the procedures usually involve a series of approximations that can provably yield a better description of the system if more data are available. Among the earliest to be used successfully, Wiener and Volterra expansions helped identify the first- and second-order kernels mapping the stimulus to response time traces in various systems (Marmarelis & Marmarelis, 1978; Recio-Spinoso, Temchin, van Dijk, Fan, & Ruggero, 2005; Sakai, 1992; Schetzen, 1989; Victor & Knight, 1979; Wiener, 1958). However, in many cases, the strong intrinsic nonlinearities underlying spike generation require a large number of terms in Wiener-Volterra expansions, even though the underlying stimulus sensitivity might be much simpler and therefore of low order.[2] Models in which the (possibly linear) projections of the stimulus in the receptive field were decoupled from the nonlinearities underlying spike generation, as in linear-nonlinear (LN) architectures illustrated in Figure 1, made further progress feasible.

LN and LN-like models have been used widely and profitably to predict the firing rate traces of single sensory neurons, because their parameters can be easily inferred under suitable conditions. However, the more intriguing cases are the ones where LN models perform poorly or fail entirely. One such failure mode is the inability to account for the statistics of neural activity beyond the mean firing rate. Specifically, real sensory neurons often have variability that is smaller than that attributed to Poisson processes (de Ruyter van Steveninck, Lewen, Strong, & Bialek, 1997); phenomena like

[2]When we speak of the order (e.g., linear, quadratic), we refer to the order of the kernel operating on the stimulus, which can be defined unambiguously. In contrast, the order of the neural processing system as a whole depends on stimulus statistics; for example, high-order statistical structure in the stimulus can conflate first- and second-order responses of the system. Likewise, aspects of the response explained by a second-order kernel inferred though gaussian noise depend on the power spectrum of the input.

Table 1: Functional Models for Single Neurons and the Related Inference Methods.

	Method	Stimulus Type	Models and Restrictions	References
A	Wiener/Volterra expansion	Gaussian white noise, sum of sinusoids	$r = r_0 + \mathbf{k} \cdot \mathbf{s} + \mathbf{s}^{\mathbf{T}} \cdot \mathbf{Q} \cdot \mathbf{s} + \cdots$	Marmarelis & Marmarelis, 1978; Schetzen, 1989; Wiener, 1958; Recio-Spinoso et al., 2005; Victor & Knight, 1979
B	Spike-triggered average (STA) (reverse correlation)	Spherically symmetric, binary noise, m-sequences	LN (single filter), isolated spikes $r = f(\mathbf{k} \cdot \mathbf{s})$	de Boer & Kuyper, 1968; Paninski, 2003; Reid, Victor, & Shapley, 1997; Schwartz, Pillow, Rust, & Simoncelli, 2006; Simoncelli, Paninski, Pillow, & Schwartz, 2004
	Debiased STA (reverse correlation)	"Gaussian-like" asym. 1-point histogram	LN (single filter), isolated spikes $r = f(\mathbf{k} \cdot \mathbf{s})$	Lesica, Ishii, Stanley, & Hosoya, 2008
	Differential reverse correlation (dRC) (reverse correlation)	Spike-triggering snippet	Linear feature that predicts spike timing $t_{spike} \propto \mathbf{k} \cdot \mathbf{s}$	Tkačik & Magnasco, 2008
	(maximum likelihood)	Any	Leaky integrate-and-fire (LIF/LN-LIF)	Gerstner & Kistler, 2002; Paninski et al., 2004; Pillow, 2007
	Error function minimization (general fitting methods)	Any	Dynamical extensions of LN $r = \Theta(h)\dot{h}, h = \mathbf{k} \cdot \mathbf{s} + \mathbf{q} \cdot \mathbf{y} + \eta$ (Keat), $\dot{A}_i = M_{ij}(f(\mathbf{k} \cdot \mathbf{s}))A_j, r = A_1$ (LNK)	Keat et al., 2001; Ozuysal & Baccus, 2012
	Generalized linear models (GLM) (maximum likelihood)	Any	Point process (dependence on past spiking) $r = f(\mathbf{k} \cdot \mathbf{s} + \mathbf{q} \cdot \mathbf{y} +$ (effect of other neurons))	Paninski, 2004; Pillow et al., 2008; Truccolo, Eden, Fellows, Donoghue, & Brown, 2004; Gerwinn, Macke, & Bethge, 2010; Pillow, 2007
	Isoresponse mapping	synthetic stimuli (parameterizable, low-D)	LNLN cascade $r = f(\mathbf{k}_1 * g(\mathbf{k}_2 * \mathbf{s}))$	Bölinger & Gollisch, 2012; Gollisch & Herz, 2005
C	Generalized linear models (GLM) (with multiple nonlinear stimulus features)	Any	point process (dependence on past spiking) $r = f(\sum_i w_i g_i(\mathbf{k}_i \cdot \mathbf{s}) + \mathbf{q} \cdot \mathbf{y} + \ldots)$	Gerwin, Macke, Seeger, & Bethge, 2008

Table 1: *Continued.*

	Method	Stimulus Type	Models and Restrictions	References
	Maximally informative dimensions (MID) (info maximization)	Any	LN (multiple filters) $r = f(\mathbf{k}_1 \cdot \mathbf{s}, \ldots, \mathbf{k}_K \cdot \mathbf{s})$	Kouh & Sharpee, 2009; Sharpee et al., 2006; Sharpee, Rust, & Bialek, 2004
	Extended projection pursuit regression (ePPR)	Any	LN (multiple filters) $r = f(\mathbf{k}_1 \cdot \mathbf{s}, \ldots, \mathbf{k}_K \cdot \mathbf{s})$	Rapela, Felsen, Touryan, Mendel, & Grzywacz, 2010
	Spike-triggered covariance (STC) (reverse correlation)	Gaussian	LN (multiple filters), isolated spikes $r = f(\mathbf{k}_1 \cdot \mathbf{s}, \ldots, \mathbf{k}_K \cdot \mathbf{s})$	de Ruyter van Steveninck & Bialek, 1988; Bialek & de Ruyter van Steveninck, 2005; Simoncelli et al., 2004; Fairhall et al., 2006; Maravall, Petersen, Fairhall, Arabzadeh, & Diamond, 2007; Schwartz et al., 2002
	iSTAC (parametric MID with conditional gaussian model)	Gaussian	LN (multiple filters) $r = f(\mathbf{k}_1 \cdot \mathbf{s}, \ldots, \mathbf{k}_K \cdot \mathbf{s})$	Pillow & Simoncelli, 2006
D	Maximization of noise entropy (convex optimization)	Any	$r = \text{logistic}(k_0 + \mathbf{k} \cdot \mathbf{s} + \mathbf{s}^\mathbf{T} \cdot \mathbf{Q} \cdot \mathbf{s})$	Fitzgerald, Sincich et al., 2011; Fitzgerald, Rowekamp et al., 2011; Globerson et al., 2009
	Bayesian STC/quadratic GLM (likelihood maximization)	Any	additive linear and quadratic contributions $r = f(\mathbf{k} \cdot \mathbf{s} + \mathbf{s}^\mathbf{T} \cdot \mathbf{Q} \cdot \mathbf{s})$	Park & Pillow, 2011
	Maximally informative stimulus energy (MISE) (info maximization)	Any	general quadratic model $r = f(\mathbf{k} \cdot \mathbf{s}, \mathbf{s}^\mathbf{T} \cdot \mathbf{Q} \cdot \mathbf{s})$	Rajan & Bialek, 2012

Notes: The table is organized into four blocks. (A) Wiener-Volterra expansions, followed by (B) methods with linear stimulus sensitivity, (C) methods designed for inferring multiple stimulus features, and (D) methods explicitly designed to capture quadratic stimulus dependence. r is the firing rate or the probability of spiking; $\mathbf{k}$ are linear filters acting on stimulus clips $\mathbf{s}$; $\mathbf{Q}$ is a matrix representing a quadratic kernel; $\mathbf{q}$ is a linear filter operating on the sequence of past spikes $\mathbf{y}$; $f(\cdot), g(\cdot)$ are arbitrary nonlinear functions; $*$ denotes a convolution; Θ is a thresholding operation (1 when the argument crosses some threshold from below, 0 otherwise); η is a white noise Langevin force. In this review, we use the term *gaussian* to denote stimuli whose components are jointly gaussian and possibly correlated (i.e., nonwhite), unless otherwise noted. While reverse correlation methods are formally simpler for uncorrelated (white) gaussian noise, it is possible to generalize them for use with correlated noise ensembles. For example, to compute an unbiased estimate of the linear (L) part of the model using STA and a correlated stimulus, we need to correct for stimulus correlations by multiplying the spike-triggered average with the inverse covariance matrix. For an extensive review of spike-triggered (reverse correlation) methods, see Schwartz et al. (2006).

refractoriness and spike frequency adaptation are not captured by LN models (Berry & Meister, 1998), and in neural populations, uncoupled LN models fail to reproduce the basic covariance structure of neural activity (Granot-Atedgi, Tkačik, Segev, & Schneidman, 2012; Pillow et al., 2008; Schneidman, Berry, Segev, & Bialek, 2006). Some of these issues can be addressed by adding suitable dynamical complexity beyond the linear filtering stage to make the nonlinearities in spike generation more realistic (Keat, Reinagel, Reid, & Meister, 2001; Paninski, Pillow, & Simoncelli, 2004) or by including interactions between neurons in models of neural firing (Granot-Atedgi et al., 2012; Pillow et al., 2008).

A different kind of failure of LN models rests on the assumption that stimulus sensitivity occurs through a single linear projection (or a small number of them). One example is contrast adaptation, where a simple LN model derived from a white noise stimulus of a certain variance fails to predict the response to a stimulus with smaller or larger variance (Baccus & Meister, 2002; Borst & Egelhaaf, 1987; van Hateren, 1992; de Ruyter van Steveninck, Zaagman, & Mastebroek, 1986; Smirnakis, Berry, Warland, Bialek, & Meister, 1997). Other examples include the failure to account for the sensitivity of retinal ganglion cells to fine spatial detail (possibly because of nonlinear summation within the receptive field; Demb, Zaghloul, Haarsma, & Sterling, 2001; Schwartz et al., 2012), or to stimulus motion (Berry, Brivanlou, Jordan, & Meister, 1999; Chen et al., 2012; Gollisch & Meister, 2010; Schwartz, Taylor, Fisher, Harris, & Berry, 2007). Generally these difficulties emerge clearly when the stimulus statistics change or increase in complexity beyond those used to infer the model, for instance, by becoming more "naturalistic"—that is, having pairwise temporal and spatial correlation, skewed first-order histograms, or statistical structure beyond second order.

