Published in: Neural Computation, 16:2533-2561, Dec 2004.

Maximum Likelihood Estimation of a Stochastic Integrate-and-Fire Neural Encoding Model

Liam Paninski[1,3], Jonathan W. Pillow[1], and Eero P. Simoncelli[1,2]

[1] Howard Hughes Medical Institute,
Center for Neural Science, and
[2] Courant Institute for Mathematical Sciences
New York University

[3] Gatsby Computational Neuroscience Unit,
University College London

http://www.cns.nyu.edu/∼liam/
{*liam, pillow, eero*}*@cns.nyu.edu*

November 30, 2004

We examine a cascade encoding model for neural response in which a linear filtering stage is followed by a noisy, leaky, integrate-and-fire spike generation mechanism. This model provides a biophysically more realistic alternative to models based on Poisson (memoryless) spike generation, and can effectively reproduce a variety of spiking behaviors seen *in vivo*. We describe the maximum likelihood estimator for the model parameters, given only extracellular spike train responses (not intracellular voltage data). Specifically, we prove that the log likelihood function is concave and thus has an essentially unique global maximum that can be found using gradient ascent techniques. We develop an efficient algorithm for computing the maximum likelihood solution, demonstrate the effectiveness of the resulting estimator with numerical simulations, and discuss a method of testing the model's validity using time-rescaling and density evolution techniques.

1 Introduction

A central issue in systems neuroscience is the experimental characterization of the functional relationship between external variables — e.g., sensory stimuli, or motor behavior — and neural spike trains. Because neural responses to identical experimental input conditions are variable, we frame the problem statistically: we want to estimate the probability of any spiking response conditioned on any input. Of course, there are typically far too many possible observable signals to measure these probabilities directly. Thus our real goal is to find a good model, some functional form that allows us to predict spiking probability even for signals we have never observed directly. Ideally, such a model will be both accurate in describing neural response, and easy to estimate from a modest amount of data.

A good deal of recent interest has focused on models of "cascade" type; these models consist of a linear filtering stage in which the observable signal is projected onto a low-dimensional subspace, followed by a nonlinear, probabilistic spike generation stage. The linear filtering stage is typically interpreted as the neuron's "receptive field," efficiently representing the relevant information contained in the possibly high-dimensional input signal, 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. Unfortunately, biophysically realistic models of spike generation, such as the Hodgkin-Huxley model or its variants (Koch, 1999), are generally quite difficult to fit given only extracellular data.

As such, it has become common to assume a highly simplified model in which spikes are generated according to an inhomogeneous Poisson process, with rate determined by an instantaneous ("memoryless") nonlinear function of the linearly filtered input (see (Simoncelli et al., 2004) for review and partial list of references). In addition to its conceptual simplicity, this Linear-Nonlinear-Poisson (LNP) cascade model is computationally tractable. In particular, 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"

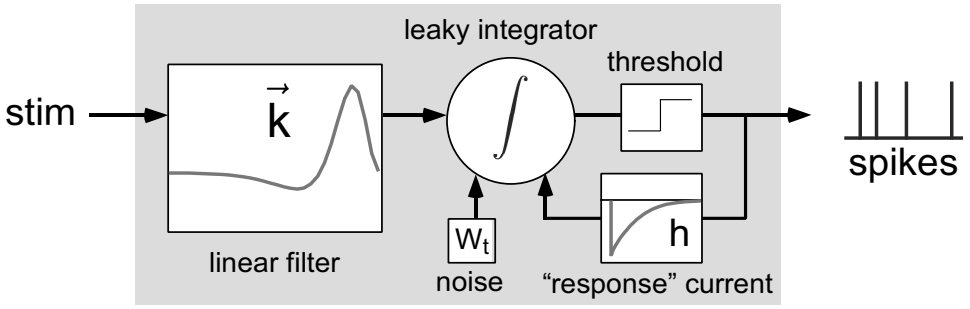

Figure 1: Illustration of the L-NLIF model.

stimuli (Paninski, 2003). Unfortunately, however, memoryless Poisson processes do not readily capture the fine temporal statistics of neural spike trains (Berry and Meister, 1998; Keat et al., 2001; Reich et al., 1998; Aguera y Arcas and Fairhall, 2003). In particular, the probability of observing a spike is not a functional of the recent stimulus alone; it is also strongly affected by the recent history of spiking. This spike-history dependence can significantly bias the estimation of the linear filter of an LNP model (Berry and Meister, 1998; Pillow and Simoncelli, 2003; Paninski et al., 2003b; Paninski, 2003; Aguera y Arcas and Fairhall, 2003).

In this paper, we consider a model that provides an appealing compromise between the oversimplified Poisson model and more biophysically realistic but intractable models for spike generation. The model consists of a linear filter (L) followed by a probabilistic, or noisy (N), form of leaky integrate-and-fire (LIF) spike generation (Koch, 1999). This "L-NLIF" model is illustrated in Fig. 1, and is essentially the standard LIF model driven by a noisy, filtered version of the stimulus; the spike history dependence introduced by the integrate-and-fire mechanism allows the model to emulate many of the spiking behaviors seen in real neurons (Gerstner and Kistler, 2002). This model thus combines the encoding power of the LNP cell with the flexible spike history dependence of the LIF model, and allows us to explicitly model neural firing statistics.

Our main result is that the estimation of the L-NLIF model parameters is computationally tractable. Specifically, we formulate the problem in terms of classical estimation theory, which provides a natural "cost function" (like-

lihood) for model assessment and estimation of the model parameters. We describe algorithms for computing the likelihood function and prove that this likelihood function contains no non-global local maxima, implying that the maximum likelihood estimator (MLE) can be computed efficiently using standard ascent techniques. Desirable statistical properties of the estimator (consistency, efficiency, etc.) are all inherited "for free" from classical estimation theory (van der Vaart, 1998). 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.

2 The Model

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

$$dV = \Big(-g(V(t) - V_{leak}) + I_{stim}(t) + I_{hist}(t) \Big) dt + W_t, \tag{1}$$

and resets instantaneously to $V_{reset} < 1$ whenever $V = 1$, the threshold potential (Fig. 2). Here, g denotes the membrane leak conductance, V_{leak} the leak reversal potential, and the stimulus current I_{stim} is defined as

$$I_{stim}(t) = \vec{k} \cdot \vec{x}(t),$$

the projection of the input signal $\vec{x}(t)$ onto the spatiotemporal linear kernel $\vec{k}$; the spike-history current I_{hist} is given by

$$I_{hist}(t) = \sum_{j=0}^{i-1} h(t - t_j),$$

where h is a post-spike current waveform of fixed amplitude and shape[1] whose value depends only on the time since the last spike t_{i-1} (with the sum above including terms back to t_0, the first observed spike); finally, W_t is an unobserved (hidden) noise process, taken here to be a standard Gaussian white noise (although we will consider more general W_t later). As usual,

[1]The letter h here was chosen to stand for "history," and should not be confused with the physiologically-defined I_h current.

in the absence of input, V decays back to V_{leak} with time constant $1/g$. Thus, the nonlinear behavior of the model is completely determined by only a few parameters, namely $\{g, V_{reset}, V_{leak}\}$, and $h(t)$. In practice, we assume the continuous aftercurrent $h(t)$ may be written as a superposition of a small number of fixed temporal basis functions; we will refer to the vector of coefficients in this basis using the vector $\vec{h}$. We should note that the inclusion of the I_{hist} current in (1) introduces additional parameters (namely, $\vec{h}$) to the model that need to be fit; in cases where there is insufficient data to properly fit these extra parameters, $\vec{h}$ could be set to zero, reducing the model (1) to the more standard LIF setting.

