ARTICLE Communicated by Paul Bressloff

Computation in a Single Neuron: Hodgkin and Huxley Revisited

Blaise Agüera y Arcas
blaisea@princeton.edu
Rare Books Library, Princeton University, Princeton, NJ 08544, U.S.A.

Adrienne L. Fairhall
fairhall@princeton.edu
NEC Research Institute, Princeton, NJ 08540, and Department of Molecular Biology, Princeton, NJ 08544, U.S.A.

William Bialek
wbialek@princeton.edu
NEC Research Institute, Princeton, NJ 08540, and Department of Physics, Princeton, NJ 08544, U.S.A.

A spiking neuron "computes" by transforming a complex dynamical input into a train of action potentials, or spikes. The computation performed by the neuron can be formulated as dimensional reduction, or feature detection, followed by a nonlinear decision function over the low-dimensional space. Generalizations of the reverse correlation technique with white noise input provide a numerical strategy for extracting the relevant low-dimensional features from experimental data, and information theory can be used to evaluate the quality of the low–dimensional approximation. We apply these methods to analyze the simplest biophysically realistic model neuron, the Hodgkin–Huxley (HH) model, using this system to illustrate the general methodological issues. We focus on the features in the stimulus that trigger a spike, explicitly eliminating the effects of interactions between spikes. One can approximate this triggering "feature space" as a two-dimensional linear subspace in the high-dimensional space of input histories, capturing in this way a substantial fraction of the mutual information between inputs and spike time. We find that an even better approximation, however, is to describe the relevant subspace as two dimensional but curved; in this way, we can capture 90% of the mutual information even at high time resolution. Our analysis provides a new understanding of the computational properties of the HH model. While it is common to approximate neural behavior as "integrate and fire," the HH model is not an integrator nor is it well described by a single threshold.

1 Introduction

On short timescales, one can conceive of a single neuron as a computational device that maps inputs at its synapses into a sequence of action potentials or spikes. To a good approximation, the dynamics of this mapping are determined by the kinetic properties of ion channels in the neuron's membrane. In the 50 years since the pioneering work of Hodgkin and Huxley, we have seen the evolution of an ever more detailed description of channel kinetics, making it plausible that the short time dynamics of almost any neuron we encounter will be understandable in terms of interactions among a mixture of diverse but known channel types (Hille, 1992; Koch, 1999). The existence of so nearly complete a microscopic picture of single-neuron dynamics brings into focus a very different question: What does the neuron compute? Although models in the Hodgkin–Huxley (HH) tradition define a dynamical system that will reproduce the behavior of the neuron, this description in terms of differential equations is far from our intuition about—or the formal description of—computation.

The problem of what neurons compute is one instance of a more general problem in modern quantitative biology and biophysics: Given a progressively more complete microscopic description of proteins and their interactions, how do we understand the emergence of function? In the case of neurons, the proteins are the ion channels, and the interactions are very simple: current flows through open channels, charging the cell's capacitance, and all channels experience the resulting voltage. Arguably, there is no other network of interacting proteins for which the relevant equations are known in such detail; indeed, some efforts to understand function and computation in other networks of proteins make use of analogies to neural systems (Bray, 1995). Despite the relative completeness of our microscopic picture for neurons, there remains a huge gap between the description of molecular kinetics and the understanding of function. Given some complex dynamic input to a neuron, we might be able to simulate the spike train that will result, but we are hard pressed to look at the equations for channel kinetics and say that this transformation from inputs to spikes is equivalent to some simple (or perhaps not so simple) computation such as filtering, thresholding, coincidence detection, or feature extraction.

Perhaps the problem of understanding computational function in a model of ion channel dynamics is a symptom of a much deeper mathematical difficulty. Despite the fact that all computers are dynamical systems, the natural mathematical objects in dynamical systems theory are very different from those in the theory of computation, and it is not clear how to connect these different formal schemes. Finding a general mapping from dynamical systems to their equivalent computational functions is a grand challenge, but we will take a more modest approach.

We believe that a key intuition for understanding neural computation is the concept of feature selectivity: while the space of inputs to a neuron—

whether we think of inputs as arriving at the synapses or being driven by sensory signals outside the brain—is vast, individual neurons are sensitive only to some restricted set of features in this vast space. The most general way to formalize this intuition is to say that we can compress (in the information-theoretic sense) our description of the inputs without losing any information about the neural output (Tishby, Pereira, & Bialek, 1999). We might hope that this selective compression of the input data has a simple geometric description, so that the relevant bits about the input correspond to coordinates along some restricted set of relevant dimensions in the space of inputs. If this is the case, feature selectivity should be formalized as a reduction of dimensionality (de Ruyter van Steveninck & Bialek, 1988), and this is the approach we follow here. Closely related work on the use of dimensionality reduction to analyze neural feature selectivity has been described in recent work (Bialek & de Ruyter van Steveninck, 2003; Sharpee, Rust, & Bialek, in press).

Here, we develop the idea of dimensionality reduction as a tool for analysis of neural computation and apply these tools to the HH model. While our initial goal was to test new analysis methods in the context of a presumably simple and well-understood model, we have found that the HH neuron performs a computation of surprising richness. Preliminary accounts of these results have already appeared (Agüera y Arcas, 1998; Agüera y Arcas, Bialek, & Fairhall, 2001).