The problems with LN models can generally be addressed in two possible ways. In the first, LN models can be extended to account for a particular phenomenon on a specific stimulus, for example, by adding a contrast gain control mechanism (Schwartz & Simoncelli, 2001; Schwartz, Chichilnisky, & Simoncelli, 2002) or by an ad hoc rescaling of nonlinearities (Brenner, de Ruyter van Steveninck, & Bialek, 2000) to account for contrast adaptation in an experiment where the variance of a gaussian input is modulated. The second approach is to find the complete (or close to complete) set of features to which the neuron responds by examining the neural responses to relatively rich noise stimuli or fully naturalistic stimuli. Noise stimulation (e.g., white noise) is analytically convenient and can provide easily obtainable unbiased estimates of linear filters, but it remains highly unnatural. While ethologically more relevant, fully natural stimuli can lead to technical obstacles in model inference, mainly due to the statistical intractability of the natural ensemble (Geisler, 2008; Simoncelli & Olshausen, 2001). The choice of the appropriate stimulus deserves a lengthier discussion (e.g., Rust & Movshon, 2005), which is beyond the scope of this review. We do,

however, wish to emphasize two points. First, it has been shown that under conditions of naturalistic stimulation, even basic filter responses of cells can change (Sharpee et al., 2006), and response mechanisms that are intractable via noise stimulation become engaged (e.g., Olveczky, Baccus, & Meister, 2007). As a consequence, finding the complete set of features that characterize the neural response across different stimulus ensembles is an elusive goal, and in practice we are often satisfied with results specific to one rich stimulus type. The second point is methodological: the applicability of some inference techniques is restricted to special stimulus types, while others permit unbiased inference with arbitrary stimuli, a distinction we make explicit in the second column of Table 1.

To characterize the sensitivity of a neuron to the selected rich stimulus ensemble more fully, one can look for multiple linear filters. A number of approaches exist for this task (see Table 1). While some (e.g. spike-triggered covariance, STC) permit us to identify several relevant stimulus dimensions, understanding how the corresponding stimulus projections influence spiking output can be difficult unless we make further simplifying assumptions. One possible anatomically motivated simplification of a multifeature LN model is a cascade LN (an LNLN) model, where the nonlinearly transformed filter outputs are linearly summed and passed through a spike-generating nonlinearity. Despite some successes (Bölinger & Gollisch, 2012; Gollisch & Herz, 2005; Schwartz et al., 2012), the general problem of inferring cascading models with multiple linear filters remains technically challenging (usually involving difficult optimizations). A somewhat simpler LNL system has proven to account for the behavior of the Y-type retinal ganglion cells very well and is tractable to infer using a sum-of-sinusoids version of the Wiener formalism (Victor & Knight, 1979; Victor & Shapley, 1979, 1980).

Models in which the stimulus sensitivity is quadratic rather than of linear order are of particular biological significance, as we briefly touched on in section 1. Quadratic stimulus dependence is formally a restricted case of LNLN models, which in turn are an instance of multifeature LN models. Taken together, biologically motivated quadratic stimulus dependence provides the necessary mathematical simplifications for the very general multifeature LN model that make the problem of inferring quadratic models tractable. Next we briefly introduce multifeature LN models and then focus specifically on the issue of quadratic stimulus dependence.

3 Multiple Linear Features

In a typical experiment, a neuron can be driven by a synthetic stimulus containing any desired statistical structure. For probing the visual system, for example, this stimulus might be a random binary checkerboard, a drifting grating, or full-field light intensity flicker. If the neuron's response depends solely on the stimulus presented in the recent past of duration T (and

possibly on its own previous spiking behavior), we can restrict our attention to stimulus clips $\mathbf{s}$ of length $\geq T$. These clips are drawn from a distribution $P(\mathbf{s})$ that characterizes the stimulus; the N components of the vector $\mathbf{s}$ represent successive stimulus values in time and optionally across space. Our task is then to infer the dependence of the instantaneous spiking probability (firing rate) at time t, on the stimulus $\mathbf{s}(t)$ presented just prior to t.

If the neuron is well described by an LN model, where the spiking rate r is an arbitrary positive, point-wise nonlinear function f of the stimulus projected onto the filter, $r(\mathbf{s}) = f(\mathbf{k} \cdot \mathbf{s})$, and the stimulus distribution is chosen to be spherically symmetric, $P(\mathbf{s}) = P(|\mathbf{s}|)$, we can use the spike-triggered average (STA) to obtain an unbiased estimate of the single linear filter $\mathbf{k}$ (de Boer & Kuyper, 1968; Simoncelli et al., 2004). Spike-triggered covariance (STC) generalizes the filter inference to cases where the firing rate depends nonlinearly on $K \geq 1$ projections of the stimulus, $r(\mathbf{s}) = f(\mathbf{k}_1 \cdot \mathbf{s}, \mathbf{k}_2 \cdot \mathbf{s}, \ldots, \mathbf{k}_K \cdot \mathbf{s})$ (de Ruyter van Steveninck & Bialek, 1988). The number of relevant linear filters K is equal to the number of nonzero eigenvalues of the spike-triggered covariance matrix. A successful application of STC requires $P(\mathbf{s})$ to be gaussian. STC has been used successfully, for example, to understand the computations performed by motion-sensitive neurons of the blowfly (Bialek & de Ruyter van Steveninck, 2005), map out the sensitivity to full-field flickering stimuli in salamander retinal ganglion cells (Fairhall et al., 2006), explore contrast gain control (Rust, Schwartz, Movshon, & Simoncelli, 2004; Schwartz et al., 2002), and understand adaptation in the rodent barrel cortex (Maravall et al., 2007).[3]

[3]Before moving on, it seems appropriate to return to the Wiener-Volterra formalism and contrast it with spike-triggered methods for recovering LN models. The underlying assumptions of the two approaches may appear to be substantially different: first, because of the presence of the nonlinear (N) transformation in the LN model, and second, because the output of the LN model is usually taken to predict the rate of a stochastic point process, while Wiener-Volterra series are intended for analyzing deterministic systems (Wiener, 1958). Nevertheless, it is easy to see that when uncorrelated (i.e., white) gaussian noise is used to extract the filters of the LN model using spike-triggered average (STA) and spike-triggered covariance (STC), STA and STC also provide unbiased estimates (up to a scaling factor) of first- and second-order Wiener-Volterra kernels. The difference arises in subsequent analysis steps. In case of LN models, STA and STC are used solely as dimensionality-reduction steps to identify the relevant subspace of the stimuli in which the nonlinear transformation acts, while in the Wiener-Volterra formalism, STA and STC are literally the first two terms in a functional expansion that provides the best least-squares fit to the observed firing rate. Victor and Johannesma (1986) have further demonstrated that the Wiener-Volterra formalism is a special case of a general probabilistic maximum entropy framework for describing the joint distributions of stimuli and the responses they evoke. In this framework, for example, the classic Wiener-Volterra formalism is recovered if the stimulus distribution is gaussian and the response variable is also gaussian with additive noise. If the output variable is binary (spike/no spike), the same maximum entropy approach reduces to identifying LN-type models with exponential nonlinearities.

While powerful and simple to use, spike-triggered covariance (STC) is guaranteed to yield unbiased results only for gaussian stimuli (Paninski, 2003).[4] Under this restriction, STC can reliably extract from $K = 1$ to $K \sim 10$ relevant linear filters (Rust et al., 2004) for realistic recording durations. It is much more difficult to directly sample the K-dimensional nonlinear function f for K larger than 2 or 3 without making additional assumptions. In case of quadratic stimulus dependence, however, such a simplification occurs naturally. Every real, symmetric matrix, including the putative quadratic filter $\mathbf{Q}$, can be spectrally decomposed into $\mathbf{Q} = \sum_{i=1}^{N} \lambda_i \mathbf{k}_i \mathbf{k}_i^{\mathrm{T}}$. The response of the quadratic model is thus

$$r = f(\mathbf{s}^{\mathrm{T}} \cdot \mathbf{Q} \cdot \mathbf{s}) = f\left[\sum_{i=1}^{N} \lambda_i (\mathbf{k}_i \cdot \mathbf{s})^2\right], \tag{3.1}$$

explicitly demonstrating that quadratic models are special cases of the LNLN cascade, where the first linear stage applies the filters $\mathbf{k}_i$, the first nonlinear stage squares the projections, the second linear stage is a summation with weights λ_i, and the final transformation applies the nonlinearity $f(\cdot)$. This means that to infer quadratic dependence, we can identify the relevant stimulus filters using STC, parameterize the nonlinearity as a 1D nonlinear function of a linear combination of squared filter projections, and use maximum likelihood to infer the parameters of the nonlinearity (see, e.g., Schwartz et al., 2002).

Unfortunately, the gaussian ensemble can be a serious restriction for neurons that do not respond well (or at all) to unstructured stimuli; furthermore, under exclusively gaussian stimulation, we are likely to miss several neural mechanisms that depend on naturalistic statistical structure such as correlations and intermittency. A versatile method should therefore be able to successfully infer the multiple-filter dependence of a neuron probed with any stimulus of arbitrary complexity. Maximally informative dimensions (MID) (Sharpee et al., 2004) or likelihood inference for single-filter generalized linear models (Gerwinn et al., 2010; Paninski, 2004; Pillow, 2007; Pillow et al., 2008; Truccolo et al., 2004) have been used to this end when the dependence is linear, but until recently, the attempts to incorporate quadratic stimulus dependence into procedures that can be used with arbitrary stimuli have been uncommon.

[4] Specifically, if stimuli are nongaussian (even if spherically symmetric), there exist nonlinear functions f for which filter estimates given by STC will be biased. An example of such bias with the binary stimulus is given in Schwartz et al. (2006).

