Network: Comput. Neural Syst. **15** (2004) 243–262 PII: S0954-898X(04)75780-2

Maximum likelihood estimation of cascade point-process neural encoding models

Liam Paninski

Gatsby Computational Neuroscience Unit, University College London, UK

E-mail: liam@gatsby.ucl.ac.uk

Received 5 February 2004, accepted for publication 13 August 2004
Published 21 September 2004
Online at stacks.iop.org/Network/15/243
doi:10.1088/0954-898X/15/4/002

Abstract

Recent work has examined the estimation of models of stimulus-driven neural activity in which some linear filtering process is followed by a nonlinear, probabilistic spiking stage. We analyze the estimation of one such model for which this nonlinear step is implemented by a known parametric function; the assumption that this function is known speeds the estimation process considerably. We investigate the shape of the likelihood function for this type of model, give a simple condition on the nonlinearity ensuring that no non-global local maxima exist in the likelihood—leading, in turn, to efficient algorithms for the computation of the maximum likelihood estimator—and discuss the implications for the form of the allowed nonlinearities. Finally, we note some interesting connections between the likelihood-based estimators and the classical spike-triggered average estimator, discuss some useful extensions of the basic model structure, and provide two novel applications to physiological data.

1. Introduction

A central issue in computational neuroscience is the experimental characterization of the functional relationship between external variables—e.g., sensory stimuli or motor behavior—and neural spike trains. Because the nervous system is probabilistic, any description we can provide will be necessarily statistical: given some experimentally observable signal x (an image or an auditory stimulus), we want to be able to provide the probability of a given spike train y. More precisely, we want to estimate the conditional probabilities $p(y|x)$, for as large a set of observable signals x as possible. Of course, there are typically far too many possible x to characterize these probabilities directly; thus our real goal is to find a model, some functional form that allows us to predict $p(y|x)$ even for x we have never observed directly. Ideally, such a model will be both accurate and easy to estimate given 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 x is projected onto a low-dimensional subspace, followed by a nonlinear, probabilistic spike generation stage (see, e.g., Simoncelli *et al* (2004) for a review). The linear filtering stage is typically interpreted as the neuron's 'spatiotemporal 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.

The most widely used model of this type is the linear–nonlinear Poisson (LNP) cascade model (see Simoncelli *et al* (2004) for a partial list of references), in which spikes are generated according to an inhomogeneous Poisson process, with rate determined by an instantaneous ('memoryless') nonlinear function of the filtered input. This model has a number of desirable features, including conceptual simplicity and computational tractability. Additionally, reverse correlation analysis provides a simple unbiased estimator for the linear filter under certain conditions (Chichilnisky 2001), and the properties of estimators (for both the linear filter and static nonlinearity) have been thoroughly analyzed, even for the case of highly non-symmetric or 'naturalistic' stimuli (Weisberg and Welsh 1994, Paninski 2003, Sharpee *et al* 2004).

More recent work has focused on extending this simple model to include spike-history effects, such as refractoriness, burstiness or adaptation (Berry and Meister 1998, Keat *et al* 2001). In particular, we have recently described a flexible, biophysically plausible model that can be fit more efficiently (Paninski *et al* 2004d). Nevertheless, the existing algorithms for consistently fitting this type of generalized cascade model are still somewhat slow.

Here we examine a cascade model whose generality is comparable to that of these recent methods, but for which much faster fitting procedures are available. The estimation procedure is based on maximum likelihood (ML), and the models we consider will have the special property that the likelihood surface for the model parameters has no non-global local minima, ensuring that ascent algorithms will compute the ML estimator (MLE) in a stable, efficient manner. We give examples of these no-local-minima models, describe the basic properties of the corresponding ML estimators, point out some connections to other models that have appeared in the literature, and provide two novel applications to physiological data.

2. The model and likelihood function

We consider the following parametric form of the LNP model,

$$p(\text{spike}|\vec{x}) = f_\theta(K\vec{x}),$$

where $\vec{x}$ is the input signal, θ is some finite-dimensional parameter that sets the properties of the nonlinearity f; K is a linear operator that projects the high-dimensional vector $\vec{x}$ into a manageable subspace of low (say m) dimensions, and hence the nonlinear properties of the cell are defined just by the m-dimensional nonlinear behavior of f_θ (where f_θ is clearly constrained to be positive for all θ). This is known in the statistics literature as a generalized linear model (GLM, McCullagh and Nelder (1989)); examples of similar models that have appeared in the neuroscience literature are reviewed in Simoncelli *et al* (2004).

The input signal $\vec{x}$ is typically assumed to be some externally observable signal: a visual signal (Brenner *et al* 2001, Chichilnisky 2001), for example, or a time-varying, intracellularly injected current (Paninski *et al* 2003) or the dynamic position of the hand in a motor behavioral context (Paninski 2003). In these cases, the linear stage K could consist of a collection of spatiotemporal receptive fields, temporal current filters or preferred hand positions and velocities, respectively, and the firing rate nonlinearity f_θ would in each case model, e.g., rectification or other nonlinear combinations of the filtered input signal, with θ setting parameters such as the rectification threshold or sharpness of the nonlinearity. Note that in the above cases the input $\vec{x}$ can be either deterministically or stochastically controlled by the experimenter (in the case of sensory or intracellular current stimulation) or even not directly controlled by the experimenter at all (the motor behavior case).

It is worth emphasizing that $\vec{x}$ can also include 'internal' signals, such as the activity of a population of simultaneously recorded neighbor cells (Tsodyks *et al* 1999, Harris *et al* 2003, Paninski *et al* 2004b, Truccolo *et al* 2003, Nykamp 2003) or the time since the last spike of the neuron under study (this last example allows us to include the Markov interval-type models of, e.g., Brown *et al* (2002) in our discussion). In this case the linear pre-filter K could model, e.g., the synaptic transfer function which converts the activity in pre-synaptic neighboring cells into post-synaptic temporal patterns of excitation and inhibition.

Finally, $\vec{x}$ does not have to be strictly linearly related to the input signal of interest; fixed nonlinear transformations (Dodd and Harris 2002, Sahani 2000) are permissible as well; this kind of fixed mapping amounts basically to a redefinition of the input signal. The gain in this redefinition is that the new inputs $\vec{x}$ can be of higher dimension than the original inputs, possibly allowing for better classification of the excitatory and inhibitory parts of the input signal (Cristianini and Shawe-Taylor 2000). For example, we could include not just the original inputs $\vec{x}$, but also the outputs of a Wiener-like expansion of $\vec{x}$: $\left\{x_i^2\right\}$, $\{x_i x_j\}$ and so on.

We can define the likelihood function for this model in a straightforward manner. The spiking process is a conditionally inhomogeneous Poisson process with rate $f_\theta(K\vec{x}(t))$ given the inputs $\vec{x}(t)$ and the parameters (K, θ); thus, according to general point-process theory (Snyder and Miller 1991) the log-likelihood of observing spikes at times $\{t_i\}$ is

$$L(K,\theta) \sim \sum_i \log f_\theta(K\vec{x}(t_i)) - \int f_\theta(K\vec{x}(t))\,\mathrm{d}t, \tag{1}$$

where the integral is over the length of the experiment (Dayan and Abbott 2001, Brown *et al* 2002). The first term penalizes low firing rates f at the times t_i when spikes were observed, while the second penalizes high firing rates at all other times (when no spike was observed). Note that the gradient of the above likelihood can be computed easily given the gradient of f_θ, implying that algorithms to ascend the likelihood surface are efficient in finding (local) maxima.

It might be helpful for the reader unfamiliar with the continuous-time formulation to compare to the discrete-time version of this model, in which the log-probability of observing a binary spike train $\{t_i\}$ in bins of width $\mathrm{d}t$ (where $\mathrm{d}t$ is assumed small enough that no more than one spike is observed per bin, and that fluctuations in the model firing rate $f(K\vec{x}(t))$ on a $\mathrm{d}t$ timescale can be neglected) is given by

$$L_{\text{discrete}}(K,\theta) = \sum_i \log(f_\theta(K\vec{x}(t_i))\,\mathrm{d}t) + \sum_{i'} \log(1 - f_\theta(K\vec{x}(t_{i'}))\,\mathrm{d}t),$$

with i' indexing the times at which no spike was observed; it is not hard to see (after expanding the logarithm) that the two formulae become similar as $\mathrm{d}t \to 0$ (Snyder and Miller 1991, Dayan and Abbott 2001). We will deal exclusively with the continuous-time formulation from now on, as the corresponding mathematics turn out to be simpler and somewhat more elegant.