It is important to emphasize that in the following, $V(t)$ itself will be considered a hidden variable; we are assuming that the spike train data we are trying to model has been collected extracellularly, without any access to the subthreshold voltage V. This implies that the parameters of the usual LIF model can only be estimated up to an unlearnable mean and scale factor. Thus, by a standard change of variables, we have not lost any generality by setting the threshold potential, V_{th}, and scale of the hidden noise process, σ, to 1 (corresponding to mapping the physical voltage $V \to 1 + (V - V_{th})/\sigma$); the relative noise level (the effective scale of W_t) can be changed by scaling $V_{leak}, V_{reset}, \vec{k}$, and h together. Of course other changes of variable are possible (for example, letting σ change freely and fixing $V_{reset} = 0$), but will not affect the analysis below.

The dynamical properties of this type of "spike response model" have been extensively studied (Gerstner and 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). We illustrate some of these diverse firing properties in Figures 2 - 2; these figures also serve to illustrate several of the important differences between the L-NLIF and LNP models. In Fig. 2, 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 and Meister, 1998; Reich et al., 1998; Keat et al., 2001). 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: by varying V_{reset} or h, for example, we can match a wide variety of interspike interval distributions and firing rate

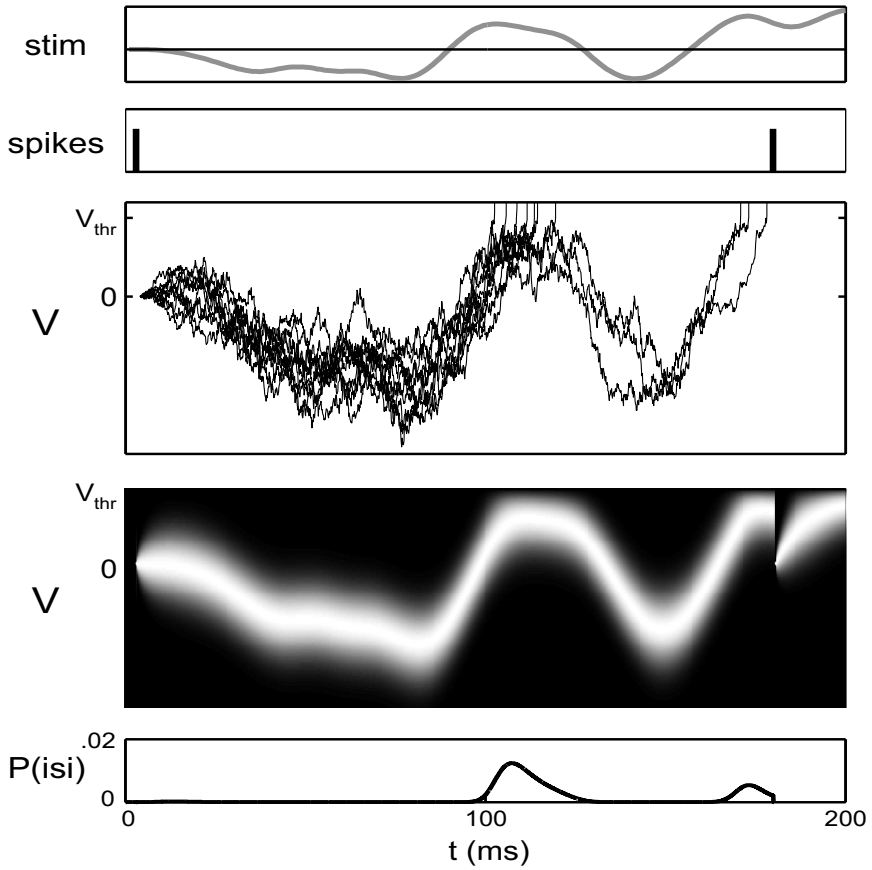

Figure 2: Behavior of the L-NLIF model during a single interspike interval, for a single (repeated) input current. **Top:** Observed stimulus $x(t)$ and response spikes. **Third panel:** Ten simulated voltage traces $V(t)$, evaluated up to the first threshold crossing, conditional on a spike at time zero ($V_{reset} = 0$). Note the strong correlation between neighboring time points, and the gradual sparsening of the plot as traces are eliminated by spiking. **Fourth:** Evolution of $P(V, t)$. Each vertical cross section represents the conditional distribution of V at the corresponding time t (i.e. for all traces that have not yet crossed threshold). Note the boundary conditions $P(V_{th}, t) = 0$ and $P(V, t_{spike}) = \delta(V - V_{reset})$ corresponding to threshold and reset, respectively; see section 4 for computational details. **Bottom:** Probability density of the interspike interval (ISI) corresponding to this particular input. Note that probability mass is concentrated at the times when input drives the mean voltage $V_0(t)$ close to threshold. Careful examination reveals, in fact, that peaks in $p(ISI)$ are sharper than peaks in the deterministic signal $V_0(t)$, due to the elimination of threshold-crossing traces which would otherwise have contributed mass to $p(ISI)$ at or after such peaks (Berry and Meister, 1998).

curves, even given a single fixed stimulus. For example, the model can mimic the FI curves of "type I" or "II" models, with either smooth or discontinuous growth of the FI curve away from 0 at threshold, respectively (Gerstner and Kistler, 2002). More generally, the L-NLIF model can exhibit adaptive behavior (Rudd and Brown, 1997; Paninski et al., 2003b; Yu and Lee, 2003) and display rhythmic, tonic, or even bistable dynamical behavior, depending on the parameter settings (Fig. 2).

3 The Estimation Problem

Our problem now is to estimate the model parameters $\theta \equiv \{\vec{k}, g, V_{leak}, V_{reset}, h\}$ from a sufficiently rich, dynamic input sequence $\vec{x}(t)$ and the response spike times $\{t_i\}$. We emphasize again that we are *not* discussing the problem of estimating the model parameters given intracellularly-recorded voltage traces (Stevens and Zador, 1998; Jolivet et al., 2003); we assume that these subthreshold responses, which greatly facilitate the estimation problem, are unknown — "hidden" — to us. A natural choice is the maximum likelihood estimator (MLE), which is easily proven to be consistent and statistically efficient here (van der Vaart, 1998). To compute the MLE, we need to compute the likelihood and develop an algorithm for maximizing the likelihood as a function of the parameters θ.

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 and Simoncelli, 2003), the voltage trace during an interspike interval $t \in [t_{i-1}, t_i]$ is given by the solution to equation (1) with the noise W_t turned off, with initial conditions $V_0(t_{i-1}) = V_{reset}$:

$$V_0(t) = V_{leak} + (V_{reset} - V_{leak}) e^{-g(t-t_{i-1})} + \int_{t_{i-1}}^{t} \left(\vec{k} \cdot \vec{x}(s) + \sum_{j=0}^{i-1} h(s - t_j) \right) e^{-g(t-s)} ds, \tag{2}$$

which is simply a linear convolution of the input current with a filter which decays exponentially with time constant $1/g$. 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 and