4 Quadratic Stimulus Dependence

Let us start by discussing a few examples of quadratic stimulus dependence. Consider a situation where the neuron has a vanishing spike-triggered average, as with complex cells, non-phase-locked auditory neurons (Recio-Spinoso et al., 2005), or motion-sensitive neurons. In these cases, a natural starting point would be a search for more than a single linear filter. For a model complex cell in the visual cortex, we would find two phase-shifted vectors $\mathbf{k}_1$ and $\mathbf{k}_2$ that together form a quadrature pair, such that the most informative variable concerning the neuron's firing is the power,

$$r(\mathbf{s}) = f[(\mathbf{k}_1 \cdot \mathbf{s})^2 + (\mathbf{k}_2 \cdot \mathbf{s})^2]. \tag{4.1}$$

Similarly, models of contrast gain control in the retina also include sensitivity to second-order features in the stimulus, with the spiking probability of the form (Schwartz et al., 2002),

$$r(\mathbf{s}) = \frac{f(\mathbf{k}_0 \cdot \mathbf{s})}{\sum_{i=1}^{M} w_i (\mathbf{k}_i \cdot \mathbf{s})^2 + \sigma^2}, \tag{4.2}$$

where the quadratic terms in the denominator scale down the gain at high contrast (in this case, however, the neuron has a nonvanishing linear filter $\mathbf{k}_0$).

A simulated model neuron showing contrast adaptation is shown in Figure 2a, featuring first- as well as second-order stimulus sensitivity. The model neuron is probed with a flickering variance stimulus, in which the variance of white noise (with a very short correlation time) is dynamically modulated by a noise process correlated over a longer timescale (Fairhall, Lewen, Bialek, & de Ruyter van Steveninck, 2001). With this synthetic stimulus, the separation of timescales allows us to partition the stimulus into chunks with approximately constant luminance variability σ_L^2. This variance is directly related to the temporal contrast, $C = \sigma_L/\bar{L}$, because the average mean light intensity $\bar{L}$ is kept constant. Within each stimulus segment, we can use STA to recover the LN model, as shown in Figures 2b and 2c. Our real goal, however, is to infer a joint model valid across the whole stimulus ensemble and ultimately to do so with naturalistic stimuli of scale-free power spectra and no clear separation between the fast fluctuations and the slow variance modulation.

We can describe these and similar examples by a generic quadratic model neuron that is sensitive to both a second-order function of the input (parameterized here by a real, symmetric matrix $\mathbf{Q}$) as well as a linear projection (parameterized by the filter $\mathbf{k}_0$):

$$r(\mathbf{s}) = f(\mathbf{k}_0 \cdot \mathbf{s}, \quad \mathbf{s}^{\mathrm{T}} \cdot \mathbf{Q} \cdot \mathbf{s}). \tag{4.3}$$

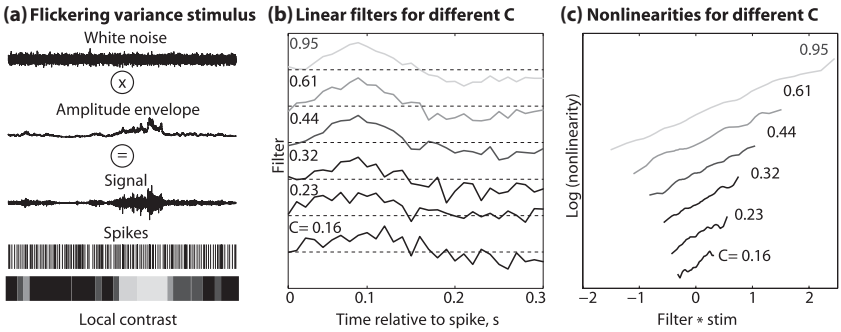

Figure 2: A synthetic contrast-adapting neuron probed with the flickering variance stimulus. The instantaneous spiking rate is given by $r(t) = f(\mathbf{k}_0 \cdot \mathbf{s}(t) + \mathbf{s}^{\mathbf{T}}(t) \cdot \mathbf{Q} \cdot \mathbf{s}(t) + \mu)$, where $f(\cdot) = \log(1 + \exp(\cdot))$, μ is an offset (bias), and the quadratic kernel $\mathbf{Q}$ is a rank 2 matrix with a quadrature pair of eigenvectors. (a) The stimulus is sampled at $\Delta = 1$ms and is given by $s(t) = \exp[A(t)]w(t)$, where $w(t)$ is an uncorrelated white noise of fixed variance and $A(t)$ is a gaussian noise process with correlation time $\tau_c = 1$s. The stimulus can be chopped into segments of duration $\tau \leq \tau_c$, which can be sorted by local contrast C (grayscale; lighter shade denotes higher contrast). Spike-triggered average analysis can be applied to recover effective LN models for all stimulus segments sharing the same local contrast. (b) The linear filters recovered at various contrast levels C (grayscale; filters are displaced along vertical axis for readability). At lower contrasts, the neuron produces fewer spikes, making the filter estimate noisier, but the shape of the filter is constant across a range of C and closely approximates the model filter $\mathbf{k}_0$. (c) The nonlinearities for different contrast levels C (grayscale; nonlinearities displaced along vertical axis for readability). The slope of the nonlinearity decreases with increasing contrast (although the adaptation is not perfect in this example) to prevent rapid saturation of the response at high C.

Graphically, while a threshold LN model with a linear filter corresponds to a classifier whose separating hyperplane is perpendicular to the filter, the proposed LN model with a threshold nonlinearity and a quadratic filter $\mathbf{Q}$ is selective for all stimuli that lie outside an N-dimensional ellipsoid whose axes correspond to the eigenvectors of $\mathbf{Q}$.

For the contrast gain control model described in equation 4.2, the matrix $\mathbf{Q}$ is of rank M, with eigenvalues w_i and eigenvectors $\mathbf{k}_i$, $i > 0$. The complex cell example described in equation 4.1 has $\mathbf{k}_0 = 0$ and $\mathbf{Q} = \sum_{i=1}^{2} \mathbf{k}_i \mathbf{k}_i^{\mathrm{T}}$; in other words, $\mathbf{Q}$ is a rank 2 matrix. While these examples feature quadratic dependences involving matrices of low rank, it is possible to extend quadratic models to biologically relevant cases where the matrix is of high rank (Rajan & Bialek, 2012). For example, the probability of spiking could

be a nonlinear function of the power $p(t)$, $r(t) = f[p(t)]$, where the power is given by

$$p(t) = \int d\tau\, f_2(\tau) \left[\int dt' f_1(t - \tau - t')\mathbf{s}(t') \right]^2. \tag{4.4}$$

Here $\mathbf{s}(t)$ is the stimulus, and f_1 and f_2 are linear filters such as those used to describe non-phase-locked auditory neurons. If the smoothing time of the second filter f_2 is larger than that of the first filter f_1, Rajan and Bialek (2012) have shown that the quadratic kernel $\mathbf{Q}$ for this model has a rich (full-rank) spectrum.

5 Inferring Quadratic Stimulus Dependence from Data

In this section we review methods that permit the inference of low- or full-rank quadratic kernels, $\mathbf{Q}$, with arbitrary stimuli.

5.1 Finding Quadratic Filters Using Information Maximization. Despite their utility and simplicity, spike-triggered methods require the use of statistically simple stimuli and, in particular, exclude the use of stimuli with naturalistic statistics (e.g., those with $1/f$ spectra, nongaussian histograms or high-order correlations). This is a big challenge when studying neurons beyond the sensory periphery that are responsible for extracting high-order structure or neurons that remain unresponsive to white noise presentations (e.g., those in the auditory pathway). To address this issue and recover filters in an unbiased manner with an arbitrary stimulus distribution, maximally informative dimensions (MID) (Kouh & Sharpee, 2009; Sharpee et al., 2004, 2006) have been developed and utilized to recover simple cell receptive fields, among other examples. MID looks for a linear filter $\mathbf{k}$ that maximizes the information between the presence or absence of a spike and the projection x of the stimulus onto $\mathbf{k}$, $x = \mathbf{k} \cdot \mathbf{s}$. Information per spike is then given by the Kullback-Leibler divergence of $P(x|\text{spike})$, the spike-triggered distribution (the distribution of stimulus projections preceding the spike), and $P(x)$, the prior distribution (the overall distribution of projections),

$$I_{\text{spike}} = D_{KL}\big[P(x|\text{spike})||P(x)\big] = \int dx\, P(x|\text{spike}) \log_2 \frac{P(x|\text{spike})}{P(x)}. \tag{5.1}$$

Given the spike train and the stimulus, finding $\mathbf{k}$ becomes an information optimization problem in I_{spike} that can be solved using various annealing methods, although the existence of local extrema could make this a nontrivial task.

Spike-triggered methods and MID do not explicitly assume a form for the nonlinearity $f(\cdot)$ in the LN model; instead, they provide unbiased

estimates of the filters, and once the filters are known, the nonlinearity can be reconstructed using the Bayes' rule from the sampled spike-triggered and prior distributions,

$$f(x) \propto P(\text{spike}|x) = \frac{P(x|\text{spike})P(\text{spike})}{P(x)}, \tag{5.2}$$

where $P(\text{spike})$ is directly proportional to the average firing rate during the experiment.

In classical MID, one finds a (set of) linear filter(s) by maximizing equation 5.1 with respect to $\mathbf{k}$. To generalize this inference method to quadratic stimulus dependence, a naive approach would make use of the spectral decomposition in equation 3.1. One would try recovering the quadratic dependence of $\mathbf{Q}$ in equation 4.3 by multidimensional MID (see Table 1), hoping to infer all $\{\mathbf{k}_i\}$ as orthogonal informative dimensions. While formally true, this is infeasible in practice because maximizing the mutual information would involve sampling N-dimensional distributions from stimulus samples that are limited in number by the number of spikes (Sharpee et al., 2004). Information-theoretic STC (iSTAC) evades this problem, but at the cost of going back to gaussian stimuli and assuming a gaussian spike-triggered distribution; under those restrictions, it can be used to infer quadratic stimulus dependence (Pillow & Simoncelli, 2006).