3. The likelihood condition

Given the definition of the LNP model, the obvious question is: Does the likelihood surface contain any local extrema that would prevent an ascent algorithm from efficiently locating the true global maximum (the MLE)? Our main result is a simple condition guaranteeing that no non-global local maxima exist in the parameters (θ, K), no matter what data $\vec{x}_i$ and t_i are observed. This condition is quite easy to derive, given the form of the likelihood expression (1):

Condition. *$f_\theta(\vec{u})$ is jointly convex in θ and $\vec{u}$, and $\log f_\theta(\vec{u})$ is jointly concave in θ and $\vec{u}$; the parameter space of possible (K, θ) is convex.*

The condition clearly implies the concavity of the likelihood (since concavity is preserved under addition); the lack of non-global local maxima follows immediately, from standard properties of concave functions. Similar conditions for other GLMs have been considered elsewhere (Wedderburn 1976, Haberman 1977, McCullagh and Nelder 1989). Note also that the condition on K is typically not a problem: K naturally takes values in a convex set, the vector space of possible linear filters (although in certain situations it is useful to restrict elements of K to be positive, for example; this does not affect the convexity).

We should note that this condition is sufficient but not necessary; it is possible, in some cases, to rule out the existence of non-global local minima in the log-likelihood without guaranteeing its concavity (as was basically our approach in deriving the condition). Nevertheless, we expect this condition to be 'nearly' necessary in the sense that any less restrictive assumption (i.e., any condition that does not imply concavity) will be a great deal more complicated and difficult to apply in practice. Of course, one could also argue quite plausibly that our condition is too strong from a physiological (if not necessarily a rigorous mathematical) sense—what we really want is to rule out the existence of 'bad' local minima for 'most' reasonable data, not necessarily *all* possible non-global local minima for *any* possible data set, as established here—but we restrict our attention in this work to statements which may be proved mathematically, postponing the discussion of these more qualitative (though in the end, more important) questions for future detailed investigations on real physiological data.

In the remainder of this section, we pursue some of the implications of the above condition, which turns out to be perhaps stronger than is at first apparent. We assume throughout that f is non-constant (that is, the cell's responses are in fact at least somewhat dependent on its inputs $\vec{x}$). Also remember that f must be non-negative, by construction.

We consider scalar $f(u)$ first, for simplicity; in particular, we ignore the θ-dependence for now, for reasons that will become clear below. Then it is not hard to see that $f(u)$ must

- be monotonic in u (we assume for simplicity that $f(u)$ grows monotonically);
- grow at least linearly (and at most exponentially) as a function of u;
- decay at least exponentially as $u \to -\infty$;
- have a derivative, $f'(u)$, which must be not only monotonically increasing, but also continuous everywhere except possibly at the point $u_0 = \sup\{u : f(u) = 0\}$, and strictly positive for all $u > u_0$;
- vanish either everywhere on the interval $(-\infty, u_0)$ or nowhere ($u_0 = -\infty$).

Thus both symmetric and saturating ('sigmoidal') nonlinearities f are not allowed under our condition (of course, the fact that a symmetric nonlinearity would induce at least two local maxima in the likelihood is obvious in retrospect). Unfortunately, this rules out both 'squashing'-type nonlinearities and the standard quadratic model for complex cells in primary visual cortex (Simoncelli and Heeger 1998). Saturating nonlinearities can be recovered using a trick described below (section 7), but nonmonotonic nonlinearities must be modeled at an earlier stage, by effectively changing the definition of $\vec{x}$ to include nonmonotonic functions of $\vec{x}$ (e.g. x_i^2; cf section 2), possibly at the expense of a larger number of parameters in K.

The condition turns out to be even more restrictive in the general case of vectors $\vec{u}$ and θ. It turns out that all $f(\vec{u}, \theta)$ which satisfy the condition must have a basically one-dimensional structure: since $f(\vec{u}, \theta)$ is convex, the sets

$$A_z \equiv \{(\vec{u}, \theta) : f_\theta(\vec{u}) \leqslant z\}$$

are convex (where $\vec{u}$ abbreviates $K\vec{x}$). However, f is also log-concave (that is, $\log f$ is a concave function), implying

$$A^z \equiv \{(\vec{u}, \theta) : f_\theta(\vec{u}) \geqslant z\}$$

are convex as well, and this forces the contours of f to be linear,

$$f_\theta(K\vec{x}) = f_0(\vec{k}\vec{x} + b)$$

for some one-dimensional f_0 satisfying the above description, a *single* vector $\vec{k}$, and a scalar b. In other words, the parameter θ has been reduced to a simple scalar offset term b (in fact, $\theta = b$ can be eliminated entirely, or at least absorbed into $\vec{k}$, by the standard trick of assuming an additional constant input to the cell and setting the corresponding element of $\vec{k}$ equal to b). This effectively one-dimensional nature of f clearly rules out any energy-type models (Simoncelli and Heeger 1998), or models with other nonlinear symmetries (invariances), without modifications in the definition of $\vec{x}$ before the LN stage of the model.

4. Examples

Despite these rather stringent constraints, it is not difficult to think of functions f for which our no-local-extrema condition holds (remember that linear functions are both convex and concave):

- $f(u) = \mathrm{e}^u$ (Martignon *et al* 2000, Truccolo *et al* 2003, Paninski *et al* 2004b);
- $f(u) = \begin{cases} \mathrm{e}^u & u < 0 \\ 1 + u & u \geqslant 0 \end{cases}$ (Harris *et al* 2003);
- $f(u) = \begin{cases} 0 & u < 0 \\ u^\alpha & u \geqslant 0 \end{cases}$ with $\alpha \geqslant 1$ (Anderson *et al* 2000, Chichilnisky 2001, Miller and Troyer 2002, Hahnloser *et al* 2003).

In addition, all products of the above functions are acceptable, since concavity (log-concavity) is closed under addition (multiplication) and the positive, increasing, convex functions are closed under multiplication. Unfortunately, however, the acceptable space (of increasing, convex and log-concave functions; call it $\mathcal{F}$) is not closed with respect to addition; $\log \mathcal{F}$, on the other hand, is closed with respect to addition, but not with respect to multiplication by constants <1 (and is therefore not a vector space).

A more convenient, not overly restrictive subset of $\mathcal{F}$ may be constructed as follows. Let

$$f(u) = \int^u \mathrm{e}^{h(v)}\, \mathrm{d}v$$

for some increasing, concave function $h(u)$. Clearly, any f of this form is convex, since $f' = \mathrm{e}^{h(u)}$ is increasing; it is also log-concave by the Prekopa–Rinott theorem (Rinott 1976). The set $\mathcal{H}$ of h satisfying the conditions forms a vector cone: $\mathcal{H}$ is closed under addition, multiplication by non-negative scalars, and translation. Finally, the remainder set $\mathcal{F} \cap \mathcal{H}^c$ is fairly small: it takes some thought to produce a log-concave, convex, increasing function f satisfying the constraint $(\log f'(u))'' > 0$ for some u.[1]

5. Sampling properties of the MLE; connections to spike-triggered averaging

In the above, we have developed a class of models for which the ML estimator is especially computationally tractable. However, we have not yet said much about how good an estimator the MLE actually is—for example, does the MLE asymptotically provide the correct $\vec{k}$? If not, how large is the asymptotic bias? Li and Duan (1989) studied conditions under which the MLE for a GLM is consistent (that is, such that the MLE provides asymptotically accurate estimates of the parameter $\vec{k}$, given enough data), even when the 'link' function f is chosen incorrectly (that is, when we fit the responses with a model f that does not correspond exactly to the true response properties of the cell under question). We adapt and extend their results here.

Assume the observed spike train is generated by a GLM with rate function g, but that we apply the MLE based on the incorrect rate function f. Our results will be stated in terms of an input probability distribution, $p(\vec{x})$, from which in the simplest case the experimenter draws independent and identically distributed inputs $\vec{x}$, but which in general is just the 'empirical distribution' of $\vec{x}$, the observed distribution of all inputs $\vec{x}$ presented to the cell during the course of the experiment. Define the important concept of an 'elliptically symmetric' density (Li and Duan 1989, Chichilnisky 2001, Paninski 2003): $p(\vec{x})$ is elliptically symmetric if $p(\vec{x}) = q(\|A\vec{x}\|_2)$ for some scalar function $q(\cdot)$, some matrix A, and the usual two-norm $\|\cdot\|_2$; that is, p is constant on the ellipses defined by fixing $\|A\vec{x}\|_2$. (The canonical example of an elliptically symmetric density is the Gaussian with mean zero and covariance C, for which A may be chosen as $C^{-1/2}$, a square root of the inverse covariance matrix; the case to keep in mind is the radially, or spherically, symmetric case, in which A is proportional to the identity and the elliptic symmetries above become spherical.) Then we have

Proposition 1. *The MLE based on any convex and log-concave f is consistent almost surely for any true underlying g, provided $p(\vec{x})$ is elliptically symmetric and the spike-triggered mean, $E_{p(\vec{x}|\mathrm{spike})}\vec{x}$, is different from zero.*

In other words, given a symmetry condition on the input distribution $p(\vec{x})$, an asymmetry condition on g and enough data, the MLE based on f will always give us the true $\vec{k}$. In particular, again, the assumption on $p(\vec{x})$ holds if $\vec{x}$ is drawn from any Gaussian distribution (e.g., in a Gaussian white-noise-type experiment). The condition on the spike-triggered mean refers to the conditional distribution $p(\vec{x}|\mathrm{spike})$ of $\vec{x}$ given that a spike was observed: the spike-triggered mean is easily computed in this case (Chichilnisky 2001, Paninski 2003) as $E_{p(\vec{x}|\mathrm{spike})}\vec{x} = Z\int p(\vec{x})\vec{x}g(\vec{k}_0\vec{x})\,\mathrm{d}\vec{x}$, where Z is a positive scalar and $\vec{k}_0$ denotes the true underlying linear projection $\vec{k}$. (Note that the original operator parameter K has been reduced to the single vector $\vec{k}_0$ here, according to our discussion in section 3.)