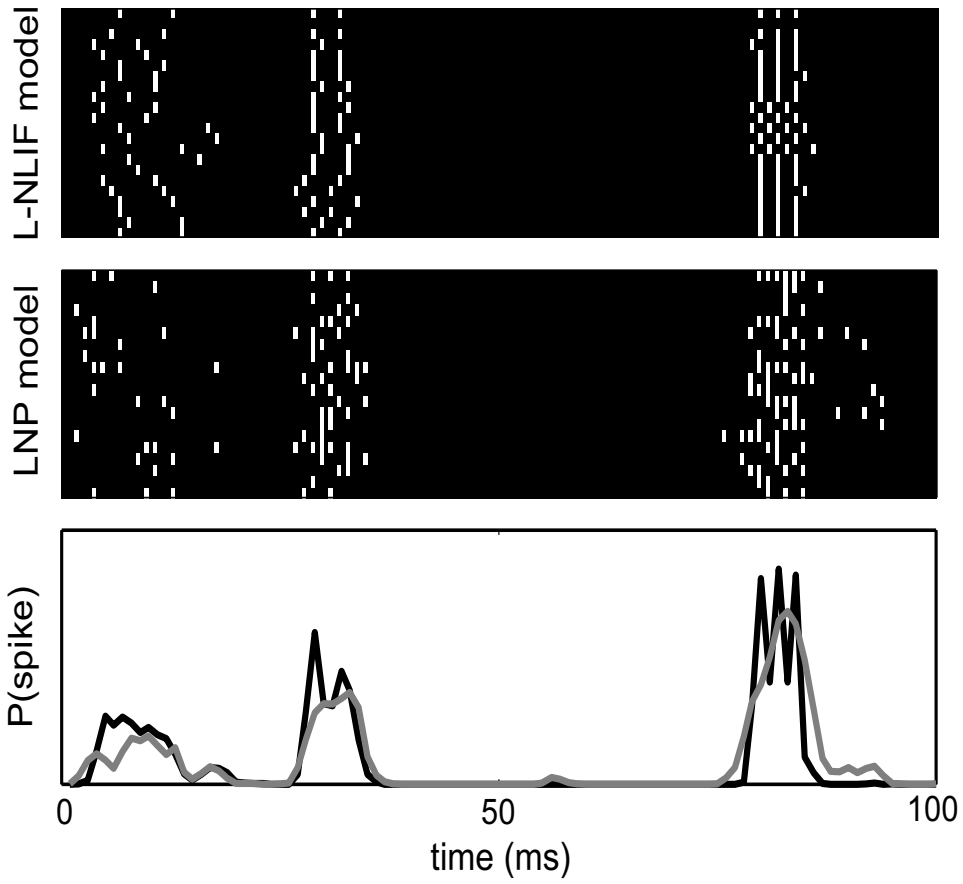

Figure 3: Simulated responses of L-NLIF and LNP models to 20 repetitions of a fixed 100-ms stimulus segment of temporal white noise. **Top:** Raster of responses of L-NLIF model to a dynamic input stimulus. The top row shows the fixed (deterministic) response of the model with the noise set to zero. **Middle:** Raster of responses of LNP model to the same stimulus, with parameters fit with standard methods from a long run of the L-NLIF model responses to non-repeating stimuli. **Bottom:** Post-stimulus time histogram (PSTH) of the simulated L-NLIF response (black line), and PSTH of the LNP model (gray line). Note that the LNP model, due to its Poisson output structure, fails to preserve the fine temporal structure of the spike trains, relative to the L-NLIF model.

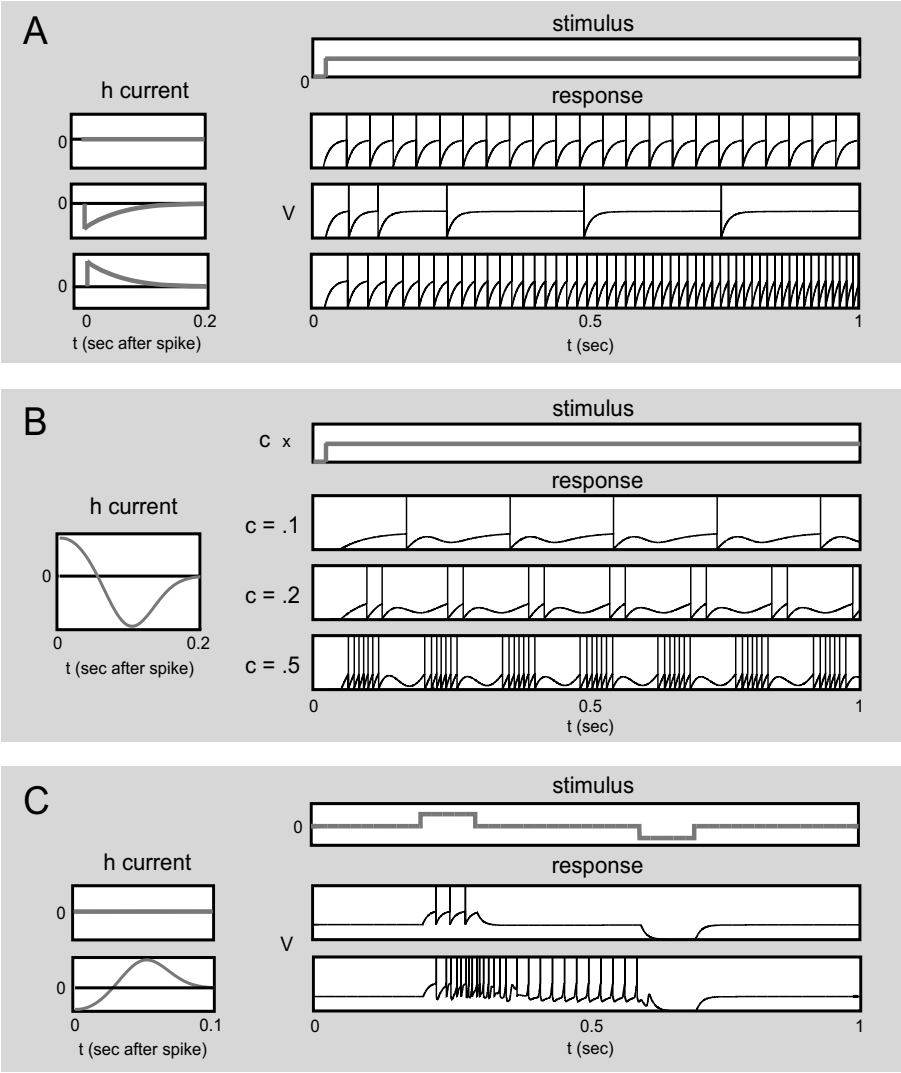

Figure 4: Diversity of NLIF model response patterns. **A.** Firing rate adaptation. A positive DC current was injected into three different NLIF cells, all with slightly different settings for h (top, $h = 0$; middle, h depolarizing; bottom, h hyperpolarizing). Note that all three voltage response traces are identical up until the time of the first spike, but adapt to the constant input in three different ways. (For clarity, noise level set to zero in all panels.) **B.** Rhythmic, bursting responses. DC current (top trace) injected into an NLIF cell with h shown at left. As amplitude c of current increases (voltage traces, top to bottom), burst frequency and duration increase. **C.** Tonic and bistable ("memory") responses. The same current (bottom trace) was injected into two different NLIF cells with different settings for h. The biphasic h in the bottom panel leads to a self-sustaining response that is inactivated only by the subsequent negative pulse.

Taylor, 1981). This density is uniquely characterized by its first two moments; the mean is given by (2), and its covariance

$$\mathrm{Cov}(t_1, t_2) = E_g E_g^T = \frac{1}{2g}\left(e^{-g|t_2-t_1|} - e^{-g(t_1+t_2)}\right), \tag{3}$$

where E_g is the convolution operator corresponding to e^{-gt}. We denote this Gaussian density $G(V(t)|\vec{x}_i, \theta)$, where index i indicates the ith spike and the corresponding stimulus segment $\vec{x}_i$ (i.e. the stimuli that influence $V(t)$ during the ith interspike interval). Note that this density is highly correlated for nearby points in time; intuitively, smaller leak conductance g leads to stronger correlation in $V(t)$ at nearby time points.