To address this problem efficiently without imposing restrictions on the prior and spike-triggered ensembles, the inference problem can be formulated by assuming quadratic dependence on the stimulus from the start, as proposed in Rajan and Bialek (2012). A quadratic filter $\mathbf{Q}$ can be reconstructed from an observed spike train by maximizing the information in equation 5.1, where x is now given by $x = \mathbf{s}^{\mathrm{T}} \cdot \mathbf{Q} \cdot \mathbf{s}$. Taking a derivative of equation 5.1 with respect to $\mathbf{Q}$ gives us a gradient,

$$\nabla_{\mathbf{Q}} I = \int dx\, P_{\mathbf{Q}}(x) \left[\left\langle \mathbf{s}\mathbf{s}^{\mathrm{T}} | x, \text{spike} \right\rangle - \left\langle \mathbf{s}\mathbf{s}^{\mathrm{T}} | x \right\rangle \right] \frac{d}{dx} \left(\frac{P_{\mathbf{Q}}(x|\text{spike})}{P_{\mathbf{Q}}(x)} \right), \tag{5.3}$$

where $\langle \cdot \rangle$ indicates averaging over the spike-triggered and prior distributions, respectively, and the subscript $\mathbf{Q}$ makes the dependence of the probability distributions explicit. Only the symmetric part of $\mathbf{Q}$ contributes to x, and the overall scale of the matrix is irrelevant to the information, making the number of free parameters $N(N+1)/2 - 1$. This approach makes the inference problem tractable even when $\mathbf{Q}$ is of high rank.

To learn the maximally informative stimulus energy (MISE) or the quadratic filter $\mathbf{Q}$, we can ascend the gradient in successive learning steps (Rajan & Bialek, 2012),

$$\mathbf{Q} \to \mathbf{Q} + \gamma\, \nabla_{\mathbf{Q}} I. \tag{5.4}$$

The probability distributions within the gradient are obtained by computing x for all stimuli, choosing an appropriate binning for the variable x, and sampling binned versions of the spike-triggered and prior distributions. The $\langle \mathbf{s}\mathbf{s}^{\mathrm{T}} \rangle$ averages are computed separately for each bin, and the integral in equations 5.1 and 5.3 and the derivative in equation 5.3 are approximated as a sum over bins and as a finite difference, respectively. To deal with local maxima in the objective function, we can use a large starting value of γ and gradually decrease γ during learning. This basic algorithm can be extended by using kernel density estimation and stochastic gradient ascent and annealing methods, but we do not report these technical improvements here.

It is also possible to select an approximate linear basis in which to expand the matrix $\mathbf{Q}$ by writing

$$\mathbf{Q} = \sum_{\mu=1}^{M} \alpha_\mu \mathbf{B}^{(\mu)}. \tag{5.5}$$

The basis can be chosen so that increasing the number of basis components M allows the reconstruction of progressively finer features in $\mathbf{Q}$. We considered as $\{\mathbf{B}^{(\mu)}\}$ a family of gaussian bumps that tile the space of the $N \times N$ matrix $\mathbf{Q}$ and whose scale (standard deviation) is inversely proportional to $\sqrt{M}$. For $M \to N^2/2$, the matrix set becomes a complete basis, allowing every $\mathbf{Q}$ to be exactly represented by the vector of coefficients α. In such a matrix basis representation, the learning rule becomes

$$\alpha_\mu \to \alpha_\mu + \gamma \sum_{i,j=1}^{N} \frac{\partial I}{\partial \mathbf{Q}_{ij}} \mathbf{B}_{ij}^{(\mu)}, \tag{5.6}$$

where applying the chain rule on $\nabla_{\mathbf{Q}} I$ yields the Trace$[\nabla_{\mathbf{Q}}(\alpha) \cdot \mathbf{B}]$ update term for each step.

We illustrate this approach with two examples. In the first example, we make use of the matrix basis expansion from equation 5.5 to infer a quadratic kernel $\mathbf{K}$ that is of high rank. For $\mathbf{K}$, we used a highly structured 500×500 matrix as shown in Figure 3a. While this is not an example of a receptive field from a real neuron, it illustrates the validity of the approach even when the response has an atypical and highly structured dependence on the stimulus. The stimuli were natural image clips from the Penn Natural Image database, flattened into a high-dimensional vector representation $\mathbf{s}$ (Tkačik et al., 2011), and the spikes were generated by thresholding the term $\mathbf{s}^{\mathrm{T}} \cdot \mathbf{K} \cdot \mathbf{s}$. Gaussian basis matrices, similar to the 225 shown in Figure 3b, were used to expand the quadratic kernel, reducing the number of free parameters from approximately 2.5×10^5 to a few hundred. We start the gradient ascent with a large γ value of 1 and progressively scale it down to 0.1 near the end of the algorithm; Figure 3e shows the information

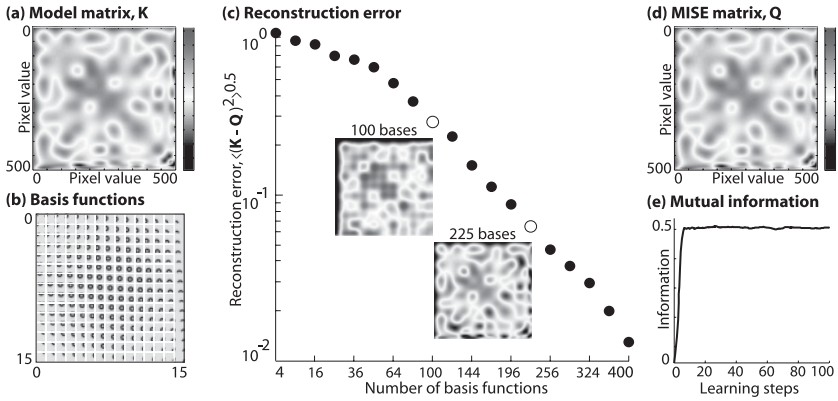

Figure 3: Reconstructing a high-rank quadratic filter using stimuli extracted from natural scenes. (a) A complex high-rank matrix $\mathbf{K}$ is used as a quadratic filter for a model neuron that fires whenever $\mathbf{s}^{\mathrm{T}} \cdot \mathbf{K} \cdot \mathbf{s}$ exceeds a fixed threshold. $\mathbf{K}$ is thus the "true" quadratic filter for our threshold quadratic model neuron. (b) The collection of 225 gaussian matrix basis functions whose peaks densely tile the matrix space is shown here. A trial matrix is constructed as a linear sum (with coefficients $\{\alpha_\mu\}$) of the basis matrices, and information optimization is performed over $\{\alpha_\mu\}$. (c) Normalized reconstruction error, shown by the filled circle, decreases as the number of basis functions M increases from 4 to 400. With enough data, perfect reconstruction is possible as the number of basis functions M approaches the number of independent pixels in $\mathbf{K}$. The two open circles show reconstructions with $M = 100$ or $M = 225$ basis functions, respectively. (d) The maximally informative stimulus energy $\mathbf{Q}$ after maximizing mutual information using 400 basis functions is shown here. (e) Mutual information increases as learning progresses according to equation 5.4, peaks at the 40th step, and remains unchanged thereafter. Learning step 100 is the point where the maximally informative $\mathbf{Q}$ is extracted and plotted in panel d.

plateauing in about 20 learning steps. The maximally informative quadratic filter reconstructed from 400 basis coefficients is shown in Figure 3d. Figure 3c demonstrates how the root-mean-squared reconstruction error systematically decreases as the number of basis functions M is increased from 4 to 400, improving precision. Insets show two inferred matrices with $M = 100$, corresponding to the first open circle, showing a marked improvement with $M = 225$, corresponding to the second open circle. Reconstruction error drops to approximately 1% for $M = 400$.

In contrast to standard MID where the number of spikes required grows exponentially in the number of filters extracted, the data requirement for this approach is proportional to the square of the stimulus dimension for a matrix kernel with no additional structural simplifications (these data requirement- and performance-related issues are explored in detail in Rajan

& Bialek, 2012). For the examples shown in this review, expansion in matrix basis reduces this number to the order of stimulus dimension, making this procedure pertinent to experimentalists.

The second example shows the MISE analysis of the synthetic neuron presented in Figure 2 where the stimulus-response relationship is more biologically realistic. Here, a smooth nonlinear function f is used, and the model has a linear and a quadratic kernel. The analysis is applied to the flickering variance stimulus without partitioning it into regions of fixed contrast. With $\sim 2 \times 10^4$ spikes, the method recovers the linear filter $\mathbf{k}_0$ as well as the quadratic kernel, which turns out to have the two dominant eigenvectors, $\mathbf{k}_1$ and $\mathbf{k}_2$, corresponding to the quadrature pair of filters used to construct $\mathbf{Q}$, as shown in Figure 4b.

These examples show that quadratic filters can be extracted using information maximization for both low-rank and full-rank matrices, under natural stimulation and with realistic numbers of spikes. Importantly, for cases where the stimulus sensitivity is both linear and quadratic, MISE does not explicitly assume that the collective effect of two filtering operations is necessarily additive, that is, that $x = \mathbf{k}_0 \cdot \mathbf{s} + \mathbf{s}^{\mathrm{T}} \cdot \mathbf{Q} \cdot \mathbf{s}$; rather, the dependence can be an arbitrary 2D nonlinear function, $f(\mathbf{k}_0 \cdot \mathbf{s}, \mathbf{s}^{\mathrm{T}} \cdot \mathbf{Q} \cdot \mathbf{s})$. Unlike the quadratic generalizations of GLM presented below, this allows MISE to fully recover forms of contrast gain control that have a parametric form similar to equation 4.2.

5.2 Finding Quadratic Filters Using Maximization of Noise Entropy. Another information-theoretic approach for inferring single-neuron sensitivities is derived from the principle of noise entropy maximization (Fitzgerald, Rowekamp et al., 2011; Fitzgerald, Sincich et al., 2011; Globerson et al., 2009). Suppose that the spiking or silence of a chosen neuron in a time bin indexed by t is represented by a binary variable $y_t \in \{0, 1\}$. From data, we can reliably estimate certain statistics of the neural response, such as the average spiking rate $\langle y_t \rangle_t$, the spike-triggered average $\langle y_t \mathbf{s}(t) \rangle_t$, or the spike-triggered covariance $\langle y_t \mathbf{s}(t) \mathbf{s}(t)^{\mathrm{T}} \rangle_t$, where the brackets $\langle \cdot \rangle_t$ denote averaging across the duration of the experiment. In general, all of these statistics are of the form $\langle O_\mu(\mathbf{s}) y_t \rangle_t$, where μ indexes the different operators whose expectation values we are computing.