[1] One simple example is $f(u) = 0, u < 0$, $f(u) = u + g(u), u \geqslant 0$, with

$$g'' = R([0, 1]),$$

with $R(\cdot)$ the windowed linear rectifier $R(u) = 0, u < 0$; $R(u) = u, u < 1$; $R(u) = 0, u \geqslant 1$. Clearly, f is increasing and convex. That f is log-concave but f' is not takes a little more effort, but is fairly straightforward, using

$$(\log f'(u))'' > 0 \to (f''/f')' > 0 \to \frac{f'f''' - (f'')^2}{(f')^2} > 0 \to f''' > \frac{(f'')^2}{f'}.$$

We present the proof here in the body of the text to emphasize both its simplicity and its similarity to the corresponding proof for the classical estimator for this type of cascade model, the spike-triggered average (STA, Chichilnisky (2001), Paninski (2003)); note that the input distribution $p(\vec{x})$ is assumed here and throughout this paper to be centered (have mean zero), which may be enforced in general via a simple change of variables.

Proof. General likelihood theory (van der Vaart 1998) states that ML estimators asymptotically maximize $E(L(\vec{k}, b))$, the expectation of the likelihood function (1) under the true data distribution. We need to prove that this function has a unique maximum at $\alpha\vec{k}_0$, for some $\alpha \neq 0$. We have

$$E(L(\vec{k}, b)) = \int p(\vec{x})[g(\vec{k}_0\vec{x}) \log f(\vec{k}\vec{x} + b) - f(\vec{k}\vec{x} + b)]\, \mathrm{d}\vec{x}.$$

(We have implicitly assumed that the integrals above exist; no further assumptions on g are necessary.) The key fact about this function is that it is concave in $(\vec{k}, b)$ and, after suitable change of variables (multiplication by a whitening matrix), symmetric with respect to reflection about the $\vec{k}_0$ axis. This immediately implies that a maximizer lies on this axis (i.e., is of the form $\alpha\vec{k}_0$ for some scalar α); the strict convexity of f or $-\log f$ implies that any such maximizer is unique.

It only remains to prove that $\alpha \neq 0$. Assume otherwise. Then the gradient of $E(L(\vec{k}, b))$ must vanish at $\vec{k} = \vec{0}$,

$$f'(b) \int p(\vec{x})\vec{x} \left[\frac{g(\vec{k}_0\vec{x})}{f(b)} - 1\right] = \vec{0},$$

this implies that either $f'(b) = 0$ for the optimal offset scalar b or

$$\int p(x_0)x_0[g(x_0) - f(b)]\, \mathrm{d}x_0 = 0,$$

where $x_0 \equiv \vec{k}_0 x$. By symmetry of $p(x_0)$,

$$\int p(x_0)x_0 f(b) = f(b) \int p(x_0)x_0 = 0,$$

hence, $\int p(x_0)x_0 g(x_0)$ must vanish as well, contradicting our assumption on the spike-triggered mean. Thus we only need to rule out $f'(b) = 0$. Recall from our discussion in the last section that f' is monotonically increasing (by convexity), and $f'(u)$ is strictly greater than 0 whenever $f(u) > 0$. Thus we only need to show that $f(b) \neq 0$ for the optimal setting of b (remembering that we have assumed temporarily that $\vec{k} = \vec{0}$); plugging in $f = 0$ above gives that the expected likelihood is negative infinity whenever g is not identically zero, an obvious contradiction. □

Li and Duan (1989) gave a slightly weaker condition guaranteeing that a maximizer lies on the $\vec{k}_0$ axis; however, their condition is only met for all possible unknown directions $\vec{k}$ if $p(\vec{x})$ is elliptically symmetric (Eaton 1986, Paninski 2003). They did not examine conditions preventing the null case $\alpha = 0$. As in section 3, this result on the MLE holds quantitatively (that is, for all g and all $p(\vec{x})$ satisfying the conditions); the conditions can presumably be weakened, at the expense of some simplicity, while still allowing for more qualitative statements (say, the MLE is at worst weakly biased for most ‘reasonable’ g and $p(\vec{x})$). (In addition, of course, one could argue that the bias induced by misspecification of g might be small compared to the error resulting from the assumption that real data may be modeled in the simple cascade form considered here; some steps toward remedying this latter, ‘large-scale’ model error will be considered in the following sections.)

As noted above, proposition 1 bears a striking similarity to the main result for the STA (Bussgang 1952, Chichilnisky 2001, Paninski 2003); the conditions ensuring their asymptotic accuracy are exactly equivalent (and by much the same symmetry argument). This leads us to study the similarities of these two methods more carefully.

We base our discussion on the solution to the equations obtained by setting the gradient of the likelihood to zero. The MLE solves

$$\frac{1}{T}\sum_i \vec{x}_i \frac{f'}{f}(\vec{k}_{\mathrm{MLE}}\,\vec{x}_i + b_{\mathrm{MLE}}) = \int p(\vec{x})f'(\vec{k}_{\mathrm{MLE}}\,\vec{x} + b_{\mathrm{MLE}})\vec{x}, \tag{2}$$

with T the length of the experiment. In the case of elliptically symmetric stimuli, the right-hand side converges to a vector proportional to $\sigma^2(p)\vec{k}_{\mathrm{MLE}}$ (recall that f' is monotonically increasing), where $\sigma^2(p)$ denotes the covariance matrix of the input distribution $p(\vec{x})$. The left-hand side, on the other hand, is itself a kind of weighted STA—an average of (weighted) spike-triggered stimuli $\vec{x}$—with the weight $\frac{f'}{f}(\vec{k}\vec{x}_i + b)$ positive but monotonically decreasing in $\vec{k}\vec{x}_i$, by the log-concavity of f. (We interpret this weight as a 'robustness' term, decreasing the strength of very large—possibly outlying—$\vec{x}$.)

Thus, denoting the left-hand side as $\vec{k}_{\mathrm{WSTA}}$, for weighted STA, we have that the MLE asymptotically behaves like

$$\vec{k}_{\mathrm{MLE}} = \sigma^2(p)^{-1}\vec{k}_{\mathrm{WSTA}},$$

this is exactly analogous to the 'rotated STA', $\vec{k}_{\mathrm{RSTA}} \equiv \sigma^2(p)^{-1}\vec{k}_{\mathrm{STA}}$, the basic correlation-corrected estimator for cascade models (Paninski 2003). Also note that, in the exponential case (the maximally convex case, recall section 3), the weight $\frac{f'}{f}(\vec{k}\vec{x}_i + b)$ is constant for all $\vec{x}$. Thus, in the case of elliptically symmetric $p(\vec{x})$, the RSTA and the exponential MLE are exactly equivalent (this, in turn, gives an interesting alternate proof of the consistency of the RSTA). More generally, the RSTA provides a useful starting point for iterative maximization of the GLM likelihood.

We can pursue this robustness idea further: How does the bias of the MLE based on f behave when $p(\vec{x})$ is asymmetric and the data are generated by a different rate function g? (Of course, as usual, the MLE is well behaved—asymptotically normal and optimal in several natural senses—when f is in fact the correct rate function. The asymptotic normality is retained more generally when f is chosen incorrectly, but optimality is not (Li and Duan 1989).) Taking the expectation of the likelihood gradient equation (2), we can see that the MLE asymptotically solves

$$\int p(\vec{x})\vec{x}f'(\vec{k}_{\mathrm{MLE}}\,\vec{x} + b_{\mathrm{MLE}})\left(\frac{g(\vec{k}_0\vec{x})}{f(\vec{k}_{\mathrm{MLE}}\,\vec{x} + b_{\mathrm{MLE}})} - 1\right) = 0.$$

It is difficult to solve this equation for general $p(\vec{x})$; however, we do know the solution for elliptically symmetric $p(\vec{x})$ (i.e., $\vec{k}_0$, up to a scale constant α), and can use a perturbative approach around this known solution to develop insight about the more general case. In particular, we want to know what happens to the solution when a symmetric $p(\vec{x})$ is perturbed by a small but in general asymmetric measure $\epsilon q(\vec{x})$.