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

$$L_{\{\vec{x}_i, t_i\}}(\theta) = \int_{V \in C_i} G(V(t)|\vec{x}_i, \theta),$$

the integral of the Gaussian density $G(V(t)|\vec{x}_i, \theta)$ over the set C_i of (unobserved) voltage paths consistent with the observed spike train data.

Spiking resets V to V_{reset}; since W_t is white noise, this means 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, define the full likelihood as

$$L_{\{\vec{x}_i, t_i\}}(\theta) = \prod_i \int_{V \in C_i} G(V(t)|\vec{x}_i, \theta),$$

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

Now that we have an expression for the likelihood, we need to be able to maximize it over the parameters θ. Our main result is that we can use simple ascent algorithms to compute the MLE without fear of becoming trapped in local maxima[2].

Theorem 1. *The likelihood $L_{\{\vec{x}_i,t_i\}}(\theta)$ has no non-global local extrema in the parameters θ, for any data $\{\vec{x}_i, t_i\}$.*

The proof of the theorem (in the appendix) is based on the logconcavity of the likelihood $L_{\{\vec{x}_i,t_i\}}(\theta)$ under a certain relabelling of the parameters (θ). The classical approach for establishing the nonexistence of local maxima of a given function is 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 local 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.

[2]More precisely, we say that a smooth function has no non-global local extrema if the set of points at which the gradient vanishes is connected and (if nonempty) contains a global extremum; thus all "local extrema" are in fact global, if a global maximum exists. (This existence, in turn, is guaranteed asymptotically by classical MLE theory whenever the model's parameters are identifiable, and guaranteed in general if we assume θ takes values in some bounded set.) Note that the L-NLIF model has parameter space isomorphic to the convex domain $\Re^{\dim(\vec{k})+\dim(\vec{h})+1} \times \Re_+^2$, with $\Re_+$ denoting the positive axis (recall that the parameter h takes values in a finite-dimensional space, $g > 0$, and $V_{reset} < 1$).

4 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 describe two ways to compute these integrals; each has its own advantages.

The first technique can be termed "density evolution" (Knight et al., 2000; Haskell et al., 2001; Paninski et al., 2003b). The method is based on the following well-known fact from the theory of stochastic differential equations (Karlin and Taylor, 1981): given the data $(\vec{x}_i, t_{i-1})$, the probability density of the voltage process $V(t)$ up to the next spike t_i satisfies the following partial differential (Fokker-Planck) equation:

$$\frac{\partial P(V,t)}{\partial t} = \frac{1}{2}\frac{\partial^2 P}{\partial V^2} + g\frac{\partial[(V - V_{rest})P]}{\partial V}, \tag{4}$$

under the boundary conditions

$$P(V, t_{i-1}) = \delta(V - V_{reset}),$$

$$P(V_{th}, t) = 0,$$

enforcing the constraints that voltage resets at V_{reset} and is killed (due to spiking) at V_{th}, respectively. $V_{rest}(t)$ is defined, as usual, as the stationary point of the noiseless subthreshold dynamics (1):

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

The integral $\int P(V,t)dV$ is simply the probability that the neuron has not yet spiked at time t, given that the last spike was at t_{i-1}; thus, $1 - \int P(V,t)dV$ is the cumulative distribution of the spike time since t_{i-1}. Therefore

$$f(t) \equiv -\frac{\partial}{\partial t}\int P(V,t)dV,$$

the conditional probability density of a spike at time t (defined at all times $t \notin \{t_i\}$), satisfies

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

Thus standard techniques (Press et al., 1992) for solving the drift-diffusion evolution equation (4) lead to a fast method for computing $f(t)$ (as illustrated in Fig. 2). Finally, the likelihood $L_{\vec{x}_i,t_i}(\theta)$ is simply $\prod_i f(t_i)$.

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 probability of spiking $f(t)$ at all times t, but rather at just a subset of times $\{t_i\}$. In fact, while we are ascending the likelihood surface (in particular, while we are far from the maximum), we do not need to know the likelihood precisely, and can trade accuracy for speed. Thus we can turn to more specialized, approximate techniques for faster performance. Our algorithm can be described in three steps.

The first is a specialized algorithm due to Genz (Genz, 1992), designed to compute exactly the kinds of integrals considered here, which works well when the orthants C_i are defined by fewer than ≈ 10 linear constraints. The number of actual constraints grows linearly in the length of the interspike interval $(t_{i+1} - t_i)$; 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 and Simoncelli, 2003), in which we discard all but a small subset of the constraints. The key point is that only a few constraints are actually needed to approximate the integrals to a high degree of precision, basically because of the strong correlations between the value of V at nearby time points.

This idea provides us with an efficient approximation of the likelihood at a single point in parameter space. To find the maximum of this function using standard ascent techniques, we obviously have to compute the likelihood at many such points. We can make this ascent process much quicker by applying a version of the coarse-to-fine idea. Let L_j denote the approximation to the likelihood given by allowing only j constraints in the above algorithm. Then we know, by a proof identical to that of Theorem 1, that L_j has no local maxima; in addition, by the above logic, $L_j \to L$ as j grows. It takes little additional effort to prove that

$$\text{argmax}_{\theta\in\Theta}\ L_j(\theta) \to \text{argmax}_{\theta\in\Theta}\ L(\theta)$$

as $j \to \infty$; thus, we can efficiently ascend the true likelihood surface by ascending the "coarse" approximants L_j, then gradually "refining" our ap-

proximation by letting j increase. The $j = \infty$ term is computed via the full density evolution method.

The last trick is a simple method for choosing a good starting point for each ascent. To do this, we borrow the jackknife idea from statistics (Efron and Stein, 1981; Strong et al., 1998): set our initial guess for the maximizer of L_{j_N} to be

$$\theta^0_{j_N} = \theta^\infty_{j_{N-1}} + \frac{j_N^{-1} - j_{N-1}^{-1}}{j_{N-1}^{-1} - j_{N-2}^{-1}} \left(\theta^\infty_{j_{N-1}} - \theta^\infty_{j_{N-2}} \right),$$

the linear extrapolant on a $1/j$ scale.

Now that we have an efficient ascent algorithm, we need to provide it with a sensible initialization of the parameters. We employ the following simple method, related to our previous work on the deterministic LIF model (Pillow and Simoncelli, 2003): we set g^0 to some physiologically plausible value (say $(50 \text{ ms})^{-1}$), then $\vec{k}^0, h^0$ and V^0_{leak} to the ML solution of the following regression problem:

$$E_g \left(\vec{k} \cdot \vec{x}_i + gV_{leak} + \sum_{j=0}^{i-1} h(t - t_j) \right) = 1 + \sigma_i \epsilon_i,$$

with ϵ_i a standard i.i.d. normal random variable scaled by

$$\sigma_i = \mathrm{Cov}(t_i - t_{i-1}, t_i - t_{i-1})^{1/2} = \frac{1}{\sqrt{2g}} e^{-g(t_i - t_{i-1})},$$

the standard deviation of the Ornstein-Uhlenbeck process V (recall expression (3)) at time $t_i - t_{i-1}$. Note that the reset potential V^0_{reset} is initially fixed at zero, away from the threshold voltage $V_{th} = 1$, to prevent the trivial $\theta = 0$ solution. The solution to this regression problem has the usual least-squares form and can thus be quickly computed analytically (see (Sahani and Linden, 2003) for a related approach), and serves as a kind of $j = 1$ solution (with the single voltage constraint placed at t_i, the time of the spike). See also (Brillinger, 1992) for a discrete-time formulation of this single-constraint approximation.