2 Dimensionality Reduction

Neurons take input signals at their synapses and give as output sequences of spikes. To characterize a neuron completely is to identify the mapping between neuronal input and the spike train the neuron produces in response. In the absence of any simplifying assumptions, this requires probing the system with every possible input. Most often, these inputs are spikes from other neurons; each neuron typically has of order $N \sim 10^3$ presynaptic connections. If the system operates at 1 msec resolution and the time window of relevant inputs is 40 msec, then we can think of a single neuron as having an input described by a $\sim 4 \times 10^4$ bit word—the presence or absence of a spike in each 1 msec bin for each presynaptic cell—which is then mapped to a one (spike) or zero (no spike). More realistically, if average spike rates are $\sim 10\,\mathrm{s}^{-1}$, the input words can be compressed by a factor of 10. In this picture, a neuron computes a Boolean function over roughly 4000 variables. Clearly one cannot sample every one of the $\sim 2^{4000}$ inputs to identify the neural computation. Progress requires making some simplifying assumption about the function computed by the neuron so that we can vastly reduce the space of possibilities over which to search. We use the idea of dimensionality reduction in this spirit, as a simplifying assumption that allows us to make progress but that also must be tested directly.

The ideas of feature selectivity and dimensionality reduction have a long history in neurobiology. The idea of receptive fields as formulated by Hartline, Kuffler, and Barlow for the visual system gave a picture of neurons as having a template against which images would be correlated (Hartline, 1940; Kuffler, 1953; Barlow, 1953). If we think of images as vectors in a high-dimensional space, with coordinates determined by the intensities of each pixel, then the simplest receptive field models describe the neuron as sensitive to only one direction or projection in this high-dimensional space. This picture of projection followed by thresholding or some other nonlinearity to determine the probability of spike generation was formalized in the linear perceptron (Rosenblatt, 1958, 1962). In subsequent work, Barlow, Hill, and Levick (1964) characterized neurons in which the receptive field has subregions in space and time such that summation is at least approximately linear in each subregion but these summed signals interact nonlinearly, for example, to generate direction selectivity and motion sensitivity. We can think of Hubel and Wiesel's description of complex and hypercomplex cells (Hubel & Wiesel, 1962) again as a picture of approximately linear summation within subregions followed by nonlinear operations on these multiple summed signals. More formally, the proper combination of linear summation and nonlinear or logical operations may provide a useful bridge from receptive field properties to proper geometric primitives in visual computation (Iverson & Zucker, 1995). In the same way that a single receptive field or perceptron model has one relevant dimension in the space of visual stimuli, these more complex cells have as many relevant dimensions as there are independent subregions of the receptive field. Although this number is larger than one, it still is much smaller than the full dimensionality of the possible spatiotemporal variations in visual inputs.

The idea that neurons in the auditory system might be described by a filter followed by a nonlinear transformation to determine the probability of spike generation was the inspiration for de Boer's development (de Boer & Kuyper, 1968) of triggered or reverse correlation. Modern uses of reverse correlation to characterize the filtering or receptive field properties of a neuron often emphasize that this approach provides a "linear approximation" to the input-output properties of the cell, but the original idea was almost the opposite: neurons clearly are nonlinear devices, but this is separate from the question of whether the probability of generating a spike is determined by a simple projection of the sensory input onto a single filter or template. In fact, as explained by Rieke, Warland, Bialek, and de Ruyter van Steveninck (1997), linearity is seldom a good approximation for the neural input-output relation, but if there is one relevant dimension, then (provided that input signals are chosen with suitable statistics) the reverse correlation method is guaranteed to find this one special direction in the space of inputs to which the neuron is sensitive. While the reverse correlation method is guaranteed to find the one relevant dimension if it exists, the method does not include

any way of testing for other relevant dimensions, or more generally for measuring the dimensionality of the relevant subspace.

The idea of characterizing neural responses directly as the reduction of dimensionality emerged from studies (de Ruyter van Steveninck & Bialek, 1988) of a motion-sensitive neuron in the fly visual system. In particular, this work suggested that it is possible to estimate the dimensionality of the relevant subspace rather than just assuming that it is small (or equal to one). More recent work on the fly visual system has exploited the idea of dimensionality reduction to probe both the structure and adaptation of the neural code (Brenner, Bialek, & de Ruyter van Steveninck, 2000; Fairhall, Lewen, Bialek, & de Ruyter van Steveninck, 2001) and the nature of the computation that extracts the motion signal from the spatiotemporal array of photoreceptor inputs (Bialek & de Ruyter van Steveninck, 2003). Here we review the ideas of dimensionality reduction from previous work; extensions of these ideas begin in section 3.

In the spirit of neural network models, we will simplify away the spatial structure of neurons and consider time-dependent currents $I(t)$ injected into a point–like neuron. While this misses much of the complexity of real cells, we will find that even this system is highly nontrivial. If the input is an injected current, then the neuron maps the history of this current, $I(t < t_0)$, into the presence or absence of a spike at time t_0. More generally, we might imagine that the cell (or our description) is noisy, so that there is a probability of spiking $P[\text{spike at } t_0 \mid I(t < t_0)]$ that depends on the current history. The dependence on the history of the current means that the input signal still is high dimensional, even without spatial dependence. Working at time resolution Δt and assuming that currents in a window of size T are relevant to the decision to spike, the input space is of dimension $D = T/\Delta t$, where D is often of order 100.

The idea of dimensionality reduction is that the probability of spike generation is sensitive only to some limited number of dimensions K within the D-dimensional space of inputs. We begin our analysis by searching for linear subspaces, that is, a set of signals $s_1, s_2, \ldots, s_K$ that can be constructed by filtering the current,

$$s_\mu = \int_0^\infty dt f_\mu(t) I(t_0 - t), \tag{2.1}$$

so that the probability of spiking depends on only this small set of signals,

$$P[\text{spike at } t_0 \mid I(t < t_0)] = P[\text{spike at } t_0] g(s_1, s_2, \ldots, s_K), \tag{2.2}$$

where the inclusion of the average probability of spiking, $P[\text{spike at } t_0]$, leaves g dimensionless. If we think of the current $I(t_0 - T < t < t_0)$ as a D-dimensional vector, with one dimension for each discrete sample at spacing Δt, then the filtered signals s_i are linear projections of this vector. In

this formulation, characterizing the computation done by a neuron involves three steps:

1. Estimate the number of relevant stimulus dimensions K, with the hope that there will be many fewer than the original dimensionality D.
2. Identify a set of filters that project into this relevant subspace.
3. Characterize the nonlinear function $g(\vec{s})$.