For simplicity, assume that b is fixed at zero, $p(\vec{x})$ is radially (not just elliptically) symmetric, and normalize $\vec{k}_0$ (these assumptions entail no loss of generality, since the first can be met by a simple redefinition of f, and the latter two by a simple change of basis of $\vec{x}$). Define α as the asymptotic length of $\vec{k}_{\mathrm{MLE}}$, given by the solution in α of

$$\int p(x_0)x_0 f'(\alpha x_0)\left(\frac{g(x_0)}{f(\alpha x_0)} - 1\right) = 0$$

(recall $x_0 \equiv \vec{k}_0 x$). Now one reasonable way of measuring the susceptibility of the MLE to bias is to compute

$$\alpha^{-1} \left. \frac{\partial \vec{k}_1(q,\epsilon)}{\partial \epsilon} \right|_0,$$

the ratio of the length α to the rate of change of $\vec{k}_{\mathrm{MLE}}\,(q,\epsilon)$ in a direction $\vec{k}_1$ orthogonal to $\vec{k}_0$; recall from proposition 1 that $\vec{k}_1(q,0) = 0$ for any $\vec{k}_1 \perp \vec{k}_0$, so this ratio measures something like the angular error induced in $\vec{k}_{\mathrm{MLE}}$ by q for small ϵ. (As usual with LN cascade-type models, angular measures of error are more appropriate than absolute error, since the model is only defined up to a scale factor: changes in the scale of $\vec{k}$ can be absorbed easily by changes in the scale of f.) A straightforward Taylor expansion gives

$$\left. \frac{\partial \vec{k}_1}{\partial \epsilon} \right|_0 = \frac{1}{2C} \int q(\vec{x}) x_1 f'(\alpha x_0) \left(\frac{g(x_0)}{f(\alpha x_0)} - 1 \right),$$

with the curvature

$$C \equiv - \int p(\vec{x}) x_1^2 \left(\frac{f f''(\alpha x_0) - f'(\alpha x_0)^2}{f(\alpha x_0)^2} g(x_0) - f''(\alpha x_0) \right).$$

(The curvature C is exactly the component of the Fisher information matrix orthogonal to $\vec{k}_0$, and therefore sets the inverse scale of the asymptotic error in the bias-free case; see, e.g., Paninski (2003) for further details.) The simplest case here is when q is supported on a single point, $\vec{y}$, in which case the above reduces to

$$\left. \frac{\partial \vec{k}_1}{\partial \epsilon} \right|_0 = \frac{y_1}{2C} f'(\alpha y_0) \left(\frac{g(y_0)}{f(\alpha y_0)} - 1 \right),$$

with y_0 and y_1 denoting the projection of $\vec{y}$ onto $\vec{k}_0$ and $\vec{k}_1$, respectively. Again, it might help to keep the exponential case in mind, for which the above simplifies to

$$\left. \frac{\partial \vec{k}_1}{\partial \epsilon} \right|_0 = \frac{1}{2C} \int q(\vec{x}) x_1 (g(x_0) - \mathrm{e}^{\alpha x_0}),$$

with

$$C = \int p(\vec{x}) x_1^2 \, \mathrm{e}^{\alpha x_0},$$

and

$$\left. \frac{\partial \vec{k}_1}{\partial \epsilon} \right|_0 = \frac{y_1}{2C} (g(y_0) - \mathrm{e}^{\alpha y_0})$$

in the point-support q case. (In any event, the complicated parenthetical term in the definition of C is always nonpositive, by f convexity and log-concavity and g positivity.) The above should be compared to

$$\frac{\int q(\vec{x}) x_1 g(x_0)}{\int p(\vec{x}) x_0 g(x_0)}, \tag{3}$$

or

$$\frac{y_1 g(y_0)}{\int p(\vec{x}) x_0 g(x_0)},$$

the bias susceptibility corresponding to the STA in the general and point-support q case, respectively.

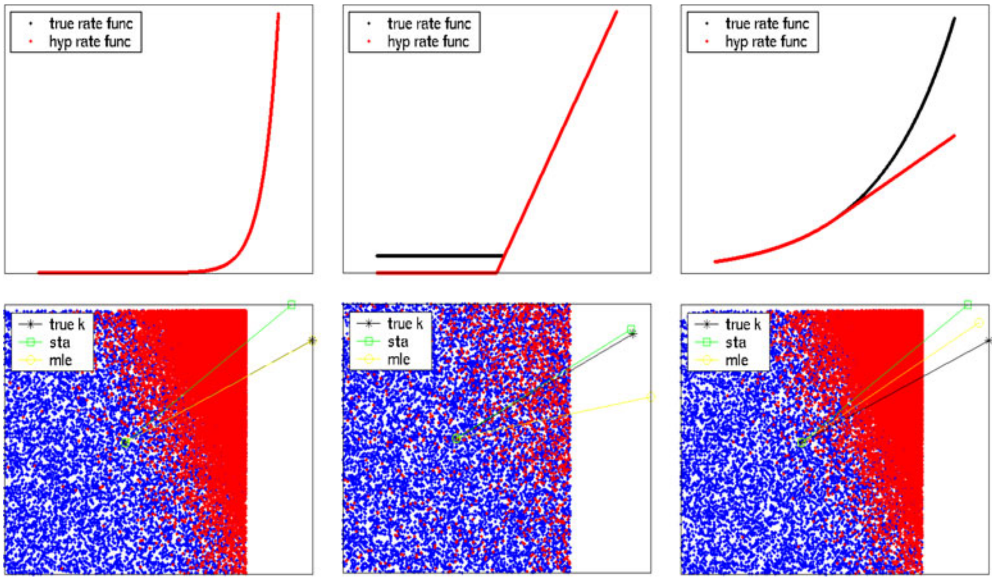

Figure 1. The bias of the MLE depends on both the true and the hypothetical rate function $f(u)$ in simulations. The true filter $\vec{k}$ is held fixed (bottom panels; black line). Left: the hypothetical f is chosen to match the true f generating the data (top panel; x-axis denotes u, with y-axis denoting $f(u)$), and the MLE recovers $\vec{k}$ almost perfectly (in contrast to the STA, which is strongly biased upward; note that $p(\vec{x})$—uniformly distributed on the square here—is not elliptically symmetric, although the covariance $\sigma^2(p)$ is proportional to the identity matrix). Red dots indicate $\vec{x}_i$ for which spikes were observed; blue dots all other $\vec{x}$ ($\vec{x}$ taken to be two-dimensional in this example, for illustrative purposes). The green and yellow lines indicate the corresponding STA and MLE, respectively. Middle: the STA is always biased toward the northeast corner in this case (as can easily be shown analytically via the STA bias formula (3) (Chichilnisky 2001, Simoncelli *et al* 2004)), but the MLE can be biased in the opposite direction if the hypothesized $f(u)$ decays more quickly than the true rate function as $u \to -\infty$. Right: the mixed exponential-linear function of Harris *et al* (2003) is less prone to bias than the STA for superlinear true rate functions f.

We have three main conclusions. First, the MLE can be biased in the opposite direction as the STA: restricting our attention to the point-mass case, the STA is always biased in the direction of y_1, while the MLE can be biased toward or away from y_1, depending on the sign of $g(y_0) - f(\alpha y_0)$. Second, the MLE based on rate functions $f(u)$ with sub-exponential growth as u grows is, roughly speaking, less bias-prone; the denominator term C in the above expressions is of larger magnitude in general for non-exponential rates, while the numerator is more tame (due to the decreasing nature of f'/f). This, in turn, implies that sub-exponential rates should be less bias-prone than the RSTA, if we recall the similarities between the exponential MLE and the RSTA. Finally, rate functions $f(u)$ which decay more quickly than exponentially as $u \to -\infty$ can lead to strongly biased ML estimators, particularly when $\frac{f'g}{f}(x_0) \gg 1$; see figure 1. (We should emphasize that the examples shown in figure 1 are meant to be strictly qualitative; in physiological applications, the bias will depend quantitatively on the specific cell's preferences, the input distribution $p(\vec{x})$ and so on.)

These robustness considerations, taken together, recommend the mixed exponential-linear function used in Harris *et al* (2003) (section 4) in case of highly asymmetric inputs $p(\vec{x})$, where bias might be a concern; recall that this function grows linearly as $u \to \infty$ (as slowly as possible given our convexity constraint, leading to a well-controlled f'/f weight term) and decays exponentially as $u \to -\infty$ (as slowly as possible under log-concavity). However, it is important to note that there is no hypothetical rate f for which the MLE is unbiased for all true rates g; as in the case of the STA, the bias is always dependent on the true g (figure 1), and to obtain a bias-free estimator for $\vec{k}$ in general we must consider the slower but more general class of semiparametric estimators which adapt to different g as well (Weisberg and Welsh 1994, Paninski 2003, Sharpee *et al* 2004).

6. Multiplicative models of spike-history dependence

In the next two sections, we will point out some connections of the above likelihood-based techniques to related models that have appeared in the literature. The first such model we will consider was developed in order to correct the main deficiency of the basic LNP model, namely its Poisson nature. A simple way to model the spike-history effects (e.g., refractoriness, burstiness) that such a Poisson model ignores is to introduce a multiplicative term, dependent on the time since the last spike, to modulate the firing rate; this leads to a kind of 'linear–nonlinear recovery' model (Berry and Meister 1998, Brown *et al* 2002, Paninski 2003). It turns out that this type of model may be solved using likelihood-based tools which are mathematically quite similar to those developed above.