To summarize, we provide pseudocode for the full algorithm in Fig. 4. One important note is that, due to its ascent structure, the algorithm can be gracefully interrupted at any time without catastrophic error. In addition,

---

- Initialize $(\vec{k}, V_l, h)$ to regression solution
- Normalize by observed scale of ϵ_i
- **for** increasing j

  Let θ_j maximize L_j

  Jackknife θ_{j+1}

  end
- Let $\theta_{MLE} \equiv \theta_\infty$ maximize L

Figure 5: Pseudocode for the L-NLIF MLE.

the time complexity of the algorithm is linear in the number of spikes. 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.

5 Time Rescaling

Once we have obtained our estimate of the parameters $(\vec{k}, g, V_{leak}, V_{reset}, h)$, 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 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 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. 2.

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

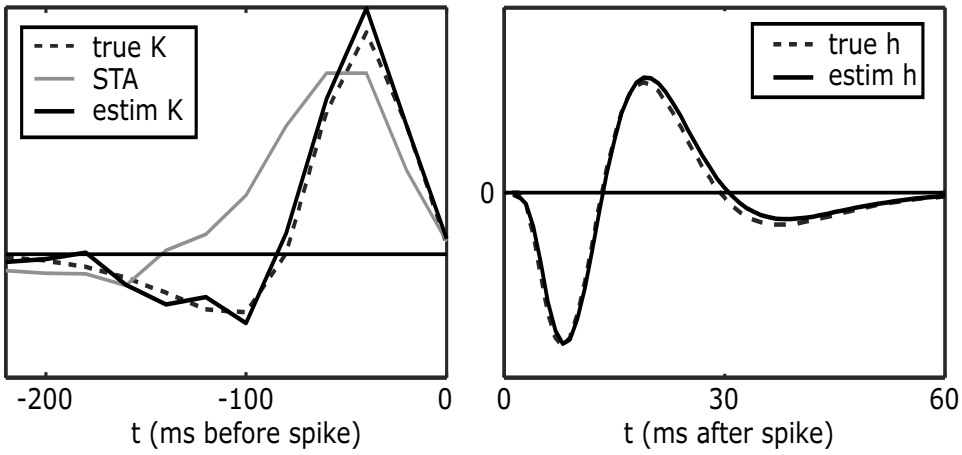

Figure 6: Demonstration of the estimator's performance on simulated data. Dashed lines show the true kernel $\vec{k}$ and aftercurrent h; $\vec{k}$ is a 12-sample function chosen to resemble the biphasic temporal impulse response of a macaque retinal ganglion cell (Chichilnisky, 2001), while h is a weighted sum of five gamma functions whose biphasic shape induces a slight degree of burstiness in the model's spike responses (c.f. Fig. 2). With only 600 spikes of output (given temporal white noise input), the estimator is able to retrieve an estimate of $\vec{k}$ which closely matches the true $\vec{k}$ and h. Note that the spike-triggered average, which is an unbiased estimator for the kernel of a LNP neuron (Chichilnisky, 2001), differs significantly from the true $\vec{k}$ (and, of course, provides no estimate for h).

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.

6 Extensions

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

6.1 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

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

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 et al., 1999; Harris et al., 2003; Paninski et al., 2003a); 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.

6.2 Nonlinear input

Next, we can use a trick from the machine learning and regression literature (Duda and Hart, 1972; Cristianini and 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 and 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 and Abbott,

2001; Dodd and 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 et al., 1998) that span a reasonable space of possible nonlinearities, such as the principal components of previously observed nonlinear tuning functions (see also (Sahani and Linden, 2003) for a similar idea, but in a purely linear setting).

6.3 Regularization

The extensions discussed in the last two subsections have made our basic model considerably more powerful, but at the cost of a larger number of parameters that must be estimated from data. This is problematic, as it is well-known that the phenomenon of "overfitting" can actually hurt the predictive power of models based on a large number of parameters (see, e.g., (Sahani and Linden, 2003; Smyth et al., 2003; Machens et al., 2003) for examples, again in a linear regression setting). How do we control for overfitting in the current context?

One simple approach is to use a maximum *a posteriori* (MAP, instead of ML) estimate for the model parameters. This entails maximizing an expression of the penalized form

$$\log L(\theta) + Q(\theta)$$

instead of just $L(\theta)$, where $L(\theta)$ is the likelihood function, as above, and $-Q$ is some "penalty" function (where in the classical Bayesian setting, e^Q is required to be a probability measure on the parameter space Θ). If Q is taken to be concave, a glance at the proof of Theorem 1 shows that the MAP estimator shares the MLE's global extrema property; as usual, simple regularity conditions on Q ensure that the MAP estimator converges to the MLE given enough data, and therefore inherits the MLE's asymptotic efficiency.

Thus we are free to choose Q as we like within the class of smooth concave functions, bounded above. If Q peaks at a point such that all the weight coefficients (a_i or $\vec{k}$, depending on the version of the model in question) are zero, the MAP estimator will basically be a more "conservative" version of the MLE, with the chosen coefficients shifted nonlinearly towards zero. This type of "shrinkage" estimator has been extremely well-studied, from a variety of viewpoints (e.g., (James and Stein, 1960; Donoho et al., 1995; Tipping, 2001) and references therein), and is known, for example, to perform strictly better than the MLE in certain contexts. Again, see (Sahani and Linden, 2003; Smyth et al., 2003; Machens et al., 2003) for some illustrations of this effect. One particularly simple choice for Q is the weighted L^1 norm

$$Q(\vec{k}) = \sum_l |b(l)k(l)|,$$

where the weights $b(l)$ set the relative scale of Q over the likelihood and may be chosen by symmetry considerations, cross-validation (Machens et al., 2003; Smyth et al., 2003), and/or evidence optimization (Tipping, 2001; Sahani and Linden, 2003). This choice for Q has the property that sparse solutions (i.e., solutions for which as many components of $\vec{k}$ as possible are set to zero) are favored; the desirability of this feature is discussed, e.g., in (Girosi, 1998; Donoho and Elad, 2003) and references therein.

6.4 Correlated noise

In some situations (particularly when the cell is poorly driven by the input signal $\vec{x}(t)$), the whiteness assumption on the noise W_t will be inaccurate. Fortunately, it is possible to generalize this part of the model as well, albeit with a bit more effort. The simplest way to introduce correlations in the noise (Fourcaud and Brunel, 2002; Moreno et al., 2002) is to replace the white W_t with an Ornstein-Uhlenbeck process N_t defined by

$$dN = -\frac{N}{\tau_N}dt + W_t. \tag{5}$$

As above, this is simply white noise convolved with a simple exponential filter of time constant τ_N (and therefore the conditional Gaussianity of $V(t)$ is retained); the original white noise model is recovered as $\tau_N \to 0$,

after suitable rescaling. (N_t here is often interpreted as synaptic noise, with τ_N the synaptic time constant, but it is worth emphasizing that N_t is not voltage-dependent, as would be necessary in a strict conductance-based model.) Somewhat surprisingly, the essential uniqueness of the global likelihood maximum is preserved for this model: for any $\tau_N \geq 0$, the likelihood has no local extrema in $(\vec{k}, g, V_{leak}, V_{reset}, h)$.