The classical perceptron–like cell of neural network theory would have only one relevant dimension, given by the vector of weights, and a simple form for g, typically a sigmoid.

Rather than trying to look directly at the distribution of spikes given stimuli, we follow de Ruyter van Steveninck and Bialek (1988) and consider the distribution of signals conditional on the response, $P[I(t < t_0) \mid \text{spike at } t_0]$, also called the response conditional ensemble (RCE); these are related by Bayes' rule,

$$\frac{P[\text{spike at } t_0 \mid I(t < t_0)]}{P[\text{spike at } t_0]} = \frac{P[I(t < t_0) \mid \text{spike at } t_0]}{P[I(t < t_0)]}. \tag{2.3}$$

We can now compute various moments of the RCE. The first moment is the spike-triggered average stimulus (STA),

$$\text{STA}(\tau) = \int [dI] P[I(t < t_0) \mid \text{spike at } t_0] I(t_0 - \tau), \tag{2.4}$$

which is the object that one computes in reverse correlation (de Boer & Kuyper, 1968; Rieke et al., 1997). If we choose the distribution of input stimuli $P[I(t < t_0)]$ to be gaussian white noise, then for a perceptron–like neuron sensitive to only one direction in stimulus space, it can be shown that the STA or first moment of the RCE is proportional to the vector or filter $f(\tau)$ that defines this direction (Rieke et al., 1997).

Although it is a theorem that the STA is proportional to the relevant filter $f(\tau)$, in principle it is possible that the proportionality constant is zero, most plausibly if the neuron's response has some symmetry, such as phase invariance in the response of high-frequency auditory neurons. It also is worth noting that what is really important in this analysis is the gaussian distribution of the stimuli, not the "whiteness" of the spectrum. For nonwhite but gaussian inputs, the STA measures the relevant filter blurred by the correlation function of the inputs, and hence the true filter can be recovered (at least in principle) by deconvolution. For nongaussian signals and nonlinear neurons, there is no corresponding guarantee that the selectivity of the neuron can be separated from correlations in the stimulus (Sharpee et al., in press).

To obtain more than one relevant direction (or to reveal relevant directions when symmetries cause the STA to vanish), we proceed to second

order and compute the covariance matrix of fluctuations around the spike-triggered average,

$$C_{\text{spike}}(\tau, \tau') = \int [dI] P[I(t < t_0) \mid \text{spike at } t_0] I(t_0 - \tau) I(t_0 - \tau') - \text{STA}(\tau)\text{STA}(\tau'). \tag{2.5}$$

In the same way that we compare the spike-triggered average to some constant average level of the signal in the whole experiment, we compare the covariance matrix C_{spike} with the covariance of the signal averaged over the whole experiment,

$$C_{\text{prior}}(\tau, \tau') = \int [dI] P[I(t < t_0)] I(t_0 - \tau) I(t_0 - \tau'), \tag{2.6}$$

to construct the change in the covariance matrix,

$$\Delta C = C_{\text{spike}} - C_{\text{prior}}. \tag{2.7}$$

With time resolution Δt in a window of duration T as above, all of these covariances are $D \times D$ matrices. In the same way that the spike-triggered average has the clearest interpretation when we choose inputs from a gaussian distribution, ΔC also has the clearest interpretation in this case. Specifically, if inputs are drawn from a gaussian distribution, then it can be shown that (Bialek & de Ruyter van Steveninck, 2003):

1. If the neuron is sensitive to a limited set of K-input dimensions as in equation 2.2, then ΔC will have only K nonzero eigenvalues.[1] In this way, we can measure directly the dimensionality K of the relevant subspace.
2. If the distribution of inputs is both gaussian and white, then the eigenvectors associated with the nonzero eigenvalues span the same space as that spanned by the filters $\{f_\mu(\tau)\}$.
3. For nonwhite (correlated) but still gaussian inputs, the eigenvectors span the space of the filters $\{f_\mu(\tau)\}$ blurred by convolution with the correlation function of the inputs.

Thus, the analysis of ΔC for neurons responding to gaussian inputs should allow us to identify the subspace of inputs of relevance and test specifically the hypothesis that this subspace is of low dimension.

[1] As with the STA, it is in principle possible that symmetries or accidental features of the function $g(\vec{s})$ would cause some of the K eigenvalues to vanish, but this is very unlikely.

Several points are worth noting. First, except in special cases, the eigenvectors of ΔC and the filters $\{f_\mu(\tau)\}$ are not the principal components of the RCE, and hence this analysis of ΔC is not a principal component analysis. Second, the nonzero eigenvalues of ΔC can be either positive or negative, depending on whether the variance of inputs along that particular direction is larger or smaller in the neighborhood of a spike. Third, although the eigenvectors span the relevant subspace, these eigenvectors do not form a preferred coordinate system within this subspace. Finally, we emphasize that dimensionality reduction—identification of the relevant subspace—is only the first step in our analysis of the computation done by a neuron.

3 Measuring the Success of Dimensionality Reduction

The claim that certain stimulus features are most relevant is in effect a model for the neuron, so the next question is how to measure the effectiveness or accuracy of this model. Several different ideas have been suggested in the literature as ways of testing models based on linear receptive fields in the visual system (Stanley, Lei, & Dan, 1999; Keat, Reinagel, Reid, & Meister, 2001) or linear spectrotemporal receptive fields in the auditory system (Theunissen, Sen, & Doupe, 2000). These methods have in common that they introduce a metric to measure performance—for example, mean square error in predicting the firing rate as averaged over some window of time. Ideally, we would like to have a performance measure that avoids any arbitrariness in the choice of metric, and such metric-free measures are provided uniquely by information theory (Shannon, 1948; Cover & Thomas, 1991).