We begin with an arbitrary model (including, but not limited to, the LNP models considered above; the following results can be stated quite generally). For a given input $\vec{x}$, this model will produce a time-varying predicted firing rate $f_{\vec{x}}(t)$. Now we try to fix up this model by including a multiplicative 'recovery' term modeling the effect of the cell's recent spike history, $r(t - t_{i-1})$, with $(t - t_{i-1})$ the time since the last spike; thus, our new model predicts a firing rate

$$f_1(t) = f_{\vec{x}}(t) r(t - t_{i-1}).$$

Now how to choose r? As before, a reasonable approach is maximum likelihood: the log-likelihood for this process, given a spike train $\{t_i\}$, is again of the form

$$\sum_i \log(f_1(t_i)) - \int f_1(t)\,\mathrm{d}t.$$

Rewrite this as

$$\sum_i \log(f_{\vec{x}}(t_i) r(t_i - t_{i-1})) - \sum_i \int_{t_{i-1}}^{t_i} f_{\vec{x}}(t) r(t - t_{i-1})\,\mathrm{d}t$$

$$= Z + \sum_i \left(\log r(t_i - t_{i-1}) - \int_{t_{i-1}}^{t_i} f_{\vec{x}}(t) r(t - t_i)\,\mathrm{d}t \right),$$

where Z is constant with respect to r whenever $f_{\vec{x}}$ is fixed (i.e., when we are maximizing over all possible history dependence terms r given a fixed base model $f_{\vec{x}}$).

This log-likelihood is quite convenient. First, it is concave in the recovery function r, implying that no non-global local maxima exist (since r takes values in a convex set, the set of all non-negative functions). In fact, by setting the functional gradient of the likelihood with respect to r to zero, we can find the ML estimate for r in closed form: $r_{\mathrm{MLE}}(\tau)$ solves

$$\sum_{i: t_i - t_{i-1} = \tau} \frac{1}{r_{\mathrm{MLE}}(\tau)} = \sum_{i: t_i - t_{i-1} \geqslant \tau} f_{\vec{x}}(t_i + \tau),$$

or written more suggestively,

$$r_{\mathrm{MLE}}(\tau) = \frac{p(\tau)}{\int \mathrm{d}p(f) f(\tau) \int_\tau^\infty p(t|f)\,\mathrm{d}t},$$

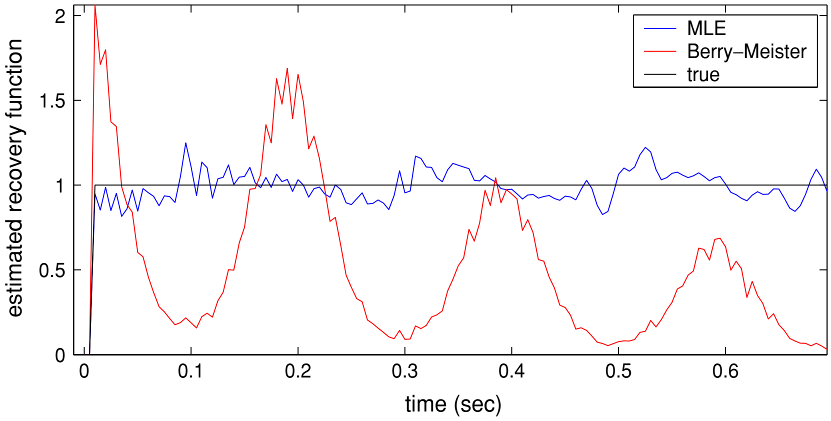

Figure 2. Comparison of estimates of recovery function $r(\tau)$ given simulated data. Black trace: true underlying r; blue: ML estimate; red: estimate due to (Berry and Meister 1998) (i.e., $r_{\text{BM}}(\tau) = c^{-1}p(\tau)\,\mathrm{e}^{c\tau}$, with $p(\tau)$ the observed inter-spike interval density and c the observed mean firing rate). True underlying firing rate $f_{\vec{x}}(t)$ chosen here to vary sinusoidally at 5 Hz; recovery function models an absolute refractory period of 10 ms, followed by complete recovery (i.e., $r(\tau) = 0$ for $\tau \leqslant 10$ ms, $r(\tau) = 1$, $\tau > 10$). Note that the MLE matches true function fairly well, but the Berry–Meister estimate is contaminated by oscillations in $f_{\vec{x}}(t)$ and is strongly biased downward for large delays τ.

where $p(\tau)$ denotes the observed inter-spike interval (ISI) density, $p(f)$ is the observed density of $f(t_i + \tau)$, and $p(t|f)$ the observed density of a spike at time t given f.

In the case that $p(f)$ is concentrated at the point $f_{\vec{x}}(t) \equiv c$ (i.e., our original model is that the spike train is a Poisson process with rate c, independent of the stimulus), we have the commonsense solution that

$$r_{\text{MLE}}(\tau) = \frac{p(\tau)}{c\,\mathrm{e}^{-c\tau}},$$

the ratio of the observed ISI density $p(\tau)$ to $c\,\mathrm{e}^{-c\tau}$, the ISI density we would expect of a Poisson process of rate c; thus, the MLE agrees in this very special case with the formula given in Berry and Meister (1998) (though in general, of course, the formulae do not agree; that is, the estimator of Berry and Meister (1998) is not generally the MLE; see figure 2 for a comparison).

Also note that if we take our model to be $f_1(t) = f_{\vec{x}}(t) r_{\text{MLE}}(t - t_{i-1})$, plugging into the formula above, we get that the new $r_{\text{MLE}}(\tau) \equiv 1$, as hoped for; thus, iterating the above procedure does not gain us anything new unless, of course, a new model for $f_{\vec{x}}(t)$ is chosen given the new estimated refractory term r_{MLE}. For example, we could alternately fit an LNP model for $\vec{k}$, given r (using the obvious modification of the original LNP log-likelihood (1),

$$L_r(\vec{k}, b) \sim \sum_i \left(\log f(\vec{k}\vec{x}(t_i) + b) + \log r(t_i - t_{i-1}) - \int_{t_{i-1}}^{t_i} f(\vec{k}\vec{x}(t) + b) r(t - t_{i-1})\,\mathrm{d}t \right),$$

which retains its concavity in $\vec{k}$ and b under our condition, since r is non-negative), then fit r given f, iterating until convergence (since both procedures ascend the likelihood surface, convergence is guaranteed).

7. Extension to integrate-and-fire-like models

It is worthwhile to consider the relationship between the generalized linear models under discussion here and the more biophysically motivated models employed in studies of intracellular dynamics (in particular, there is some evidence (Reich *et al* 1998) that models of the multiplicative form considered in the last section do not adequately capture the inter-spike interval statistics of visual neurons). One connection is provided by the following model: consider the inhomogeneous Poisson process with rate given by $f(V(t)+b)$, where f is a convex, log-concave scalar function, b is a scalar, and $V(t)$ is the solution of the 'leaky-integrator' differential equation

$$\frac{\partial V}{\partial t} = -gV(t) + \vec{k}\cdot\vec{x}(t) + \sum_{j=0}^{i-1} h(t-t_j),$$

with initial value

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

namely

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

the linear convolution of the input current with a negative exponential of time constant $1/g$. Here g denotes the membrane leak conductance, $\vec{k}\cdot\vec{x}(t)$ the projection of the input signal $\vec{x}(t)$ onto the spatiotemporal linear kernel $\vec{k}$, and h is a post-spike current waveform which is summed over the previously observed spikes. As usual, in the absence of input, V decays back to 0 with time constant $1/g$. We allow h, in turn, to take values in some low-dimensional vector space; this allows the shape and magnitude of h to be varied to fit different intrinsic spiking patterns (including burstiness, adaptation, saturation, etc) (Gerstner and Kistler 2002, Paninski *et al* 2004d).

$V(t)$ in the above is the subthreshold, and therefore unobserved ('hidden'), solution of the usual leaky integrate-and-fire (LIF) equation (Dayan and Abbott 2001), for which the voltage resets to V_{reset} after the spike is emitted at t_i. Alternatively, we could define $V\left(t_i^+\right) = V\left(t_i^-\right) - V_r$, that is, subtract a fixed amount V_r from V after each spike. This induces a conditional positive dependence between spikes—if V is large before the spike, it will remain large after the spike—which is not present in the standard LIF model, but which may be more accurate in certain cases. (Note that really this latter case is just a mixed finite- and infinite-impulse-response linear filter model, with the spike train $\sum_i \delta(t-t_i)$ treated as an additional input in $\vec{x}$.)

We have written the above equations to emphasize the similarity to the form of the 'spike-response model' introduced by Gerstner and colleagues (Gerstner and Kistler 2002, Jolivet *et al* 2003) and employed in Paninski *et al* (2004d), (2004c) to model extracellularly recorded spike-train responses. The combined IF–GL model described above is conceptually identical to a simple version of the 'escape-rate' approximation to the noisy LIF-type model given in Plesser and Gerstner (2000), Gerstner and Kistler (2002) (see also Stevens and Zador (1996)); this escape-rate approximation, in turn, was introduced to partially alleviate the difficulties associated with computing the passage time density and firing rate of the LIF model driven by noise (again, see Paninski *et al* (2004d) for more details).