Of course, we do have to make a few changes in the computational schemes associated with this new model. Most of the issues arise from the loss of the conditional renewal property of the interspike intervals for this model: the conditional probability of a spike given the input $\vec{x}$ is *not* conditionally independent of the last interspike interval (this is one of the main reasons we are interested in this correlated noise model). Instead, we have to write our likelihood $L_{\{\vec{x}_i,t_i\}}(\theta, \tau_N)$ as

$$\int p(N_t, \tau_N) \prod_i 1\Big(V_t(\vec{x}_i, \theta, N_t) \in C_i\Big) dN_t,$$

where the integral is over all noise paths N_t, under the Gaussian measure $p(N_t, \tau_N)$ induced on N_t by expression (5); the multiplicands on the right are 1 or 0 according to whether the voltage trace V_t, given the noise path N_t, the stimulus $\vec{x}_i$, and the parameters θ, was in the constraint set C_i or not, respectively.

Despite the loss of the renewal property, N_t is still a Gauss-Markov diffusion process, and we can write the Fokker-Planck equation (now in two dimensions, V and N):

$$\frac{\partial P(V,N,t)}{\partial t} = \frac{1}{2}\frac{\partial^2 P}{\partial N^2} + g\frac{\partial[(V - V_{rest} - \frac{N}{g})P]}{\partial V} + \frac{1}{\tau_N}\frac{\partial[NP]}{\partial N},$$

under the boundary conditions

$$P(V_{th}, N, t) = 0,$$

$$P(V, N, t_{i-1}^+) = -\frac{1}{Z}\delta(V - V_{reset})\frac{\partial P(V,N,t_{i-1}^-)}{\partial V}\bigg|_{V=V_{th}} R\Big(\frac{N}{g} - V_{th} + V_{rest}(t_{i-1}^-)\Big),$$

with R the usual linear rectifier

$$R(u) = \begin{cases} 0 & u \leq 0, \\ u & u > 0 \end{cases}$$

and Z the normalization factor

$$Z = -\int \frac{\partial P(V, N, t_{i-1}^-)}{\partial V}\bigg|_{V=V_{th}} R\bigg(\frac{N}{g} - V_{th} + V_{rest}(t_{i-1}^-)\bigg) dN;$$

the threshold condition here is the same as in equation (4), while the reset condition reflects the fact that V is reset to V_{reset} with each spike, but N is not (the complicated term on the right is obtained from the usual expression by conditioning on $V(t_{i-1}^-) = V_{th}$ and $\frac{\partial V(t_{i-1}^-)}{\partial t} > 0$). Note that the relevant discretized differential operators are still extremely sparse, allowing for efficient density propagation, although the density must now be propagated in two dimensions, which does make the soultion significantly more computationally costly than in the white noise case. Simple approximative approaches like those described in section 4 (via the Genz algorithm) are available as well.

6.5 Subthreshold resonance

Finally, it is worth examining how easily generalizable our methods and results might be to subthreshold dynamics more interesting than the (linear) leaky integrator employed here. While the density evolution methods developed in section 4 can be generalized easily to nonlinear and even time-varying subthreshold dynamics, the Genz algorithm obviously depends on the Gaussianity of the underlying distributions (which is unfortunately not preserved by nonlinear dynamics), and the proof of Theorem 1 appears to depend fairly strongly on the linearity of the transformation between input current and subthreshold membrane voltage (although linear filtering by non-exponential windows is allowed).

Perhaps the main generalization worth noting here is the extension from purely "integrative" to "resonant" dynamics. We can accomplish this by the simple trick of allowing the membrane conductance g to take complex values (see, e.g., (Izhikevich, 2001) and references therein for further details and background on subthreshold resonance). This transforms the low-pass exponential filtering of equation (2) to a band-pass filtering by a damped sinusoid: a product of an exponential and a cosine whose frequency is determined, as usual, by the imaginary part of g. All of the equations listed above remain otherwise unchanged if we ignore the imaginary part of this

new filter's output, and Theorem 1 continues to hold for complex g, with g restricted to the upper-right quadrant $(\mathrm{real}(g), \mathrm{imag}(g) \geq 0)$ to eliminate the conjugate symmetry of the filter corresponding to g. The only necessary change is in the density evolution method, where we need to propagate the density in an extra dimension to account for the imaginary part of the resulting dynamics (importantly, however, the Markov nature of model (1) is retained, preserving the linear diffusion nature of equation (4)).

7 Discussion

We have shown here that the L-NLIF model, which couples a filtering stage to a biophysically plausible and flexible model of neuronal spiking, can be efficiently estimated from extracellular physiological data. In particular, we proved that the likelihood surface for this model has no local peaks, ensuring the essential uniqueness of the maximum likelihood and maximum *a posteriori* estimators in some generality. This result leads directly to reliable algorithms for computing these estimators, which are known by general likelihood theory to be statistically consistent and efficient. Finally, we showed that the model lends itself directly to analysis via tools from the modern theory of point processes, such as time-rescaling tests for model validation. As such, we believe the L-NLIF model could become a fundamental tool in the analysis of neural data, a kind of canonical "encoding model."

Our primary goal was an elaboration of the LNP model to include spike-history (e.g. refractory) effects. As detailed in (Simoncelli et al., 2004), the basic LNP model provides a powerful framework for analyzing neural encoding of high-dimensional signals; however, it is well-known that the Poisson spiking model is inadequate to capture the fine temporal properties of real spike trains. Previous attempts to address this shortcoming have fallen into two classes: "multiplicative" models (Snyder and Miller, 1991; Miller and Mark, 1992; Iyengar and Liao, 1997; Berry and Meister, 1998; Brown et al., 2002; Paninski, 2003), of the basic form

$$p(\text{spike}(t) \mid \text{stimulus, spike history}) = F(\text{stimulus})H(\text{history})$$

— in which H encodes purely spike-history dependent terms like refractory

or burst effects — and "additive" models like

$$p(\text{spike}(t) \mid \text{stimulus, history}) = F\left(\text{stimulus} + H(\text{history})\right),$$

(Brillinger, 1992; Joeken et al., 1997; Keat et al., 2001; Truccolo et al., 2003), in which the spike history is basically treated as a kind of additional input signal; the L-NLIF model is of the latter form, with the post-spike current h injected directly into expression (1) with the filtered input $\vec{k}\cdot\vec{x}(t)$. It is worth noting that one popular form of the multiplicative history-dependence functional $H(\cdot)$ above, the "inverse-Gaussian" density model (Seshardri, 1993; Iyengar and Liao, 1997; Brown et al., 2002), arises as the first-passage time density for the Wiener process, effectively the time of the first spike in the L-NLIF model given constant input at no leak ($g = 0$); see (Stevens and Zador, 1996; Plesser and Gerstner, 2000) for further such multiplicative-type approximations. It seems that the treatment of history effects as simply another form of "stimulus" might make the additive class slightly easier to estimate (this was certainly the case here, for example); however, any such statement remains to be verified via systematic comparison of the accuracy of these two classes of models, given real data.

We based our model on the LIF cell in an attempt to simultaneously maximize two competing objectives: flexibility (explanatory power) and tractability (in particular, ease of estimation, as represented by Theorem 1). We attempted to make the model as general as possible without violating the conditions necessary to ensure the validity of this theorem: thus, we included the h current and the various extensions described in section 6 but did not, for example, attempt to model postsynaptic conductances directly, or permit any nonlinearity in the subthreshold dynamics (Brunel and Latham, 2003), or allow any rate-dependent modulations of the membrane conductance g (Stevens and Zador, 1998; Gerstner and Kistler, 2002); it is unclear at present whether Theorem 1 can be extended to these cases.