The crucial step is to look for maximum entropy approximations to $P(y|\mathbf{s})$, the distribution of the (binary) neural response given the stimulus. Maximum entropy distributions are as unstructured (random, therefore parsimonious) as possible with the constraint that they exactly reproduce the measured expectation values of a chosen set of statistics, $\{O_\mu\}$ (Jaynes, 1957a, 1957b). When the variable y is binary, it can easily be shown that these distributions have the form of the logistic function,

$$P(y = 1|\mathbf{s}) = \frac{1}{1 + e^{-F(\mathbf{s})}}, \tag{5.7}$$

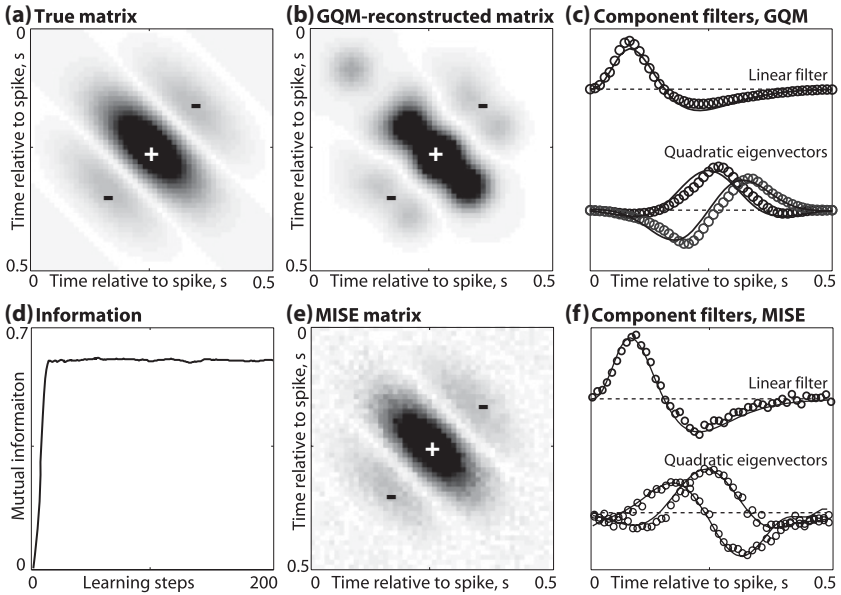

Figure 4: Recovering the synthetic model of the contrast gain control neuron using a flickering variance stimulus. The spikes were simulated using the same model presented in Figure 2. (a) The true quadratic kernel of the model is a matrix of rank 2 with the two filters combining in quadrature to estimate the signal power or variance. (b) The reconstructed kernel using the quadratic extension of the GLM. The space of matrices was spanned by an 85-dimensional basis of gaussian bumps (some resulting granularity is obvious in the reconstruction). The dominant eigenvectors of the inferred matrix are plotted in (c) with circles (solid black lines show the true values) along with the recovered linear filter (circles) and its true value (solid black line). Inferring of the same model using MISE shows convergence in (d) and the recovered quadratic kernel is plotted in (e). (f) The linear filter and the eigenvectors of the quadratic kernel recovered with MISE (circles), compared to the true values (black solid line). Note that quadratic filter eigenvectors are only determined up to a sign.

where F resembles the free energy in statistical physics,

$$F(\mathbf{s}) = \sum_{\mu} \lambda_{\mu} O_{\mu}(\mathbf{s}), \tag{5.8}$$

and λ_{μ} are the Lagrange multipliers that have to be set such that the set of statistics measured in the data equals the expectation values of the same operators under distribution P, that is, $\langle O_{\mu}(\mathbf{s})y\rangle_{P} = \langle O_{\mu}(\mathbf{s})y\rangle_{t}$. To apply this general framework to the inference of quadratic filters, Fitzgerald,

Rowekamp et al. (2011) choose the mean firing rate, STA and STC as constraints, which yields the following response distribution:

$$P(y=1|\mathbf{s}) = \frac{1}{1+\exp(\mu + \mathbf{k}_0 \cdot \mathbf{s} + \mathbf{s}^{\mathrm{T}} \cdot \mathbf{Q} \cdot \mathbf{s})}, \tag{5.9}$$

where $\{\mu, \mathbf{k}_0, \mathbf{Q}\}$ act as the Lagrange multipliers λ_μ conjugated to the operators $\{y, y\mathbf{s}, y\mathbf{s}\mathbf{s}^{\mathrm{T}}\}$. Numerically, the task is to solve for parameters $\{\mu, \mathbf{k}_0, \mathbf{Q}\}$ that satisfy a set of constraints: $\langle y \rangle_t = \langle y \rangle_P$ (matching the measured mean firing rate to that of the model), $\langle y\mathbf{s} \rangle_t = \langle y\mathbf{s} \rangle_P$ (matching the measured STA to that of the model), and $\langle y\mathbf{s}\mathbf{s}^{\mathrm{T}} \rangle_t = \langle y\mathbf{s}\mathbf{s}^{\mathrm{T}} \rangle_P$ (matching the measured STC to that of the model). This is a convex optimization task and can be solved by conjugate gradient descent.

An attractive feature of this approach emerges when we rewrite the information per spike $I(\text{spike}; \mathbf{s})$ as a difference between the total and the noise entropy,

$$\begin{aligned} I(\text{spike}; \mathbf{s}) &= \sum_{\mathbf{s}} P(\mathbf{s}) \sum_{y} P(y|\mathbf{s}) \log_2 \frac{P(y|\mathbf{s})}{P(y)} \\ &= S[P(y)] - \langle S[P(y|\mathbf{s})] \rangle_{\mathbf{s}}, \end{aligned} \tag{5.10}$$

where $S[P(x)] = -\sum_x P(x) \log_2 P(x)$ is the entropy of $P(x)$. The first term (total entropy) is fully determined by the probability of spiking $\langle y \rangle_t$, $S[P(y)] = -\langle y \rangle_t \log_2 \langle y \rangle_t - (1 - \langle y \rangle_t) \log_2 (1 - \langle y \rangle_t)$, because y is a binary variable. The mean firing rate is one of the statistics constrained in the model for $P(y|\mathbf{s})$, ensuring consistency. Since our model for $P(y|\mathbf{s})$ has maximum entropy given the observed constraints, we are effectively setting an upper bound on the noise entropy $\langle S[P(y|\mathbf{s})] \rangle_{\mathbf{s}}$ and therefore a lower bound on the mutual information I. As more and more statistics $O(\mathbf{s})$ are included as constraints into the maximum entropy model for equation 5.7, the noise entropy must progressively drop and information must increase toward the true value (which is bounded by the output entropy). At the point where this lower bound on information meets the actual information per spike (which can be empirically estimated from repeated stimulation; see, e.g., Brenner et al., 2000), we obtain the complete set of relevant stimulus statistics $\{O_\mu\}$ that characterize the sensitivity of the neuron.

Fitzgerald, Rowekamp et al. (2011) show that this framework is applicable for inferring quadratic neural filters on synthetic and real data and compare it to MID. This method is applicable to any stimulus ensemble but requires assumptions beyond those needed for MID or MISE—namely, that the nonlinear function is logistic and that the contributions of the linear and quadratic filters combine additively. The advantage of this method is that the optimization problem remains convex, does not suffer from the exponential curse of dimensionality (like multidimensional MID), and is flexible,

allowing different constraints (beyond the STA and STC) to be used for constructing models of the stimulus-conditional distribution $P(y|\mathbf{s})$.

5.3 Finding Quadratic Filters in a Likelihood Framework: Extensions to GLM and Bayesian STC. A powerful technique for modeling neural spiking behavior is the generalized linear model (GLM) framework (Paninski, 2004; Truccolo et al., 2004). Recently, GLM has been used to account for the stimulus sensitivity, dependence on spiking history, and connectivity in a population of 27 retinal ganglion cells in the macaque retina (Pillow et al., 2008). For a single neuron, the model assumes that the instantaneous spiking rate $r(t)$ is a nonlinear function f of a sum of contributions,

$$r(t) = f\left[\mathbf{k} \cdot \mathbf{s}(t) + \mathbf{q} \cdot \mathbf{y}(t\_) + \mu\right], \tag{5.11}$$

where $\mathbf{k}$ is a linear filter acting on the stimulus $\mathbf{s}$, $\mathbf{q}$ is a linear filter acting on the spiking history $\mathbf{y}(t\_)$ of the neuron, and μ is an offset or an intrinsic bias toward spiking or silence. When the stimulus and the spike train are discretized into time bins of duration Δ, the probability of observing (an integral number of) y_t spikes is Poisson, with a mean given by $r_t\Delta$ (where the subscript indexes the time bin). Here, we neglect the history dependence of the spikes (with no loss of generality) and focus instead on the stimulus dependence; since each time bin is conditionally independent given the stimulus (and past spiking), the log likelihood for any spike train $\{y_t\}$ is (Pillow, 2007),

$$\log P(\{y_t\}|\mathbf{s}) = \sum_t y_t \log r_t - \Delta \sum_t r_t + c, \tag{5.12}$$

where c is independent of both μ and $\mathbf{k}$. This likelihood can be maximized with respect to μ and $\mathbf{k}$ (and, optionally, with respect to $\mathbf{g}$) given adequate spikes, providing an estimate of the filters from neural responses to complex, even natural stimuli. In contrast to maximally informative approaches, such as the stimulus energy derived in section 5.1 (Rajan & Bialek, 2012), the functional form of the nonlinearity f is an explicit assumption in likelihood-based methods like GLM. For specific classes of the function f, such as $f(z) = \log[1 + \exp(z)]$, $\exp(z)$, or $\lfloor z \rfloor$, the likelihood optimization problem is convex, and gradient ascent is guaranteed to find a unique global maximum.

While the tractability consequent to convexity of the objective function is a big strength of this approach, the disadvantage is that if the chosen nonlinearity f is different from the true function f' used by the neuron, the filters inferred by maximizing likelihood in equation 5.12 could be biased. If we relax the stringent requirement for convexity, we can choose more general nonlinear functions for the model, for example, by parameterizing

the nonlinearity in a point-wise fashion and inferring it jointly with the filters. For this discussion, however, we assume that f has been selected from the specific class of nonlinearities guaranteed to yield a convex likelihood function.