Thus this 'IF–GLM' can be seen as a direct approximation to the noisy LIF model developed in Paninski *et al* (2004d) (which in turn is a tractable approximation to more detailed biophysical models for spiking dynamics). The main result of interest in the context of this paper is that this IF–GL model shares a great deal of the tractability of the noisy LIF cell: the log-likelihood for the IF–GLM is jointly concave in the parameters $\{\vec{k}, h, V_{\text{reset}}, b\}$, for any data $\{\vec{x}, t_i\}$, ensuring the tractability of the MLE. (The log-likelihood is not necessarily concave in g here, but one-dimensional maximizations are eminently tractable; it is worth noting that the convolution kernel e^{-gt} can be generalized as well, to possibly nonstationary kernels (Stevens and Zador 1998, Jolivet *et al* 2003), at the expense of the possible addition of a few more parameters.) See section 10 for a simple application of the IF–GLM to physiological data.

8. Regularization of GLM ML estimates

As emphasized in section 2, the ease of computation and maximization in the GLMs considered here make it possible to include a large number of (possibly nonlinearly transformed) input variables $\vec{x}$ which might increase the modeling flexibility of the basic linear–nonlinear cascade model. However, 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 in a linear regression setting). How do we control for overfitting in the current context?

The likelihood-based methods discussed here can be adapted quite easily for this purpose. 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

$$L(\vec{k}) - Q(\vec{k})$$

instead of just $L(\vec{k})$, where $L(\vec{k})$ is the log-likelihood, 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). If Q is taken to be convex, 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 behavior.

Thus we are free to choose Q as we like within the class of smooth convex functions, bounded below. If $-Q$ peaks at the point $\vec{k} = \vec{0}$, the MAP estimator will basically be a more 'conservative' version of the MLE, with the chosen coefficients shifted nonlinearly toward zero. (Of course, other prior information—for example, we might believe $\vec{k}$ has a certain degree of smoothness—can be built into Q as well.) This type of 'shrinkage' estimator has been extremely well studied, from a variety of viewpoints (James and Stein 1960, Donoho *et al* 1995, Klinger 1998, Tipping 2001, Ng 2004), 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), Harris *et al* (2003) for some illustrations of this effect. One particularly simple choice for Q is the weighted L_α^α-norm

$$Q(\vec{k}) = \sum_j b_j |k_j|^\alpha, \qquad \alpha \geqslant 1,$$

where j indexes the elements of $\vec{k}$, while the weights b_j 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).

9. Bayesian spike-train decoding

The convexity condition introduced in section 3 also has some useful implications for Bayesian decoding of spike-train information. This is interesting especially in the context of neural prosthetic design (Serruya *et al* 2002, Donoghue 2002), but also more generally in the analysis of neural codes (Rieke *et al* 1997).

Spike-train decoding concerns the following problem: given a spike-train $\{t_i\}$ (we consider only single spike-trains for simplicity, but the generalization to population spike-train decoding should be clear), how can we estimate the input signal $\vec{x}$? The Bayesian approach is to examine the *a posteriori* distribution

$$p(\vec{x}|\{t_i\}) \sim p(\vec{x})L_{\vec{k}}(\{t_i\}|\vec{x}),$$

with $L_{\vec{k}}(\{t_i\}|\vec{x})$ denoting the likelihood of observing $\{t_i\}$ given $\vec{x}$ under the model defined by $\vec{k}$. The important thing to note here is that if $\vec{x}$ is drawn from a log-concave prior distribution $p(\vec{x})$, then the posterior distribution is also log-concave, and thus has no non-local global maxima. This log-concavity follows from arguments nearly identical to those in section 3 applied to $\vec{x}$ instead of $\vec{k}$ (plus the fact that log-concavity is preserved under multiplication), and as in the model estimation context, this lack of local minima greatly simplifies decoding via the MAP estimator for $\vec{x}$. In addition, log-concavity makes Laplace approximation—quadratic approximation of the log-posterior density (e.g. for construction of confidence intervals)—a reasonable approach (Barbieri *et al* 2004).

10. Demonstration on real data

We give a few applications to physiological data here to further illustrate the ideas discussed above.

10.1. Prediction of in vitro responses

Our first example illustrates the utility and ease of modeling refractory behavior in the context of these cascade models. For simplicity, we examine data discussed in Paninski *et al* (2003); in an *in vitro* rat cortical slice preparation, we injected a white-noise current into a cell and recorded the voltage responses (under dual whole-cell patch clamp; see Paninski *et al* (2003) and references therein for full physiological details). Spike times were defined by thresholding the bandpass-filtered voltage trace. (Recall that the temporal details of spike-trains under these stimulus conditions are extremely reproducible (Mainen and Sejnowski 1995, Tiesinga *et al* 2003).)

We then fit two models: the standard LNP model and the IF–GLM, both using the same time-varying stimulus current as the input $\vec{x}$. Results are shown in figure 3; the main conclusion is that the IF–GLM permits quite accurate prediction of the observed spike-train; moreover, the predictions of the IF–GLM are much more precise than are those of the simple LNP model, in large part because the refractory effects of the post-spike current h serve to enhance the reproducibility of the resulting spike-train (Joeken *et al* 1997, Berry and Meister 1998, Reich *et al* 1998, Brown *et al* 2002).

10.2. Regularization of predictions from multineuronal in vivo responses

Our second example is somewhat more subtle: instead of a simple stimulus current, now we take the input $\vec{x}$ to be the activity of other simultaneously active cells in the cortical network

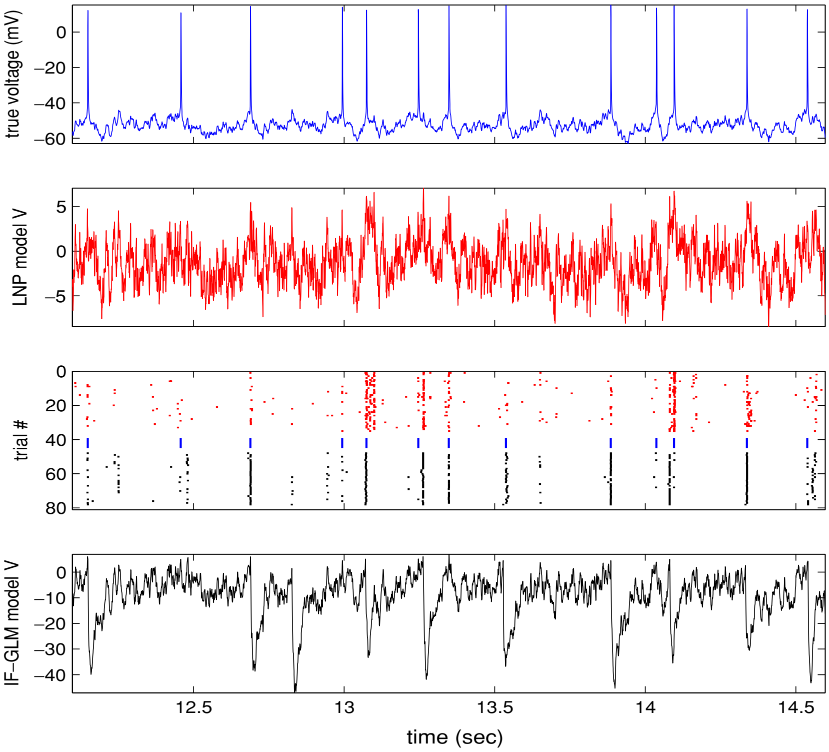

Figure 3. Predicting *in vitro* responses using the LNP and IF–GLM models given intracellularly injected current stimulus. Top: true observed intracellular voltage (display is a randomly chosen segment of a 100 s long white-noise experiment). B: $V(t)$ signal of estimated LNP model for corresponding time period; $V(t) = \vec{k}\vec{x}(t)$ here, with $\vec{k}$ estimated from training data and $\vec{x}(t)$ the vector formed by collecting time-delayed copies of the white-noise stimulus current $I(t)$ (not shown). An exponential model for the rate function $f(V) = p(\text{spike}|\vec{x})$ was used here, so $V(t)$ corresponds to log-firing rate at time t. Note that the model captures the subthreshold voltage fluctuations fairly well (remember scale differences). C: predicted and true spike time rasters. Red is LNP model (repeated independent, identically distributed draws from the inhomogeneous Poisson model with rate $f(V(t))$), blue is true data (single trial; this is just a thresholded version of the top voltage trace), and black is IF–GLM (repeated draws from Poisson model with rate given by IF–GLM $V(t)$, which, unlike the LNP $V(t)$, is dependent on spike times, due to the after-current h, and therefore changes stochastically from trial to trial). Recall that rasters should, on average, give exponential of red and black $V(t)$ signals; note the greatly improved spike timing in the IF–GLM model. Bottom: $V(t)$ signal of estimated IF–GLM model, for a single trial (corresponding to the bottom raster in third panel). Note the post-spike episodes of extremely low log-firing rate (i.e., relative refractory periods) induced by the hyperpolarizing post-spike current h (recall that both h and $\vec{k}$ are estimated from data for this model). Also note the approximately fourfold difference in scales between the LNP and IF–GLM $V(t)$ traces, indicating a comparable difference in the relative log-likelihoods of the two models.

in vivo (Tsodyks *et al* 1999, Maynard *et al* 1999, Harris *et al* 2003). We recorded simultaneously from 21 isolated single cells in the arm area of the primary motor cortex (MI) of a monkey performing a visually guided manual tracking task (Paninski *et al* 2004a), then attempted to predict the activity of each cell given only the concurrent activity of the cell's 20 neighbors (where, for simplicity, the jth element of the input $\vec{x}$ is defined here as the number of spikes from cell j in a single, relatively large 500 ms window).