Of course, due largely to its simplicity, the LIF cell has become the *de facto* canonical model in cellular neuroscience (Koch, 1999). Although the 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 and Simoncelli, 2003). We and others have explored other "adaptive" properties of the LIF model (Rudd and Brown, 1997; Paninski et al., 2003b; Yu and Lee, 2003). We provided a brief sampling of the flexibility of the L-NLIF model in Figures 2-2; of course, similar behaviors have been noted elsewhere (Gerstner and Kistler, 2002), although the spiking diversity of this particular model (with no additional time-varying conductances, etc.) has not, to our knowledge, been previously collected in one place, and some aspects of this flexibility (e.g. Fig. 2C) might come as a surprise in such a simple model.

The probabilistic nature of the L-NLIF model provides several important advantages over the deterministic version we have considered previously (Pillow and Simoncelli, 2003). First, clearly, this probabilistic formulation is necessary for our entire likelihood-based presentation; moreover, use of an explicit noise model greatly simplifies the discussion of spiking statistics. Second, the simple subthreshold noise source employed here could provide a rigorous basis for a metric distance between spike trains, useful in other contexts (Victor, 2000). Finally, this type of noise influences the behavior of the model itself (c.f. Fig. 2), giving rise to phenomena not observed in the purely deterministic model (Levin and Miller, 1996; Rudd and Brown, 1997; Burkitt and Clark, 1999; Miller and Troyer, 2002; Paninski et al., 2003b; Yu and Lee, 2003).

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 (Warland et al., 1997; Tsodyks et al., 1999; Harris et al., 2003; Paninski et al., 2003a).

Appendix A: Proof of Theorem 1

Proof. We prove the main result indirectly, by establishing the more general statement in section 6.4: for any $\tau_N \geq 0$, the likelihood function for the L-NLIF model has no local extrema in $\theta = (\vec{k}, g, V_{leak}, V_{reset}, h)$ (including possibly complex g); the theorem will be recovered in the special case that $\tau_N \to 0$ and g is real.

As discussed in the text, we need only establish that the likelihood function is logconcave in a certain smoothly invertible reparameterization of θ. The proof is based on the following fact (Bogachev, 1998):

Theorem (Integrating out log-concave functions). *Let $f(x, y)$ be jointly log-concave in $x \in \Re^j$ and $y \in \Re^k$, $j, k < \infty$, and define*

$$f_0(x) \equiv \int f(x, y)dy;$$

then f_0 is log-concave in x.

To apply this theorem, we write the likelihood in the following "path integral" form:

$$L_{\{\vec{x}_i,t_i\}}(\theta) = \int p(N_t, \tau_N) \prod_i 1\Big(V_t(\vec{x}_i, \theta, N_t) \in C_i\Big) dN_t \tag{6}$$

where we are integrating over each possible path of the noise process N_t, $p(N_t, \tau_N)$ is the (Gaussian) probability measure induced on N_t under the parameter τ_N, and $1(V_t(\vec{x}_i, \theta, N_t) \in C_i)$ is the indicator function for the event that $V_t(\vec{x}_i, \theta, N_t)$ — the voltage path driven by the noise sample N_t under the model settings θ and input data $\vec{x}_i$ — is in the set C_i. Recall that C_i is defined as the convex set satisfying a collection of linear inequalities that must be satistfied by any $V(t)$ path consistent with the observed spike train $\{t_i\}$; however, the precise identity of these inequalities will not play any role below (in particular, C_i only depends on the real part of $V(t)$ and is independent of τ_N and θ).

The logic of the proof is as follows. Since the product of two log-concave functions is log-concave, $L(\theta)$ will be log-concave under some reparameterization if p and 1 are both log-concave under the same reparameterization of the variables N and θ, for any fixed τ_N. This follows by 1) approximating the full path integral by (finite-dimensional) integrals over suitably

time-discretized versions of path space, 2) applying the above integrating-out theorem, 3) noting that the pointwise limit of a sequence of (log)concave functions is (log)concave, and 4) applying the usual separability/continuity limit argument to lift the result from the arbitrarily-finely-discretized (but still finite) setting to the full (infinite-dimensional) path space setting.

To discretize time, we simply sample $V(t)$ and $N(t)$ (and bin t_i) at regular intervals Δt, where $\Delta t > 0$ is an arbitrary small parameter we will send to zero at the end of the proof. We prove the log-concavity of p and 1 in the reparameterization

$$(g, V_{leak}) \to (\alpha, I_{DC}) \equiv (e^{-g\Delta t}, gV_{leak});$$

this map is clearly smooth, but due to aliasing effects, the map $g \to \alpha$ is smoothly invertible only if the imaginary part of g satisfies $g\Delta t < 2\pi$; thus we restrict the parameter space further to $(0 \leq \text{real}(g), 0 \leq \text{imag}(g) \leq \pi(\Delta t)^{-1})$, an assumption that becomes negligible as $\Delta t \to 0$. Finally, importantly, note that this reparameterization preserves the convexity of the parameter space Θ.

Now to the proof of the log-concavity of the components of the integrand in (6). Clearly, p is the easy part: $p(N, \tau_N)$ is independent of all variables but N and τ_N; p is Gaussian in N and is thus the prototypical log-concave function.

Now for the function $1(V_t(\vec{x}_i, \theta, N_t) \in C_i)$. First, note that this function is independent of τ_N given N. Next, an indicator function for a set is log-concave if and only if the set is convex. Thus it is sufficient to prove that the set (N, θ) such that $V_t(N, \theta) \in C$ is convex, for any convex C. To see this, we write out the dependence of V_t on N and θ in operator form:

$$V_t = E_g\big[V_{reset}\delta(0) + I_{DC} + \vec{k} \cdot \vec{x}(t) + \sum_j h(t - t_j) + N_t\big],$$

where E_g, recall, is the exponential convolution operator corresponding to g. Now, the key fact is that E_g^{-1} depends linearly on α:

$$E_g = \begin{bmatrix} 1 & & & & \\ \alpha & 1 & & & \\ \alpha^2 & \alpha & 1 & & \\ & & \ddots & \ddots & \\ & \dots & \alpha^2 & \alpha & 1 \end{bmatrix},$$

while

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

as can be shown by direct computation. Thus the set (N, θ) such that $V_t(N, \theta) \in C$ can be written as the set $N \in A(\theta)C$, with $A(\theta)$ an invertible operator, affine in θ, namely

$$A(\theta)z(t) = E_g^{-1}V(t) - V_{reset}\delta(0) - I_{DC} - k \cdot \vec{x}(t) - \sum_j h(t - t_j)$$

for any $z(t) \in C$; since C, Θ, and the set of all possible N are convex, the proof is complete, because the union of the graphs of a convex set of nonsingular affine translates of a convex set is itself convex. □

We have Theorem 1 as a corollary upon restricting α (or equivalently, g) to the real axis and letting $\tau_N \to 0$, rescaling, and again noting that the pointwise limit of a sequence of (log-)concave functions is (log-)concave.

In a previous version of this manuscript, we gave a different proof, in which the key log-concavity property was established not by the result on integrating out, but rather by an appeal to the Prekopa-Rinott theorem (Bogachev, 1998; Rinott, 1976) on log-concave measures; this earlier proof relied on a somewhat complex construction of convex translations of sets and required a more involved reparameterization; the current proof seems simpler. In addition, the current proof clarifies the generality of the result, in at least two directions. First, it is clear that the proof is valid for any fixed log-concave noise measure $p(N)$ (possibly including correlations, non-Gaussianity, and nonstationarities), not just Gaussian white noise. Second, integrating over hyperparameters (e.g. in a Bayesian model selection setting (Sahani and Linden, 2003)) does not induce any local maxima as long as the log-concavity of the integrands is undisturbed. Finally, it is interesting to note that a nearly identical proof demonstrates that the likelihood of the model introduced in (Keat et al., 2001) contains no non-global local maxima, in all parameters except for the time constant τ_p of the after-potential introduced in equation (7) in (Keat et al., 2001); however, this

proof does not extend in any obvious way to the non-likelihood-based cost function minimized by Keat et al.