Observing the arrival time t_0 of a single spike provides a certain amount of information about the input signals. Since information is mutual, we can also say that knowing the input signal trajectory $I(t < t_0)$ provides information about the arrival time of the spike. If "details are irrelevant," then we should be able to discard these details from our description of the stimulus and yet preserve the mutual information between the stimulus and spike arrival times (for an abstract discussion of such selective compression, see Tishby et al., 1999). In constructing our low-dimensional model, we represent the complete (D-dimensional) stimulus $I(t < t_0)$ by a smaller number ($K < D$) of dimensions $\vec{s} = (s_1, s_2, \ldots, s_K)$.

The mutual information $\mathcal{I}[I(t < t_0); t_0]$ is a property of the neuron itself, while the mutual information $\mathcal{I}[\vec{s}; t_0]$ characterizes how much our reduced description of the stimulus can tell us about when spikes will occur. Necessarily, our reduction of dimensionality causes a loss of information, so that

$$\mathcal{I}[\vec{s}; t_0] \leq \mathcal{I}[I(t < t_0); t_0], \tag{3.1}$$

but if our reduced description really captures the computation done by the neuron, then the two information measures will be very close. In particular,

if the neuron were described exactly by a lower-dimensional model—as for a linear perceptron or for an integrate-and-fire neuron (Agüera y Arcas & Fairhall, 2003)—then the two information measures would be equal. More generally, the ratio $\mathcal{I}[\vec{s}; t_0]/\mathcal{I}[I(t < t_0); t_0]$ quantifies the efficiency of the low-dimensional model, measuring the fraction of information about spike arrival times that our K dimensions capture from the full signal $I(t < t_0)$.

As shown by Brenner, Strong, Koberle, Bialek, and de Ruyter van Steveninck (2000), the arrival time of a single spike provides an information,

$$\mathcal{I}[I(t < t_0); t_0] \equiv I_{\text{one spike}} = \frac{1}{T}\int_0^T dt \frac{r(t)}{\bar{r}} \log_2\left[\frac{r(t)}{\bar{r}}\right], \tag{3.2}$$

where $r(t)$ is the time-dependent spike rate, $\bar{r}$ is the average spike rate, and $\langle\cdots\rangle$ denotes an average over time. In principle, information should be calculated as an average over the distribution of stimuli, but the ergodicity of the stimulus justifies replacing this ensemble average with a time average. For a deterministic system like the HH equations, the spike rate is a singular function of time: given the inputs $I(t)$, spikes occur at definite times with no randomness or irreproducibility. If we observe these responses with a time resolution Δt, then for Δt sufficiently small, the rate $r(t)$ at any time t either is zero or corresponds to a single spike occurring in one bin of size Δt, that is, $r = 1/\Delta t$. Thus, the information carried by a single spike is

$$I_{\text{one spike}} = -\log_2 \bar{r}\Delta t. \tag{3.3}$$

On the other hand, if the probability of spiking really depends on only the stimulus dimensions $s_1, s_2, \ldots, s_K$, we can substitute

$$\frac{r(t)}{\bar{r}} \to \frac{P(\vec{s} \mid \text{spike at } t)}{P(\vec{s})}. \tag{3.4}$$

Replacing the time averages in equation 3.2 with ensemble averages, we find

$$\mathcal{I}[\vec{s}; t_0] \equiv I^{\vec{s}}_{\text{one spike}} = \int d^K s P(\vec{s} \mid \text{spike at } t) \log_2\left[\frac{P(\vec{s} \mid \text{spike at } t)}{P(\vec{s})}\right] \tag{3.5}$$

(for details of these arguments, see Brenner, Strong, et al., 2000). This allows us to compare the information captured by the K-dimensional reduced model with the true information carried by single spikes in the spike train.

For reasons that we will discuss in the following section, and as was pointed out in Agüera y Arcas et al. (2001) and Agüera y Arcas and Fairhall (2003), we will be considering isolated spikes—those separated from previous spikes by a period of silence. This has important consequences for our analysis. Most significantly, as we will be considering spikes that occur

on a background of silence, the relevant stimulus ensemble, conditioned on the silence, is no longer gaussian. Further, we will need to refine our information estimate.

The derivation of equation 3.2 makes clear that a similar formula must determine the information carried by the occurrence time of any event, not just single spikes; we can define an event rate in place of the spike rate and then calculate the information carried by these events (Brenner, Strong, et al., 2000). In the case here, we wish to compute the information obtained by observing an isolated spike, or equivalently by the event silence+spike. This is straightforward: we replace the spike rate by the rate of isolated spikes, and equation 3.2 will give us the information carried by the arrival time of a single isolated spike. The problem is that this information includes both the information carried by the occurrence of the spike and the information conveyed in the condition that there were no spikes in the preceding $t_{\rm silence}$ msec (for an early discussion of the information carried by silence, see de Ruyter van Steveninck & Bialek, 1988). We would like to separate these contributions, since our idea of dimensionality reduction applies only to the triggering of a spike, not to the temporally extended condition of nonspiking.