How can we extend GLM to situations where the neuron's response is more complex than a single linear projection of the stimulus? We start with a proposal and follow up with a review of Park and Pillow (2011), which has provided a fuller analysis and several interesting extensions. One possibility is to expand the stimulus clip $\mathbf{s}$ of dimension N into a larger space first, for instance, by forming $\mathbf{s}\mathbf{s}^{\mathrm{T}}$ (of dimension $N \times N$), and then operate on this object with a filter: $\sum_{i,j=1}^{N}(s_i s_j)Q_{ij}$. Such a term can be added to the argument of f in the model exemplified in equation 5.11. Specifically, we propose a generalized quadratic model of the following form,

$$r(t) = f[\mathbf{k}\cdot\mathbf{s}(t) + \mathbf{s}^{\mathrm{T}}(t)\cdot\mathbf{Q}\cdot\mathbf{s}(t) + \mathbf{g}\cdot\mathbf{y}(t_{\_}) + \mu]. \tag{5.13}$$

If we want to retain convexity, we cannot simply expand $\mathbf{Q}$ in its eigenbasis and infer its vectors by maximizing the likelihood directly, because the eigenvectors appear quadratically. However, we can expand $\mathbf{Q}$ into a weighted sum of matrix basis functions, as in equation 5.5, making the argument of f a linear function of basis coefficients α_μ,

$$r(t) = f\left(\mathbf{k}\cdot\mathbf{s}(t) + \sum_{\mu=1}^{M}\left[\mathbf{s}^{\mathrm{T}}(t)\cdot\mathbf{B}^{(\mu)}\cdot\mathbf{s}(t)\right]\alpha_\mu + \mathbf{g}\cdot\mathbf{y}(t_{\_}) + \mu\right). \tag{5.14}$$

Existing methods for inferring GLM parameters (Pillow et al., 2008) can be used to learn the linear and the quadratic filter $\mathbf{Q}$ efficiently. After extracting $\mathbf{Q}$, we can check if a few principal components account for most of its structure (this is equivalent to checking whether $\mathbf{Q}$ is indeed a low-rank matrix). To summarize, this procedure provides a way of extracting multiple filters with GLM that is analogous to diagonalizing the spike-triggered covariance matrix on the gaussian stimulus ensemble.

We have implemented such a generalized quadratic model using the flickering variance stimulus shown in Figure 2. The results are shown in Figure 4a. The recovered quadratic kernel decomposes into a quadrature pair of filters, and we recover the correct linear filter $\mathbf{k}_0$. While this method is restricted to a linear combination of first- and second-order filters within the nonlinearity, the distinct advantage over MISE is that the inference problem is always convex with the appropriate choice of nonlinear functions.

Park and Pillow (2011) consider an exponentiated general quadratic function of the following form (rewritten in the notation of this review),

$$r(\mathbf{s}) = \exp(\mathbf{s}^{\mathrm{T}}\cdot\mathbf{Q}\cdot\mathbf{s} + \mathbf{k}_0\cdot\mathbf{s} + \mu). \tag{5.15}$$

They first show that under a gaussian stimulus ensemble, the expected log likelihood can be expressed in terms of the STA, STC, and the covariance matrix of the stimulus and derive the closed-form expressions for maximum likelihood estimates of the quadratic kernel, linear filter, and bias term. Next, the generalization to arbitrary stimuli is achieved by numerically optimizing the true (as opposed to expected) likelihood. In contrast to our suggestion of using the matrix basis expansion (which becomes an implicit regularizer on choosing the dimensionality of the basis), Park and Pillow implement a Bayesian regularizer by imposing a prior on the quadratic kernel as follows.

The matrix is first decomposed into the eigensystem, $\mathbf{Q} = \sum_{i=1}^{N} \sigma_i \mathbf{w}_i \mathbf{w}_i^{\mathrm{T}}$, where $\mathbf{w}$ are not forced to have an L_2 norm of 1 and $\sigma_i = \pm 1$ to indicate whether the filter i is excitatory or suppressive (as in STC; Schwartz et al., 2006). Then a zero-mean gaussian prior $\mathcal{N}(0, \alpha_i^{-1} I)$ is put on each eigenvector $\mathbf{w}_i$, where the hyperparameter α_i determines the variance of the elements of eigenvector i ($\alpha_i \to \infty$ corresponds to eliminating the direction i from the quadratic kernel and reducing its rank by 1). Next, an iterative algorithm alternates between optimizing the likelihood with respect to model parameters and optimizing the evidence given the parameters with respect to α_i. This procedure efficiently and accurately identifies the rank of the quadratic kernels in synthetic examples, providing an automatic alternative for distinguishing the significant from sampling-noise-induced eigenvectors in the STC and quadratic kernel inference. Finally, the authors show that equation 5.15 can be further generalized from the exponentiated quadratic function to a wider class of elliptic nonlinearities at no additional computational cost.

To summarize this section, the reviewed work shows that Bayesian generalization of STC and the generalization of GLMs to quadratic stimulus dependence yield equal probabilistic models for neural encoding that can be efficiently inferred for a restricted class of nonlinear functions. However, attention needs to paid to maintain the convexity of the optimization procedure and deal with the large number of free parameters in the quadratic kernel. Basis expansions as well as regularization with Bayesian priors seem like feasible candidates to this end.

6 Discussion

While powerful conceptually, the notion that neurons respond to multiple projections of the stimulus onto orthogonal filters is difficult to turn into a tractable inference procedure when the number of filters is large. To address this concern, alternative encoding models have been proposed where the neuron can be sensitive to high-order features in the stimulus. Instead of being described by multiple linear filters, the neuron's sensitivity properties are captured by a single quadratic filter (and, optionally, an additional linear filter). We have reviewed several inference methods for

such quadratic stimulus dependence based on information maximization as well as maximizing likelihood. With MISE, no assumptions are made about how the projection onto the quadratic filter combines with the linear filter projection and how both map onto the spiking probability. This approach yields unbiased filter estimates under any stimulus ensemble but requires optimization in a rugged information landscape. Noise entropy maximization is a flexible, maximum-entropy-based framework for modeling the probability of spiking given a stimulus. It is computationally tractable and provides a convenient bound on the information per spike but assumes a specific form of the nonlinearity. Alternatively, with a specific choice of nonlinearity and filter basis, likelihood inference within the GLM class can be extended to quadratic stimulus dependence while retaining the convexity of the objective function. By formulating the problem as Bayesian inference and choosing sparsifying priors for the quadratic filter, the true rank of the quadratic filter can also be inferred from data.

All of these approaches for inferring quadratic stimulus dependence are complementary; as we show in the appendix, both information-maximization- and maximum-likelihood-based inference methods provide consistent filter estimates under defined conditions. A possible way to benefit from the tractability of likelihood formulations and maximization of noise entropy could be to use them to initialize a more general search using information maximization. This could potentially help avoid optimization problems in the rugged information landscape and remove the additive restrictions on the combination of linear and quadratic features.

Examples of recent work establishing the connection between the high-order structure of natural scenes and neural mechanisms beyond the sensory periphery (Karklin & Lewicki, 2009; Tkačik, Prentice, Victor, & Balasubramanian, 2010) make the development of methods for neural characterization such as the ones presented here very timely. Phenomena like phase invariance, adaptation to local contrast, or sensitivity to the signal envelope are widespread features of sensory neural responses (Baccus & Meister, 2004; Hubel & Wiesel, 1965; Touryan, Lau, & Dan, 2002). Moreover, as our abilities to record in vivo from the sensory systems of awake and behaving animals expand, so should the methods to analyze such recordings, where the relevant stimuli may no longer be perfectly controllable because of the animal's interaction with the environment (Kerr & Nimmerjahn, 2012). The methods presented here will help us systematically elucidate sensitivity to high-order statistical features from responses of sensory neurons to natural stimuli.

Appendix: Formal Relationship Between Information-Theoretic and Likelihood-Based Inference

We now demonstrate analytically that under rather general assumptions, the linear or quadratic filters obtained by maximizing mutual information

match the filters inferred by maximizing likelihood. We extend a reasoning we used previously in the context of inferring protein-DNA sequence-specific interactions in Kinney, Tkačik, and Callan (2007), to neural responses (see also Kouh & Sharpee, 2009; Williamson, Sahani, & Pillow, 2011).

In the following, x remains the projection of the stimulus $\mathbf{s}$ onto the linear ($x_t = \mathbf{k} \cdot \mathbf{s}_t$) or quadratic ($x_t = \mathbf{s}_t^{\mathrm{T}} \cdot \mathbf{Q} \cdot \mathbf{s}_t$) filter, with time discretized in bins of duration Δ and indexed by subscript t. We collect all the parameters describing the filter into a vector θ_1. Given a single x_t, y_t spikes are generated according to a conditional probability distribution $\pi(y_t|x_t)$. This probability distribution is typically assumed to be Poisson with a mean given by $f(x_t)$ in the case of GLM, but we take a different approach.

We discretize x_t into $x = 1, \ldots, K$ bins and parameterize $\pi(y_t|x_t)$, which is a $Y_{\max} \times K$ matrix, by a vector θ_2. Apart from assuming a cutoff value for the number of spikes per bin $Y_{\max}$ (which can always be chosen large enough to assign an arbitrarily low probability of observing $>Y_{\max}$ spikes in any real data set) and a particular discretization of the projection variable x, we leave the probabilistic relationship $\pi(y|x)$ between the projection and spike count completely unconstrained. The transformation from the stimulus to the spikes is then a Markov chain, fully specified by $\theta = \{\theta_1, \theta_2\}$,

$$\mathbf{s}_t \xrightarrow[\mathbf{k} \text{ or } \mathbf{Q}]{\theta_1} x_t \xrightarrow[\pi]{\theta_2} y_t. \tag{A.1}$$

The likelihood of the spike train $\{y_t\}$ given the stimulus $\mathbf{s}$ is $P(\{y_t\}|\mathbf{s}) = \prod_{t=1}^{T} \pi(y_t|x_t)$, where T is the total number of time bins in the data set. With x discretized into K bins, any data set can be summarized by the count matrix $c_{yx} = \sum_{t=1}^{T} \delta(y, y_t)\delta(x, x_t)$, where δ is the Kronecker delta. Note that $c_{yx} = T\tilde{p}(y, x)$, where $\tilde{p}$ is simply the empirical distribution of the probability of observing y spikes jointly with the projection x. In terms of c, the likelihood of the observed spike train is $P(\{y_t\}|\mathbf{s}) = \prod_{y=0}^{Y_{\max}} \prod_{x=1}^{K} \pi(y|x)^{c_{yx}}$. Assuming that x is adequately discretized and π is Poisson with mean $f(x)$, we will recover the generalized likelihood of equation 5.12.