In particular, we were interested in the effect of regularization on the predictive power of a GLM based on this neighbor activity. Therefore we compared the cross-validated likelihood

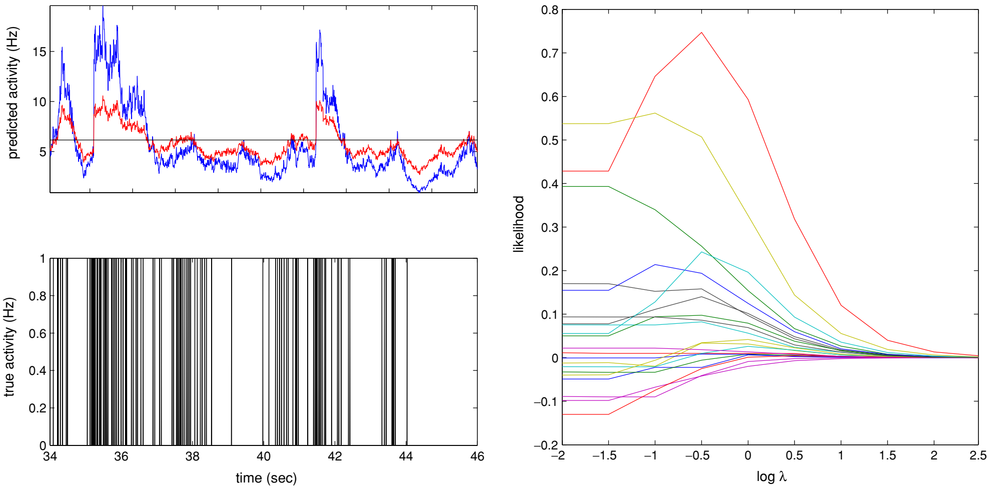

Figure 4. Regularization of predictions from multineuronal *in vivo* data. Left: illustration of shrinkage effect in MAP estimate. Blue and red traces indicate firing rate prediction as a function of time, of the MLE and MAP ($\lambda = 0.5$) estimator, respectively, for a single cell; flat black line gives baseline firing rate. Bottom panel gives observed spike-train on this trial. The red curve (MAP) clearly fluctuates less than the blue (MLE), which for this cell turns out to lead to better predictions, in a likelihood sense (this cell corresponded to the top red trace in the right panel). Right: likelihoods of the models estimated via the ML and MAP rules, for various values of λ (with λ multiplied by $(\log T)/T$, with $T \approx 400$ the length of the experiment in seconds; this scaling was roughly inspired by the Schwarz information criterion for model selection (Schwarz 1978)). Likelihoods are shown on a log scale, with baseline firing rate prediction subtracted off (i.e., $L \leqslant 0$ implies that it would have been better just to set $\vec{k} = 0$ than to use $\vec{k}_{\text{MLE}}$ or $\vec{k}_{\text{MAP}}$). Each curve corresponds to a different cell; all logs indicated are of natural base.

of the MLE on test spike-train data to that of the MAP estimator, constructed (as in section 8) as the maximizer of the penalized likelihood

$$L(\vec{k}) - \lambda \sum_j b_j |k_j|^{\alpha},$$

with λ a free parameter we varied systematically, α fixed at 2 for simplicity, and b chosen to standardize the data, $b_j = \sigma(p)_j$, the standard deviation of $\vec{x}_j$ (Klinger 1998) (the data were centered before applying the analysis to remove mean effects). Clearly, if $\lambda = 0$, the MLE is recovered (the zero-regularization case); as $\lambda \to \infty$, the MAP estimator $\vec{k}_{\text{MAP}}(\lambda)$ is over-regularized and shrinks to $\vec{0}$ (cf figure 4).

Three different behaviors are evident when we plot the cross-validated likelihood of the MAP estimator as a function of λ: we see strictly decreasing likelihood curves (in which regularization always hurts the estimator); strictly increasing curves (in which the unregularized ML estimator is complete noise, performing worse than the baseline $L = 0$, and regularization simply drops the estimator to zero as λ increases); and, most interestingly, curves with a well-defined peak for $0 < \lambda < \infty$, for which regularization increases predictive accuracy (in some cases quite a bit) and there is an optimal degree of regularization. In particular, some cells appear to be completely uninformative with no regularization ($L(0) < 0$), but reveal their structure (i.e., the log-likelihood ratio becomes positive) as λ increases. It is also interesting to note that the optimal λ, when it exists, takes values within a relatively restricted range ($\lambda \approx 0.5$), and could easily be chosen by cross-validation, as in Machens *et al* (2003), Smyth *et al* (2003). In short, these regularization techniques can significantly increase the predictive power of this type of multineuronal model, even in the case of relatively few cells (by the most recent simultaneous recording standards (Warland *et al* 1997, Harris *et al* 2003, Nicolelis *et al* 2003)); we expect these techniques to play a larger role in future studies of population coding *in vivo*.

11. Discussion

We have provided a theory for likelihood-based estimation of point-process models of 'generalized linear' or 'cascade' form. This theory turns out to be somewhat more subtle than the corresponding theory for, say, Wiener-series estimation (which is basically described by the usual least-squares theory (Dayan and Abbott 2001)), but nevertheless most of our results turned out to be built on a single main concept: the concavity of the log-likelihood function. This simple concavity idea leads directly to the form of rate functions f we allowed (section 3); the asymptotic bias properties of the MLE, and the connections between the MLE and STA—in particular the result that the GLM MLE is in some sense a generalization of the STA, with an exact correspondence in the case of exponential f (section 5); and the simple and direct approach to shrinkage-type regularization of the MLE via MAP estimation with log-concave priors (sections 8 and 10.2).

These concavity results lead directly to (in fact, were motivated by the desire for) highly efficient techniques for computing the ML and MAP estimators: the likelihoods of the GLMs studied here have no non-global local maxima and can be computed (along with their gradients) quite rapidly, ensuring the success of the usual ascent algorithms. It is worthwhile comparing the resulting estimators to other commonly employed techniques for modeling spike-train data. Each method has its own drawbacks: linear and Wiener-series methods (Joeken *et al* 1997, Sahani and Linden 2003, Machens *et al* 2003, Smyth *et al* 2003) are fast but inflexible (since sampling concerns typically restrict the Wiener polynomial expansion to second order, which is often insufficient to capture the nonlinear behavior of real cells), and the fact that these models are not defined probabilistically leads to some awkwardness when attempting to sample from the models, or to precisely define the likelihood of a given spike-train over another; semiparametric techniques for learning f together with $\vec{k}$ (Weisberg and Welsh 1994, Paninski 2003, Sharpee *et al* 2004) are extremely flexible but suffer from the existence of many local maxima, as does the filter-based model estimator proposed by (Keat *et al* 2001); correlation-based methods depend on overly restrictive conditions on their validity (Berry and Meister 1998, Chichilnisky 2001, Paninski 2003); the approach developed in Paninski *et al* (2004d), while bypassing the previous concerns, requires the iterative computation of a likelihood function consisting of high-dimensional probability integrals which, while tractable (they can be computed without resorting to Monte Carlo, or multiple passes over the data), are roughly an order of magnitude slower to calculate than the GLM likelihoods considered here. Finally, while the GLM MLE can, like the correlation-based STA, be significantly biased, the analysis of section 5 indicates that these bias problems can be made somewhat less problematic than for the STA (and are nonexistent in at least two cases: either when the observed rate function happens to match the chosen f well or, as for the STA, in the case of elliptically symmetric inputs $p(\vec{x})$).

In short, we see the GLM approach—especially when augmented with the integrate-and-fire-like spike-history dependence discussed in section 7 and applied in section 10.1—as a very attractive alternative to the available techniques for modeling spike-train responses given high-dimensional inputs in cases where the underlying rate function can reasonably be modeled as monotonic. We are currently in the process of further applications of these models to physiological data recorded both *in vivo* and *in vitro*, in order to assess whether they accurately account for the stimulus preferences and spiking statistics of real neurons.