To separate the information carried by the isolated spike itself, we have to ask how much information we gain by seeing an isolated spike given that the condition for isolation has already been met. As discussed by Brenner, Strong, et al. (2000), we can compute this information by thinking about the distribution of times at which the isolated spike can occur. Given that we know the input stimulus, the distribution of times at which a single isolated spike will be observed is proportional to $r_{\rm iso}(t)$, the time-dependent rate or peristimulus time histogram for isolated spikes. With proper normalization, we have

$$P_{\rm iso}(t \mid \text{inputs}) = \frac{1}{T} \cdot \frac{1}{\bar{r}_{\rm iso}} r_{\rm iso}(t), \tag{3.6}$$

where T is duration of the (long) window in which we can look for the spike and $\bar{r}_{\rm iso}$ is the average rate of isolated spikes. This distribution has an entropy,

$$S_{\rm iso}(t \mid \text{inputs}) = -\int_0^T dt\, P_{\rm iso}(t \mid \text{inputs}) \log_2 P_{\rm iso}(t \mid \text{inputs}) \tag{3.7}$$

$$= -\frac{1}{T}\int_0^T dt \frac{r_{\rm iso}(t)}{\bar{r}_{\rm iso}} \log_2 \left[\frac{1}{T} \cdot \frac{r_{\rm iso}(t)}{\bar{r}_{\rm iso}}\right] \tag{3.8}$$

$$= \log_2(T\bar{r}_{\rm iso}\Delta t) \text{ bits}, \tag{3.9}$$

where again we use the fact that for a deterministic system, the time-dependent rate must be either zero or the maximum allowed by our time

resolution Δt. To compute the information carried by a single spike, we need to compare this entropy with the total entropy possible when we do not know the inputs.

It is tempting to think that without knowledge of the inputs, an isolated spike is equally likely to occur anywhere in the window of size T, which leads us back to equation 3.3 with $\bar{r}$ replaced by $\bar{r}_{\text{iso}}$. In this case, however, we are assuming that the condition for isolation has already been met. Thus, even without observing the inputs, we know that isolated spikes can occur only in windows of time whose total length is $T_{\text{silence}} = T \cdot P_{\text{silence}}$, where P_{silence} is the probability that any moment in time is at least t_{silence} after the most recent spike. Thus, the total entropy of isolated spike arrival times (given that the condition for silence has been met) is reduced from $\log_2 T$ to

$$S_{\text{iso}}(t \mid \text{silence}) = \log_2(T \cdot P_{\text{silence}}), \tag{3.10}$$

and the information that the spike carries beyond what we know from the silence itself is

$$\Delta I_{\text{iso spike}} = S_{\text{iso}}(t \mid \text{silence}) - S_{\text{iso}}(t \mid \text{inputs}) \tag{3.11}$$

$$= \frac{1}{T}\int_0^T dt \frac{r_{\text{iso}}(t)}{\bar{r}_{\text{iso}}} \log_2 \left[\frac{r_{\text{iso}}(t)}{\bar{r}_{\text{iso}}} \cdot P_{\text{silence}}\right] \tag{3.12}$$

$$= -\log_2(\bar{r}_{\text{iso}}\Delta t) + \log_2 P_{\text{silence}} \text{ bits}. \tag{3.13}$$

This information, which is defined independent of any model for the feature selectivity of the neuron, provides the benchmark against which our reduction of dimensionality will be measured. To make the comparison, however, we need the analog of equation 3.5.

Equation 3.12 provides us with an expression for the information conveyed by isolated spikes in terms of the probability that these spikes occur at particular times; this is analogous to equation 3.2 for single (nonisolated) spikes. If we follow a path analogous to that which leads from equation 3.2 to equation 3.5, we find an expression for the information that an isolated spike provides about the K stimulus dimensions $\vec{s}$:

$$\Delta I^{\vec{s}}_{\text{iso spike}} = \int d\vec{s}\, P(\vec{s} \mid \text{iso spike at } t) \log_2 \left[\frac{P(\vec{s} \mid \text{iso spike at } t)}{P(\vec{s} \mid \text{silence})}\right] + \langle \log_2 P(\text{silence} \mid \vec{s})\rangle, \tag{3.14}$$

where the prior is now also conditioned on silence: $P(\vec{s} \mid \text{silence})$ is the distribution of $\vec{s}$ given that $\vec{s}$ is preceded by a silence of at least t_{silence}. Notice that this silence-conditioned distribution is not knowable a priori, and in particular it is not gaussian; $P(\vec{s} \mid \text{silence})$ must be sampled from data.

The last term in equation 3.14 is the entropy of a binary variable that indicates whether particular moments in time are silent given knowledge

of the stimulus. Again, since the HH model is deterministic, this conditional entropy should be zero if we keep a complete description of the stimulus. In fact, we are not interested in describing those features of the stimulus that lead to silence, and it is not fair (as we will see) to judge the success of dimensionality reduction by looking at the prediction of silence, which necessarily involves multiple dimensions. To make a meaningful comparison, then, we will assume that there is a perfect description of the stimulus conditions leading to silence and focus on the stimulus features that trigger the isolated spike. When we approximate these features by the K-dimensional space $\vec{s}$, we capture an amount of information,

$$\Delta I^{\vec{s}}_{\text{iso spike}} = \int d\vec{s}\, P(\vec{s} \mid \text{iso spike at } t) \log_2 \left[\frac{P(\vec{s} \mid \text{iso spike at } t)}{P(\vec{s} \mid \text{silence})} \right]. \quad (3.15)$$

This is the information that we can compare with $\Delta I_{\text{iso spike}}$ in equation 3.13 to determine the efficiency of our dimensionality reduction.

4 Characterizing the Hodgkin-Huxley Neuron

For completeness, we begin with a brief review of the dynamics of the space-clamped HH neuron (Hodgkin & Huxley, 1952). Hodgkin and Huxley modeled the dynamics of the current through a patch of membrane with ion-specific conductances:

$$C\frac{dV}{dt} = I(t) - \bar{g}_{\text{K}} n^4 (V - V_{\text{K}}) - \bar{g}_{\text{Na}} m^3 h (V - V_{\text{Na}}) - \bar{g}_l (V - V_l), \quad (4.1)$$

where $I(t)$ is injected current, K and Na subscripts denote potassium– and sodium–related variables, respectively, and l (for "leakage") terms include all other ion conductances with slower dynamics. C is the membrane capacitance. V_{K} and V_{Na} are ion-specific reversal potentials, and V_l is defined such that the total voltage V is exactly zero when the membrane is at rest. $\bar{g}_{\text{K}}$, $\bar{g}_{\text{Na}}$, and $\bar{g}_l$ are empirically determined maximal conductances for the different ion species, and the gating variables n, m, and h (on the interval $[0, 1]$) have their own voltage-dependent dynamics:

$$\begin{aligned} dn/dt &= (0.01V + 0.1)(1 - n)\exp(-0.1V) - 0.125n\exp(V/80), \\ dm/dt &= (0.1V + 2.5)(1 - m)\exp(-0.1V - 1.5) - 4m\exp(V/18), \\ dh/dt &= 0.07(1 - h)\exp(0.05V) - h\exp(-0.1V - 4). \end{aligned} \quad (4.2)$$