Suppose we are only interested in inferring the filter (parameterized by θ_1) but not the filter-to-spike mapping π (parameterized by θ_2). While avoiding any explicit assumptions about the structure of π, we can integrate the likelihood over θ_2 with some prior $P_p(\theta_2)$ such that

$$P(\{y_t\}|\mathbf{s}) = \int d\theta_2 \, P_p(\theta_2) \prod_{y,x} \pi(y|x)^{c_{yx}}. \tag{A.2}$$

This resulting likelihood, called the model-averaged likelihood, is now only a function of θ_1. The prior $P_p(\theta_2)$ can take many forms, but since we

discretized x, thereby making $\pi(y|x)$ into a (conditional probability) matrix, the simplest choice is the uniform prior. In this case, we set θ_2 equal to the entries in the $\pi(y|x)$ matrix and choose $P(\theta_2)$ to be uniform over all valid matrices π, such that the matrix entries are positive and the normalization constraint, $\sum_y \pi(y|x) = 1$ for every x, is enforced.

For any choice of priors, we can rewrite equation A.2 as

$$P(\{y_t\}|\mathbf{s}) = \int d\theta_2 \; P_p(\theta_2) \exp\left[T \sum_{y,x} \tilde{p}(y,x) \log \pi(y|x) \right], \tag{A.3}$$

which can be reorganized into

$$P(\{y_t\}|\mathbf{s}) = \int d\theta_2 P_p(\theta_2) \exp\big[T\big\{\tilde{I}(y;x) - \tilde{S}(y) - \langle D_{KL}(\tilde{p}(y|x) \,||\, \pi(y|x))\rangle_{\tilde{p}(x)}\big\}\big]. \tag{A.4}$$

Here $\tilde{I}(y;x) = \sum_{y,x} \tilde{p}(y,x) \log \frac{\tilde{p}(y,x)}{\tilde{p}(y)\tilde{p}(x)}$ is the empirical mutual information between spike counts y and the projection x, $\tilde{S}(y)$ is the empirical spike count entropy, and the "correction" term in brackets measures the average Kullback-Leibler divergence (D_{KL}) between the empirical and model conditional distributions. Importantly, only this correction term is a function of the π and thus of θ_2 and is affected by the prior $P_p(\theta_2)$, which is being integrated over; the other terms can therefore be pulled outside the integral. We can write the log likelihood per time bin as

$$\mathcal{L} = \frac{1}{T} \log P(\{y_t\}|\mathbf{s}) = \tilde{I}(y;x) - \tilde{S}(y) - \Lambda, \tag{A.5}$$

where the correction is

$$\Lambda = -\frac{1}{T} \log \int d\theta_2 \; P_p(\theta_2) e^{-T\langle D_{KL}(\tilde{p}(y|x) \,||\, \pi(y|x))\rangle_{\tilde{p}(x)}}. \tag{A.6}$$

It is necessary to show that as the number of data T grows, the correction Λ decreases for a given choice of prior distribution $P_p(\theta_2)$, and for the choice of uniform prior this is analytically tractable (Kinney et al., 2007). Intuitively, it is clear that as $T \to \infty$, the empirical distribution $\tilde{p}(y|x)$ converges to the true underlying distribution $p(y|x)$, and the integral becomes dominated by the extremal point θ_2^*, such that, in the saddle point approximation,

$$\Lambda(T \to \infty) \sim \langle D_{KL}(p(y|x) \,||\, \pi^*(y|x))\rangle_{p(x)}. \tag{A.7}$$

The distribution $\pi^*(y|x)$ is the closest distribution to $p(y|x)$ in the space over which the prior $P_p(\theta_2)$ is nonzero. As long as the prior assigns a nonzero probability to any (normalized) distribution, the divergence in Λ will decrease and Λ will vanish as T grows. The case in which Λ does not decay occurs when the prior completely excludes certain distributions by assigning a zero probability to them, while the data $p(y|x)$ precisely favor those excluded distributions.

Returning to the log likelihood per time bin $\mathcal{L}$ in equation A.5, as we decrease the time bin Δ, we enter a regime where there is only 0 or 1 spike per bin, that is, $y \in \{0, 1\}$. Then the empirical information per time bin $\tilde{I}(y; x)$ can be written as

$$\begin{aligned}\tilde{I}(y; x) = {} & \tilde{p}(y = 0) D_{KL}(\tilde{p}(x|y = 0)||\tilde{p}(x)) \\ & + \tilde{p}(y = 1) D_{KL}(\tilde{p}(x|y = 1)||\tilde{p}(x)),\end{aligned} \tag{A.8}$$

that is,

$$\tilde{I}(y; x) = \tilde{p}(\text{silence})\tilde{I}_{\text{silence}} + \tilde{p}(\text{spike})\tilde{I}_{\text{spike}}. \tag{A.9}$$

If the information in the spike train is dominated by the information carried in spikes (Brenner et al., 2000), then the likelihood from equation A.5 becomes

$$\mathcal{L} = \tilde{p}(\text{spike})\tilde{I}_{\text{spike}} + \ldots, \tag{A.10}$$

where $\ldots$ are terms that either do not depend on the filter parameters θ_1 (i.e., entropy of the spike counts $\tilde{S}(y)$) or vanish entirely as the size of data set grows (Λ).

The identity in equation A.10 is the sought-after connection between the inference using information maximization and the likelihood-based approach. In the limit of small time bins, maximizing the information per spike I_{spike} on the right-hand side of the identity (in maximally informative approaches, as in Rajan and Bialek (2012), Sharpee et al. (2004), and section 5.1 of this review) is the same as maximizing the model-averaged likelihood $\mathcal{L}$ of equation A.5 on the left-hand side of the identity.

Acknowledgments

We thank William Bialek and Michael J Berry II for insightful discussions and providing critical scientific input during the course of this project. We especially thank Jonathan Victor for helpful comments on the manuscript. This work was supported in part by the Human Frontiers Science Program,

the Swartz Foundation, NSF grants PHY–0957573 and CCF–0939370, the WM Keck Foundation, and the ANR grant OPTIMA.

References

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

Agüera y Arcas, B., Fairhall, A. L., & Bialek, W. (2003). Computation in a single neuron: Hodgkin and Huxley revisited. *Neural Comput.*, *15*, 1715–1749.

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

Baccus, S. A., & Meister, M. (2004). Retina versus cortex: Contrast adaptation in parallel visual pathways. *Neuron*, *42*, 5–7.

Berry, M. J. II, Brivanlou, I. H., Jordan, T. A., & Meister, M. (1999). Anticipation of moving stimuli by the retina. *Nature*, *398*, 334–338.

Berry, M. J. II, & Meister, M. (1998). Refractoriness and neural precision. *J. Neurosci.*, *18*, 2200–2211.

Bialek, W., & de Ruyter van Steveninck, R. R. (2005). *Features and dimensions: Motion estimation in fly vision*. arxiv.org:q-bio/0505003

Bölinger, D., & Gollisch, T. (2012). Closed-loop measurements of iso-response stimuli reveal dynamic nonlinear stimulus integration in the retina. *Neuron*, *73*, 333–346.

Borst, A., & Egelhaaf, M. (1987). Temporal modulation of luminance adapts time constant of fly movement detectors. *Biol. Cybern.*, *56*, 209–215.

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

Chen, E., Marre, O., Fisher, C., Schwartz, G., Levy, J., da Silveira, R., et al. (2012). Alert response to motion onset in the retina. *J. Neurosci.*, *33*, 120–132.

Dayan, P., & Abbott, L. F. (2001). *Theoretical neuroscience*. Cambridge, MA: MIT Press.

de Boer, E., & Kuyper, P. (1968). Triggered correlation. *IEEE Trans. Biomed. Eng.*, *15*, 169–179.

Demb, J. B., Zaghloul, K., Haarsma, L., & Sterling, P. (2001). Bipolar cells contribute to nonlinear spatial summation in the brisk-transient (Y) ganglion cell in mammalian retina. *J. Neurosci.*, *21*, 7447–7454.

de Ruyter van Steveninck, R. R., & Bialek, W. (1988). Real-time performance of a movement sensitive in the blowfly visual system: Information transfer in short spike sequences. *Proc. Roy. Soc. Lond. B*, *234*, 379–414.

de Ruyter van Steveninck, R. R., Lewen, G. D., Strong, S. P., & Bialek, W. (1997). Reproducibility and variability in neural spike trains. *Science*, *275*, 1805–1808.

de Ruyter van Steveninck, R. R., Zaagman, W. H., & Mastebroek, H.A.K. (1986). Adaptation of transient responses of a movement-sensitive neuron in the visual system of the blowfly *Calliphora erythrocephala*. *Biol Cybern*, *54*, 223–226.

Fairhall, A. L., Burlingame, C. A., Narasimhan, R., Harris, R. A., Puchalla, J. L., & Berry, M. J. II (2006). Selectivity for multiple stimulus features in retinal ganglion cells. *J. Neurophysiol.*, *96*, 2724–2738.

Fairhall, A. L., Lewen, G. D., Bialek, W., & de Ruyter van Steveninck, R. R. (2001). Efficiency and ambiguity in an adaptive neural code. *Nature*, *412*, 787–792.

Fitzgerald, J. D., Rowekamp, R. J., Sincich, L. C., & Sharpee, T. O. (2011). Second order dimensionality reduction using minimum and maximum mutual information models. *PLoS Comput. Biol.*, *7*, e1002249.

Fitzgerald, J. D., Sincich, L. C., & Sharpee, T. O. (2011). Minimal models of multidimensional computations. *PLoS Comput. Biol.*, *7*, e1001111.

Geisler, W. S. (2008). Visual perception and the statistical properties of natural scenes. *Annu. Rev. Psychol.*, *59*, 167–192.