Acknowledgments

We thank J Pillow, W Truccolo, and E Simoncelli for interesting discussions, and P Latham for helpful comments on an earlier draft. The slice data were collected in collaboration with B Lau and A Reyes; the motor cortical data were collected in collaboration with M Fellows, N Hatsopoulos and J Donoghue. This work was supported by the Howard Hughes Medical Institute and the Gatsby Charitable Trust.

References

Anderson J, Lampl I, Gillespie D and Ferster D 2000 The contribution of noise to contrast invariance of orientation tuning in cat visual cortex *Science* **290** 1968–72

Barbieri R, Frank L, Nguyen D, Quirk M, Solo V, Wilson M and Brown E 2004 Dynamic analyses of information encoding in neural ensembles *Neural Computation* **16** 277–307

Berry M and Meister M 1998 Refractoriness and neural precision *J. Neurosci.* **18** 2200–11

Brenner N, Bialek W and de Ruyter van Steveninck R 2001 Adaptive rescaling optimizes information transmission *Neuron* **26** 695–702

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** 325–46

Bussgang J 1952 Crosscorrelation functions of amplitude-distorted Gaussian signals *RLE Tech. Rep.* 216

Chichilnisky E 2001 A simple white noise analysis of neuronal light responses *Network: Comput. Neural Syst.* **12** 199–213

Cristianini N and Shawe-Taylor J 2000 *An Introduction to Support Vector Machines* (Cambridge: Cambridge University Press)

Dayan P and Abbott L 2001 *Theoretical Neuroscience* (Cambridge, MA: MIT Press)

Dodd T and Harris C 2002 Identification of nonlinear time series via kernels *Int. J. Syst. Sci.* **33** 737–50

Donoghue J 2002 Connecting cortex to machines: recent advances in brain interfaces *Nature Neurosci.* **5** 1085–8

Donoho D L, Johnstone I M, Kerkyacharian G and Picard D 1995 Wavelet shrinkage: Asymptopia? *J. R. Stat. Soc.* B **57** 301–37

Eaton M 1986 A characterization of spherical distributions *J. Multivariate Anal.* **20** 272–6

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

Haberman S 1977 Maximum likelihood estimation in exponential response models *Ann. Stat.* **5** 815–41

Hahnloser R, Seung S and Slotine J 2003 Permitted and forbidden sets in symmetric threshold-linear networks *Neural Computation* **15** 621–38

Harris K, Csicsvari J, Hirase H, Dragoi G and Buzsaki G 2003 Organization of cell assemblies in the hippocampus *Nature* **424** 552–6

James W and Stein C 1960 Estimation with quadratic loss *Proc. 4th Berkeley Symp. on Mathematical Statistics and Probability* vol 1, pp 361–79

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 *Biol. Cybern.* **76** 153–62

Jolivet R, Lewis T and Gerstner W 2003 The spike response model: a framework to predict neuronal spike-trains *Artificial Neural Networks and Neural Information Processing (ICANN/ICONIP 2003, Istanbul, 26–29 June 2003)* ed O Kaynak *et al* (*Springer Lecture Notes in Computer Science* vol 2714) (Berlin: Springer) pp 846–53

Keat J, Reinagel P, Reid R and Meister M 2001 Predicting every spike: a model for the responses of visual neurons *Neuron* **30** 803–17

Klinger A 1998 High-dimensional generalized linear models *PhD Thesis* University of Munich

Li K and Duan N 1989 Regression analysis under link violation *Ann. Stat.* **17** 1009–52

Machens C, Wehr M and Zador A 2003 Spectro-temporal receptive fields of subthreshold responses in auditory cortex *Neural Information Processing Systems*

Mainen Z and Sejnowski T 1995 Reliability of spike timing in neocortical neurons *Science* **268** 1503–6

Martignon L, Deco G, Laskey K, Diamond M, Freiwald W and Vaadia E 2000 Neural coding: higher-order temporal patterns in the neuro-statistics of cell assemblies *Neural Computation* **12** 2621–53

Maynard E, Hatsopoulos N, Ojakangas C, Acuna B, Sanes J, Normann R and Donoghue J 1999 Neuronal interactions improve cortical population coding of movement direction *J. Neurosci.* **19** 8083–93

McCullagh P and Nelder J 1989 *Generalized Linear Models* (London: Chapman and Hall)

Miller K and Troyer T 2002 Neural noise can explain expansive, power-law nonlinearities in neural response functions *J. Neurophysiol.* **87** 653–9

Ng A 2004 Feature selection, L_1 vs. L_2 regularization, and rotational invariance *Proc. 21st. Int. Conf. on Machine Learning* to be published

Nicolelis M, Dimitrov D, Carmena J, Crist R, Lehew G, Kralik J and Wise S 2003 Chronic, multisite, multielectrode recordings in macaque monkeys *Proc. Natl Acad. Sci. USA* **100** 11041–6

Nykamp D 2003 Reconstructing stimulus-driven neural networks from spike times *Neural Information Processing Systems* vol 15, pp 309–16

Paninski L 2003 Convergence properties of some spike-triggered analysis techniques *Netw. Comput. Neural Syst.* **14** 437–64

Paninski L, Fellows M, Hatsopoulos N and Donoghue J 2004a Spatiotemporal tuning properties for hand position and velocity in motor cortical neurons *J. Neurophysiol.* **91** 515–32

Paninski L, Fellows M, Shoham S, Hatsopoulos N and Donoghue J 2004b Superlinear population encoding of dynamic hand trajectory in primary motor cortex *J. Neuroscience* at press

Paninski L, Lau B and Reyes A 2003 Noise-driven adaptation: in vitro and mathematical analysis *Neurocomputing* **52** 877–83

Paninski L, Pillow J and Simoncelli E 2004c Comparing integrate-and-fire-like models estimated using intracellular and extracellular data *Neurocomputing* at press

Paninski L, Pillow J and Simoncelli E 2004d Maximum likelihood estimation of a stochastic integrate-and-fire neural model *Neural Computation* at press

Plesser H and Gerstner W 2000 Noise in integrate-and-fire neurons: from stochastic input to escape rates *Neural Computation* **12** 367–84

Reich D, Victor J and Knight B 1998 The power ratio and the interval map: spiking models and extracellular recordings *J. Neurosci.* **18** 10090–104

Rieke F, Warland D, de Ruyter van Steveninck R and Bialek W 1997 *Spikes: Exploring the Neural Code* (Cambridge, MA: MIT Press)

Rinott Y 1976 On convexity of measures *Ann. Probab.* **4** 1020–6

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 *Neural Information Processing Systems* vol 15

Schwarz G 1978 Estimating the dimension of a model *Ann. Stat.* **7** 461–4

Serruya M, Hatsopoulos N, Paninski L, Fellows M and Donoghue J 2002 Instant neural control of a movement signal *Nature* **416** 141–2

Sharpee T, Rust N and Bialek W 2004 Analyzing neural responses to natural signals: maximally informative dimensions *Neural Computation* **16** 223–50

Simoncelli E, Paninski L, Pillow J and Schwartz O 2004 Characterization of neural responses with stochastic stimuli *The Cognitive Neurosciences* 3rd edn, ed M Gazzaniga (Cambridge, MA: MIT Press)

Simoncelli E P and Heeger D J 1998 A model of neuronal responses in visual area MT *Vis. Res.* **38** 743–61

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 *J. Neurosci.* **23** 4746–59

Snyder D and Miller M 1991 *Random Point Processes in Time and Space* (Berlin: Springer)

Stevens C and Zador A 1996 When is an integrate-and-fire neuron like a Poisson neuron? *Neural Information Processing Systems* vol 8, pp 103–9

Stevens C and Zador A 1998 Novel integrate-and-fire-like model of repetitive firing in cortical neurons *Proc. 5th Joint Symp. on Neural Computation (University of California, San Diego)* available at http://www.sloan.salk.edu/∼zador/rep_fire_inc/rep_fire_inc.html

Tiesinga P, Fellous J and Sejnowski T 2003 Attractor reliability reveals deterministic structure in neuronal spike-trains *Neural Computation* **14** 1629–50

Tipping M 2001 Sparse Bayesian learning and the relevance vector machine *J. Mach. Learn. Res.* **1** 211–44

Truccolo W, Eden U, Fellows M, Donoghue J and Brown E 2003 Multivariate conditional intensity models for motor cortex *Society for Neuroscience Abstracts* **29** 607.11

Tsodyks M, Kenet T, Grinvald A and Arieli A 1999 Linking spontaneous activity of single cortical neurons and the underlying functional architecture *Science* **286** 1943–6

van der Vaart A 1998 *Asymptotic Statistics* (Cambridge: Cambridge University Press)

Warland D, Reinagel P and Meister M 1997 Decoding visual information from a population of retinal ganglion cells *J. Neurophysiol.* **78** 2336–50

Wedderburn R 1976 On the existence and uniqueness of the maximum likelihood estimator for certain generalized linear models *Biometrika* **63** 27–32

Weisberg S and Welsh A 1994 Adapting for the missing link *Ann. Stat.* **22** 1674–700