We have used the original values for these parameters, except for changing the signs of the voltages to correspond to the modern sign convention: $C = 1\ \mu\text{F/cm}^2$, $\bar{g}_{\text{K}} = 36\ \text{mS/cm}^2$, $\bar{g}_{\text{Na}} = 120\ \text{mS/cm}^2$, $\bar{g}_l = 0.3\ \text{mS/cm}^2$, $V_{\text{K}} = -12$ mV, $V_{\text{Na}} = +115$ mV, $V_l = +10.613$ mV. We have taken our system to be

a $\pi \times 30^2$ μm^2 patch of membrane. We solve these equations numerically using fourth-order Runge–Kutta integration.

The system is driven with a gaussian random noise current $I(t)$, generated by smoothing a gaussian random number stream with an exponential filter to generate a correlation time τ. It is convenient to choose τ to be longer than the time steps of numerical integration, since this guarantees that all functions are smooth on the scale of single time steps. Here, we will always use $\tau = 0.2$ msec, a value that is both less than the timescale over which we discretize the stimulus for analysis and far less than the neuron's capacitative smoothing timescale $RC \sim 3$ msec. $I(t)$ has a standard deviation σ, but since the correlation time is short, the relevant parameter usually is the spectral density $S = \sigma^2\tau$; we also add a DC offset I_0. In the following, we will consider two parameter regimes, $I_0 = 0$ and I_0 a finite value, which leads to more periodic firing.

The integration step size is fixed at 0.05 msec. The key numerical experiments were repeated at a step size of 0.01 msec with identical results. The time of a spike is defined as the moment of maximum voltage, for voltages exceeding a threshold (see Figure 1), estimated to subsample precision by quadratic interpolation. As spikes are both very stereotyped and very large compared to subspiking fluctuations, the precise value of this threshold is unimportant; we have used +20 mV.

4.1 Qualitative Description of Spiking. The first step in our analysis is to use reverse correlation, equation 2.4, to determine the average stimulus feature preceding a spike, the STA. In Figure 1(top), we display the STA in a regime where the spectral density of the input current is 6.5×10^{-4} nA^2 msec. The spike-triggered averages of the gating terms n^4 (proportion of open potassium channels) and m^3h (proportion of open sodium channels) and the membrane voltage V are plotted in Figure 1 (middle and bottom). The error bars mark the standard deviation of the trajectories of these variables.

As expected, the voltage and gating variables follow highly stereotyped trajectories during the ~5 msec surrounding a spike. First, the rapid opening of the sodium channels causes a sharp membrane depolarization (or rise in V); the slower potassium channels then open and repolarize the membrane, leaving it at a slightly lower potential than rest. The potassium channels close gradually, but meanwhile the membrane remains hyperpolarized and, due to its increased permeability to potassium ions, at lower resistance. These effects make it difficult to induce a second spike during this ~15 msec "refractory period." Away from spikes, the resting levels and fluctuations of the voltage and gating variables are quite small. The larger values evident in Figure 1(middle and bottom) by $\pm$15 msec are due to the summed contributions of nearby spikes.

The spike-triggered average current has a largely transient form, so that spikes are on average preceded by an upward swing in current. On the

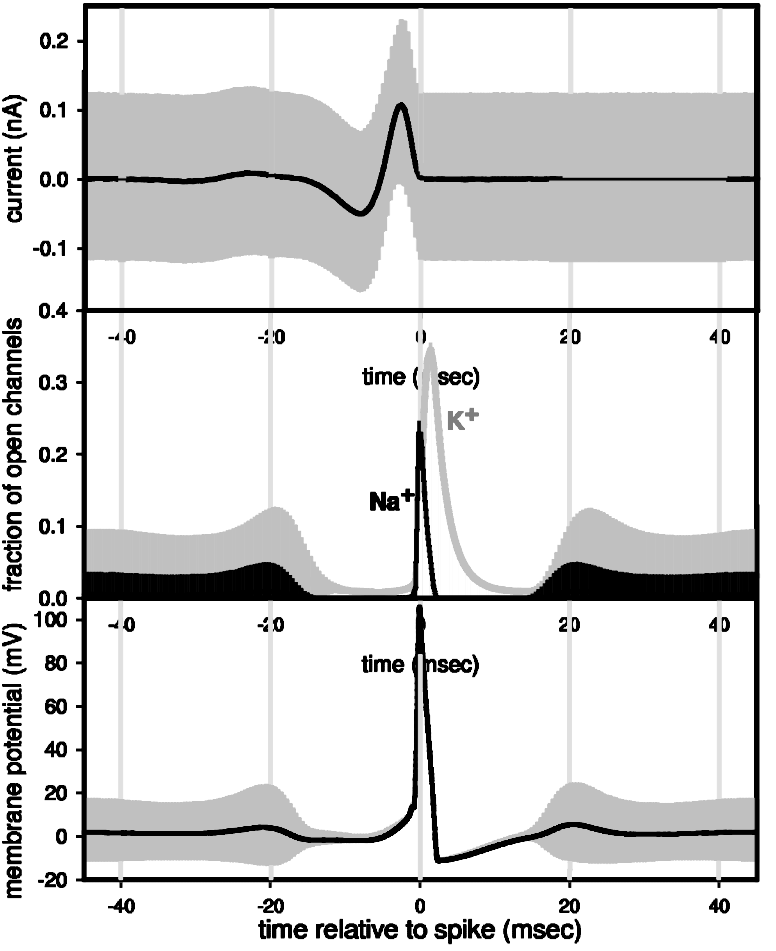

Figure 1: Spike-triggered averages with standard deviations for (top) the input current I, (middle) the fraction of open K^+ and Na^+ channels, and (bottom) the membrane voltage V, for the parameter regime $I_0 = 0$ and $S = 6.50 \times 10^{-4}$ nA^2 sec.

other hand, there is no obvious bottleneck in the current trajectories, so that the current variance is almost constant throughout the spike. This is qualitatively consistent with the idea of dimensionality reduction: if the neuron ignores most of the dimensions along which the current can vary, then the variance, which is shared almost equally among all dimensions for this near white noise, can change by only a small amount.