It is also worth noting that this proof can not directly give us log-concavity in τ_N for Gaussian densities. In fact, no Gaussian density with diagonal covariance of the form

$$\begin{bmatrix} f_1(\tau_N) & & & \\ & f_2(\tau_N) & & \\ & & \ddots & \\ & & & f_i(\tau_N) \end{bmatrix}$$

(we have in mind the covariance operator of a stationary process, expressed in the Fourier basis) can be jointly log-concave in (N, τ_N). To see this, set $N = 0$; this implies that f_i^{-1} must be of the form e^h, for h a concave function. Since the determinant of the Hessian of the function $-N^2/f_i(\tau_N) = -e^{h(\tau_N)}N^2$,

$$2e^{2h}N^2\Big(h'' - (h')^2\Big),$$

is nonpositive in general (since h is concave, i.e., $h'' \leq 0$), $-e^h N^2$ cannot be jointly concave, and this implies that the Gaussian can not be jointly log-concave, either (to see this, let $N \to \infty$). Nevertheless, it is not difficult to think of reasonable densities which are jointly log-concave in N and additional parameters like τ_N; this may prove useful in other contexts (Williams and Barber, 1998; Seeger, 2002).

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 (c.f. sections 6.1 and 6.2), 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.

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 $\{t_k\}_{1 \le k \le j}$. 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 (recall expression (3)). 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_{leak}.

Acknowledgments

We thank E.J. Chichilnisky, W. Gerstner, Z. Ghahramani, B. Lau, and S. Shoham for helpful suggestions. LP was partially supported by pre- and postdoctoral fellowships from HHMI; JWP was partially supported by a predoctoral fellowship from NSF and by an NYU Dean's Dissertation Fellowship.

References

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

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

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

Brillinger, D. (1992). Nerve cell spike train data analysis: a progression of technique. *Journal of the American Statistical Association*, 87:260–271.

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

Brunel, N. and Latham, P. (2003). Firing rate of the noisy quadratic integrate-and-fire neuron. *Neural Computation*, 15:2281–2306.

Burkitt, A. and Clark, G. (1999). Analysis of integrate-and-fire neurons: Synchronization of synaptic input and spike output. *Neural Computation*, 11:871–901.

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

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

Dayan, P. and Abbott, L. (2001). *Theoretical Neuroscience*. MIT Press.

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

Donoho, D. and Elad, M. (2003). Optimally sparse representation in general (nonorthogonal) dictionaries via l^1 minimization. *PNAS*, 100:2197–2202.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., and Picard, D. (1995). Wavelet shrinkage: Asymptopia? *J. R. Statist. Soc. B.*, 57(2):301–337.

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

Efron, B. and Stein, C. (1981). The jackknife estimate of variance. *Annals of Statistics*, 9:586–596.

Fourcaud, N. and Brunel, N. (2002). Dynamics of the firing probability of noisy integrate-and-fire neurons. *Neural Computation*, 14:2057–2110.

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

Gerstner, W. and Kistler, W. (2002). *Spiking Neuron Models: Single Neurons, Populations, Plasticity.* Cambridge University Press.

Girosi, F. (1998). An equivalence between sparse approximation and support vector machines. *Neural Computation*, 10:1455–1480.

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

Haskell, E., Nykamp, D., and Tranchina, D. (2001). Population density methods for large-scale modelling of neuronal networks with realistic synaptic kinetics. *Network*, 12:141–174.

Iyengar, S. and Liao, Q. (1997). Modeling neural activity using the generalized inverse Gaussian distribution. *Biological Cybernetics*, 77:289–295.

Izhikevich, E. (2001). Resonate-and-fire neurons. *Neural Networks*, 14:883–894.

James, W. and Stein, C. (1960). Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1:361–379.

Joeken, S., Schwegler, H., and Richter, C. (1997). Modeling stochastic spike train responses of neurons: An extended wiener series analysis of pigeon auditory nerve fibers. *Biological Cybernetics*, 76:153–162.

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

Karlin, S. and Taylor, H. (1981). *A Second Course in Stochastic Processes*. Academic Press, New York.

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

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

Koch, C. (1999). *Biophysics of Computation*. Oxford University Press.

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

Machens, C., Wehr, M., and Zador, A. (2003). Spectro-temporal receptive fields of subthreshold responses in auditory cortex. *NIPS*.

Miller, K. and Troyer, T. (2002). Neural noise can explain expansive, power-law nonlinearities in neural response functions. *Journal of Neurophysiology*, 87:653–659.

Miller, M. and Mark, K. (1992). A statistical study of cochlear nerve discharge patterns in response to complex speech stimuli. *Journal of the Acoustical Society of America*, 92:202–209.

Moreno, R., de la Rocha, J., Renart, A., and Parga, N. (2002). Response of spiking neurons to correlated inputs. *Physical Review Letters*, 89:288101.

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

Paninski, L., Fellows, M., Shoham, S., Hatsopoulos, N., and Donoghue, J. (2003a). Nonlinear population models for the encoding of dynamic hand position signals in primary motor cortex. *Annual Computational Neuroscience Meeting, Alicante, Spain*, Poster presentation.

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

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

Plesser, H. and Gerstner, W. (2000). Noise in integrate-and-fire neurons: From stochastic input to escape rates. *Neural Computation*, 12:367–384.

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

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

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

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

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

Seeger, M. (2002). PAC-Bayesian generalisation error bounds for Gaussian process classifiers. *Journal of Machine Learning Research*, 3:233–269.

Seshardri, V. (1993). *The inverse Gaussian distribution*. Clarendon, Oxford.

Simoncelli, E., Paninski, L., Pillow, J., and Schwartz, O. (to appear 2004). Characterization of neural responses with stochastic stimuli. In Gazzaniga, M., editor, *The Cognitive Neurosciences*. MIT Press, 3rd edition.

Smyth, D., Willmore, B., Baker, G., Thompson, I., and Tolhurst, D. (2003). The receptive-field organization of simple cells in primary visual cortex of ferrets under natural scene stimulation. *Journal of Neuroscience*, 23:4746–4759.

Snyder, D. and Miller, M. (1991). *Random Point Processes in Time and Space*. Springer-Verlag.

Stevens, C. and Zador, A. (1996). When is an integrate-and-fire neuron like a Poisson neuron? *NIPS*, 8:103–109.

Stevens, C. and Zador, A. (1998). Novel integrate-and-fire-like model of repetitive firing in cortical neurons. *Proceedings of the 5th joint symposium on neural computation, UCSD*.

Strong, S. Koberle, R., de Ruyter van Steveninck R., and Bialek, W. (1998). Entropy and information in neural spike trains. *Physical Review Letters*, 80:197–202.

Tipping, M. (2001). Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244.

Truccolo, W., Eden, U., Fellows, M., Donoghue, J., and Brown, E. (2003). Multivariate conditional intensity models for motor cortex. *Society for Neuroscience Abstracts*.

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

van der Vaart, A. (1998). *Asymptotic statistics*. Cambridge University Press, Cambridge.

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

Warland, D., Reinagel, P., and Meister, M. (1997). Decoding visual information from a population of retinal ganglion cells. *Journal of Neurophysiology*, 78:2336–2350.

Williams, C. and Barber, D. (1998). Bayesian classification with Gaussian processes. *IEEE PAMI*, 20:1342–1351.

Yu, Y. and Lee, T. (2003). Dynamical mechanisms underlying contrast gain control in single neurons. *Physical Review E*, 68:011901.

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