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

Gerwinn, S., Macke, J., & Bethge, M. (2010). Bayesian inference for generalized linear models for spiking neurons. *Frontiers in Comput. Neurosci.*, *4*, 12.

Gerwin, S., Macke, J., Seeger, M., & Bethge, M. (2008). Bayesian inference for spiking neuron models with a sparsity prior. In D. Köller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in neural information processing systems*, *20*. Cambridge, MA: MIT Press.

Globerson, A., Stark, E., Vaadia, E., & Tishby, N. (2009). The minimum information principle and its application to neural code analysis. *Proc. Natl. Acad. Sci. USA*, *106*, 3490–3495.

Gollisch, T., & Herz, A. V. M. (2005). Disentangling sub-millisecond process within an auditory transduction chain. *PLoS Comp. Biol.*, *9*(3), e1002922.

Gollisch, T., & Meister, M. (2010). Eye smarter than scientists believed: Neural computations in circuits of the retina. *Neuron*, *65*, 150–164.

Granot-Atedgi, E., Tkačik, G., Segev, R., & Schneidman, E. (2013). Stimulus-dependent maximum entropy models of neural population codes. *PLoS Comput. Biol.*, *9*, e1002922.

Hartline, H. K. (1940). The receptive fields of optic nerve fibers. *Am. J. Physiol.*, *130*, 690–699.

Hong, S., Agüera y Arcas, B., & Fairhall, A. L. (2007). Single neuron computation: From dynamical system to feature detector. *Neural Comput.*, *19*, 3133–3172.

Hubel, D. H., & Wiesel, T. H. (1965). Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *J. Physiol.*, *28*, 229–289.

Jaynes, E. T. (1957a). Information theory and statistical mechanics. *Phys. Rev.*, *106*, 620–630.

Jaynes, E. T. (1957b). Information theory and statistical mechanics II. *Phys. Rev.*, *108*, 171–190.

Karklin, Y., & Lewicki, M. S. (2009). Emergence of complex cell properties by learning to generalize in natural scenes. *Nature*, *457*, 83–86.

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

Kerr, J.N.D., & Nimmerjahn, A. (2012). Functional imaging in freely moving animals. *Curr. Op. Neurobiol.*, *22*, 45–53.

Kinney, J. B., Tkačik, G., & Callan, C. G. Jr (2007). Precise physical models of protein-DNA interaction from high-throughput data. *Proc. Nat. Acad. Sci. USA*, *104*, 501–506.

Kouh, M., & Sharpee, T. O. (2009). Estimating linear-nonlinear models using Renyi divergences. *Network*, *20*, 49–68.

Lesica, N. A., Ishii, T., Stanley, G. B., & Hosoya, T. (2008). Estimating receptive fields from responses to natural stimuli with asymmetric intensity distributions. *PLoS ONE*, *3*, e3060.

Lundstrom, B. N., Hong, S., & Fairhall, A. L. (2008). Two computational regimes of a single-compartment neuron separated by a planar boundary in conductance space. *Neural Comput.*, *20*, 1239–1260.

Maravall, M., Petersen, R. S., Fairhall, A. L., Arabzadeh, E., & Diamond, M. E. (2007). Shifts in coding properties and maintenance of information transmission during adaptation in barrel cortex. *PLoS Biol.*, *5*, e19.

Marmarelis, P. Z., & Marmarelis, V. Z. (1978). *Analysis of physiological systems: The white-noise approach*. New York: Plenum.

Olveczky, B. P., Baccus, S. A., & Meister, M. (2007). Retinal adaptation to object motion. *Neuron*, *56*, 689–700.

Ostojic, S., & Brunel, N. (2011). From spiking neuron models to linear-nonlinear models. *PLoS Comput. Biol.*, *7*, e1001056.

Ozuysal, Y., & Baccus, S. A. (2012). Linking the computational structure of variance adaptation to biophysical mechanisms. *Neuron*, *73*, 1002–1015.

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

Paninski, L. (2004). Maximum likelihood estimation of cascade point-process neural encoding models. *Network Comp. Neural Syst.*, *15*, 243–262.

Paninski, L., Pillow, J. W., & Simoncelli, E. P. (2004). Maximum likelihood estimation of a stochastic integrate-and-fire neural encoding model. *Neural Comput.*, *16*, 2533–2561.

Park, I., & Pillow, J. W. (2011). Bayesian spike-triggered covariance. In J. Shawe-Taylor, R. S. Zemel, P. Bartlett, F. Pereira, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems, 24* (pp. 1692–1700). Red Hook, NY: Curran.

Pillow, J. W. (2007). Likelihood-based approaches to modeling the neural code. In K. Daya, S. Ishii, A. Pouget, & R. Rao (Eds.), *Bayesian brain: Probabilistic approaches to neural coding* (pp. 53–70). Cambridge, MA: MIT Press.

Pillow, J. W., Shlens, J., Paninski, L., Sher, A., Litke, A. M., Chichilnisky, E. J., et al. (2008). Spatio-temporal correlations and visual signalling in a complete neural population. *Nature*, *454*, 995–999.

Pillow, J. W., & Simoncelli, E. P. (2006). Dimensionality reduction in neural models: An information-theoretic generalization of spike-triggered average and covariance analysis. *J. Vis.*, *6*, 414–428.

Rajan, K., & Bialek, W. (2012). *Maximally informative "stimulus energies" in the analysis of neural responses to natural signals*. arXiv.org:1201.0321.

Rapela, J., Felsen, G., Touryan, J., Mendel, J. M., & Grzywacz, N. M. (2010). ePPR: A new strategy for the characterization of sensory cells from input/output data. *Network*, *21*, 35–90.

Recio-Spinoso, A., Temchin, A. N., van Dijk, P., Fan, Y. H., & Ruggero, M. A. (2005). Wiener-kernel analysis of responses to noise of chinchilla auditory-nerve fibers. *J. Neurophys.*, *93*, 3615–3634.

Reid, R. C., Victor, J. D., & Shapley, R. M. (1997). The use of m-sequences in the analysis of visual neurons: Linear receptive field properties. *Vis. Neurosci.*, *14*, 1015–1027.

Rust, N. C., & Movshon, J. A. (2005). In praise of artifice. *Nat. Neurosci.*, *8*, 1647–1650.

Rust, N. C., Schwartz, O., Movshon, J. A., & Simoncelli, E. P. (2004). Spike-triggered characterization of excitatory and suppressive stimulus dimensions in monkey V1. *Neurocomputing*, *5860*, 793–799.

Sakai, H. M. (1992). White-noise analysis in neurophysiology. *Physiol. Rev.*, *72*, 491–505.

Schetzen, M. (1989). *The Volterra and Wiener theories of nonlinear systems*. Malabar, FL: Krieger.

Schneidman, E., Berry, M. J. II, Segev, R., & Bialek, W. (2006). Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature*, *440*, 1007–1012.

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.), *Advances in neural information processing systems*, *14* (pp. 269–276). Cambridge, MA: MIT Press.

Schwartz, O., Pillow, J. W., Rust, N. C., & Simoncelli, E. P. (2006). Spike-triggered neural characterization. *J. Vis.*, *6*, 484–507.

Schwartz, O., & Simoncelli, E. P. (2001). Natural signal statistics and sensory gain control. *Nat. Neurosci.*, *4*, 819–825.

Schwartz, G. W., Okawa, H., Dunn, F. A., Morgan, J. L., Kerschensteiner, D., Wong, R. O., et al. (2012). The spatial structure of a nonlinear receptive field. *Nature Neurosci.*, *15*, 1572–1580.

Schwartz, G., Taylor, S., Fisher, C., Harris, R., & Berry, M. J. II. (2007). Synchronized firing among retinal ganglion cells signals motion reversal. *Neuron*, *55*, 958–969.

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

Sharpee, T. O., Sugihara, H., Kurgansky, A. V., Rebrik, S. P., Stryker, M. P., & Miller, K. D. (2006). Adaptive filtering enhances information transmission in visual cortex. *Nature*, *439*, 936–942.

Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annu. Rev. Neurosci.*, *24*, 1193–1216.

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

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

Tkačik, G., Garrigan, P., Ratliff, C., Milčinski, G., Klein, J. M., Seyfarth, L. H., et al. (2011). Natural images from the birthplace of the human eye. *PLoS ONE*, *6*, e20409.

Tkačik, G., & Magnasco, M. O. (2008). Decoding spike timing: The differential reverse-correlation method. *Biosystems*, *93*, 90–100.

Tkačik, G., Prentice, J. S., Victor, J. D., & Balasubramanian, V. (2010). Local statistics in natural scenes predict the saliency of synthetic textures. *Proc. Natl. Acad. Sci. USA*, *107*, 18149–18154.

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

Truccolo, W., Eden, U. T., Fellows, M. R., Donoghue, J. P., & Brown, E. N. (2004). A point process framework for relating neural spiking activity to spiking history, neural ensemble and extrinsic covariate effects. *J. Neurophysiol.*, *93*, 1074–1089.

van Hateren, J. H. (1992). Theoretical predictions of spatiotemporal receptive felds of fly LMCs, and experimental validation. *J. Comp. Physiol. A*, *171*, 157–170.

Victor, J. D., & Johannesma, P. (1986). Maximum-entropy approximations of stochastic nonlinear transfunctions: An extension of the Wiener theory. *Biol. Cybern.*, *54*, 289–300.

Victor, J. D., & Knight, B. W. (1979). Nonlinear analysis with an arbitrary stimulus ensemble. *Quart. Appl. Math.*, *2*, 113–136.

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

Victor, J. D., & Shapley, R. M. (1980). The effect of contrast on the non-linear response of the Y cell. *J. Physiol.*, *302*, 535–547.

Wiener, N. (1958). *Nonlinear problems in random theory*. New York: Wiley.

Williamson, R. S., Sahani, M., & Pillow, J. W. (2011). *On information-theoretic and likelihood-based methods for spike-triggered neural characterization*. Poster at COSYNE 2011.

Wu, M.C.K., David, S. V., & Gallant, J. L. (2006). Complete functional characterization of sensory neurons by system identification. *Annu. Rev. Neurosci.*, *29*, 477–505.

Zetzsche, C., & Nuding, U. (2005). Nonlinear and higher-order approaches to the encoding of natural scenes. *Network*, *16*, 191–221.

Received September 1, 2012; accepted January 22, 2013.