4.2 Interspike Interaction. Although the STA has the form of a differentiating kernel, suggesting that the neuron detects edge-like events in the current versus time, there must be a DC component to the cell's response. We recall that for constant inputs, the HH model undergoes a bifurcation to constant frequency spiking, where the frequency is a function of the value of the input above onset. Correspondingly, the STA does not sum precisely to zero; one might think of it as having a small integrating component that allows the system to spike under DC stimulation, albeit only above a threshold.

The system's tendency to periodic spiking under DC current input also is felt under dynamic stimulus conditions and can be thought of as a strong interaction between successive spikes. We illustrate this by considering a different parameter regime with a small DC current and some added noise ($I_0 = 0.11$ nA and $S = 0.8 \times 10^{-4}$ nA^2 sec). Note that the DC component puts the neuron in the metastable region of its $f - I$ curve (see Figure 2). In this regime, the neuron tends to fire quasi-regular trains of spikes intermittently, as shown in Figure 3. We will refer to these quasi-regular spike sequences as "bursts" (note that this term is often used to refer to compound spikes in neurons with additional channels; such events do not occur in the HH model).

Spikes can be classified into three types: those initiating a spike burst, those within a burst, and those ending a burst. The minimum length of

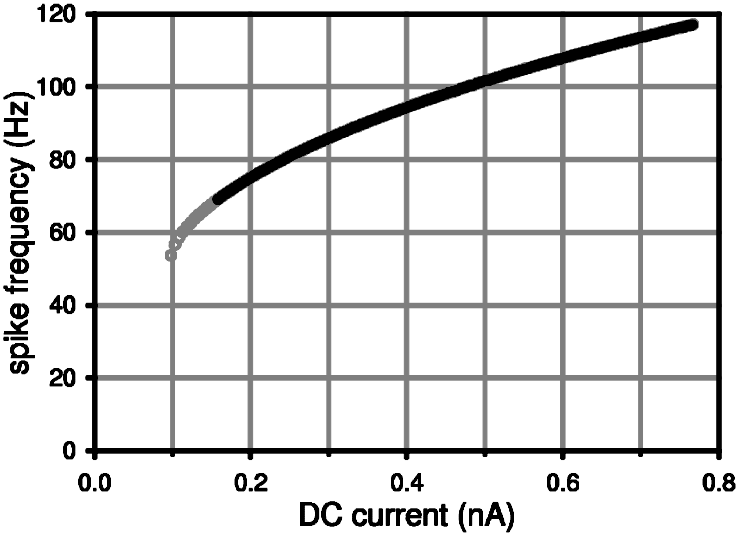

Figure 2: Firing rate of the HH neuron as a function of injected DC current. The empty circles at moderate currents denote the metastable region, where the neuron may be either spiking or silent.

Figure 3: Segment of a typical spike train in a "bursting" regime.

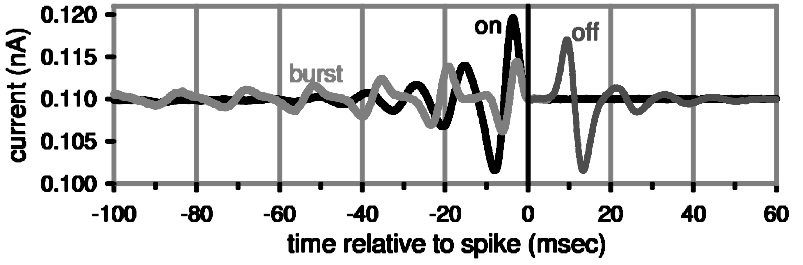

Figure 4: Spike-triggered averages, derived from spikes leading ("on"), inside ("burst"), and ending ("off") a burst. The parameters of this bursting regime are $I_0 = 0.11$ nA and $S = 0.8 \times 10^{-4}$ nA^2 sec. Note that the burst-ending spike average is, by construction, identical to that of any other within-burst spike for $t < 0$.

the silence between bursts is taken in this case to be 70 msec. Taking these three categories of spike as different "symbols" (de Ruyter van Steveninck & Bialek, 1988), we can determine the average stimulus for each. These are shown in Figure 4 with the spike at $t = 0$.

In this regime, the initial spike of a burst is preceded by a rapid oscillation in the current. Spikes within a burst are affected much less by the current; the feature immediately preceding such spikes is similar in shape to a single "wavelength" of the leading spike feature, but is of much smaller amplitude and is temporally compressed into the interspike interval. Hence, although it is clear that the timing of a spike within a burst is determined largely by the timing of the previous spike, the current plays some role in affecting the precise placement. This also demonstrates that the shape of the STA is not the same for all spikes; it depends strongly and nontrivially on the time to the previous spike, and this is related to the observation that subtly different patterns of two or three spikes correspond to very different average stimuli (de Ruyter van Steveninck & Bialek, 1988). For a reader of the spike code, a spike within a burst conveys a different message about the input than the spike at the onset of the burst. Finally, the feature ending a burst has a very similar form to the onset feature, but reversed in time. Thus, to a good approximation, the absence of a spike at the end of a burst can be read as the opposite of the onset of the burst.

In summary, this regime of the HH neuron is similar to a "flip-flop," or 1-bit memory. Like its electronic analog, the neuron's memory is preserved

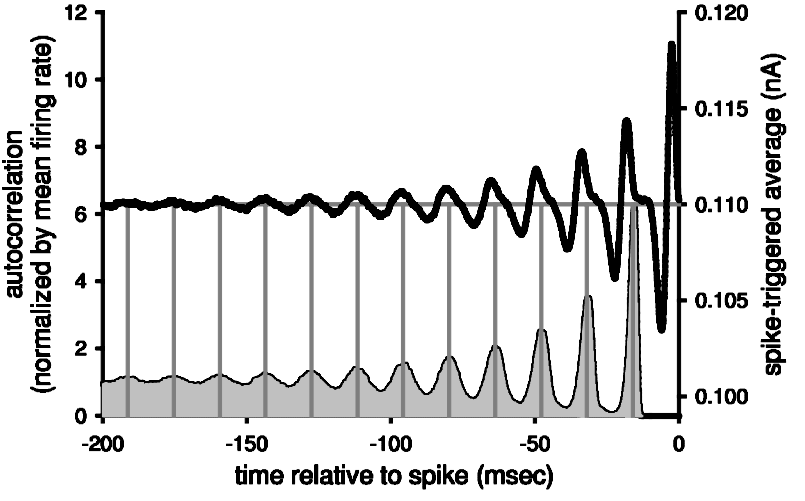

Figure 5: Overall spike-triggered average in the bursty regime, showing the ringing due to the tendency to periodic firing. Plotted in gray is the spike autocorrelation, showing the same oscillations.

by a feedback loop, here implemented by the interspike interaction. Large fluctuations in the input current at a certain frequency "flip" or "flop" the neuron between its silent and spiking states. However, while the neuron is spiking, further details of the input signal are transmitted by precise spike timing within a burst. If we calculate the spike-triggered average of all spikes for this regime, without regard to their position within a burst, then as shown in Figure 5, the relatively well-localized leading spike oscillation of Figure 4 is replaced by a long-lived oscillating function resulting from the spike periodicity. This is shown explicitly by comparing the overall STA with the spike autocorrelation, also shown in Figure 5. This same effect is seen in the STA of the burst spikes, which in fact dominates the overall average. Prediction of spike timing using such an STA would be computationally difficult due to its extension in time, but, more seriously, unsuccessful, as most of the function is an artifact of the spike history rather than the effect of the stimulus.

While the effects of spike interaction are interesting and should be included in a complete model for spike generation, we wish here to consider only the current's role in initiating spikes. Therefore, as we have argued elsewhere, we limit ourselves initially to the cases in which interspike interaction plays no role (Agüera y Arcas et al., 2001; Agüera & Fairhall, 2003). These "isolated" spikes can be defined as spikes preceded by a silent period t_{silence} long enough to ensure decoupling from the timing of the previous spike. A reasonable choice for t_{silence} can be inferred directly from the inter-

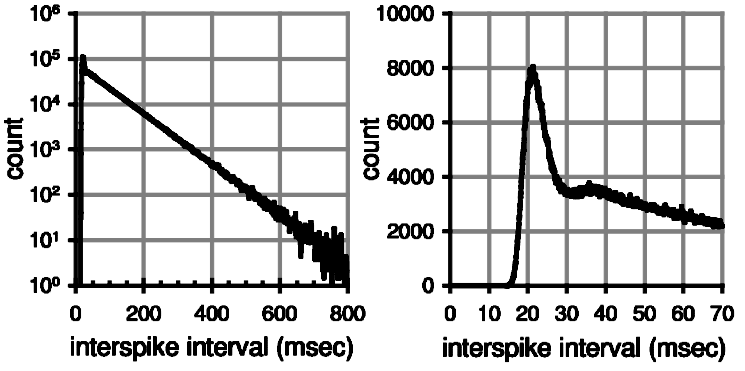

Figure 6: Hodgkin-Huxley interspike interval histogram for the parameters $I_0 = 0$ and $S = 6.5 \times 10^{-4}$ nA^2 sec, showing a peak at a preferred firing frequency and the long Poisson tail. The total number of spikes is $N = 5.18 \times 10^6$. The plot to the right is a closeup in linear scale.

spike interval distribution $P(\Delta t)$, illustrated in Figure 6. For the HH model, as in simpler models and many real neurons (Brenner, Agam, Bialek, & de Ruyter van Steveninck, 1998), the form of $P(\Delta t)$ has three noteworthy features: a refractory "hole" during which another spike is unlikely to occur, a strong mode at the preferred firing frequency, and an exponentially decaying, or Poisson, tail. The details of all three of these features are functions of the parameters of the stimulus (Tiesinga, José, & Sejnowski, 2000), and certain regimes may be dominated by only one or two features. The emergence of Poisson statistics in the tail of the distribution implies that these events are independent, so we can infer that the system has lost memory of the previous spike. We will therefore take isolated spikes to be those preceded by a silent interval $\Delta t \geq t_{\text{silence}}$, where t_{silence} is well into the Poisson regime. The burst-onset spikes of Figure 4 are isolated spikes by this definition.

Note that the bursty behavior evident in Figure 3 is characteristic of "type II" neurons, which begin firing at a well-defined frequency under DC stimulus; "type I" neurons, by contrast, can fire at arbitrarily low frequency under constant input. Nonetheless the analysis that follows is equally applicable to either type of neuron: spikes still interact at close range and become independent at sufficient separation. In what follows, we will be probing the HH neuron with current noise that remains below the DC threshold for periodic firing.

5 Isolated Spike Analysis

Focusing now on isolated spikes, we proceed to a second-order analysis of the current fluctuations around the isolated spike triggered average (see

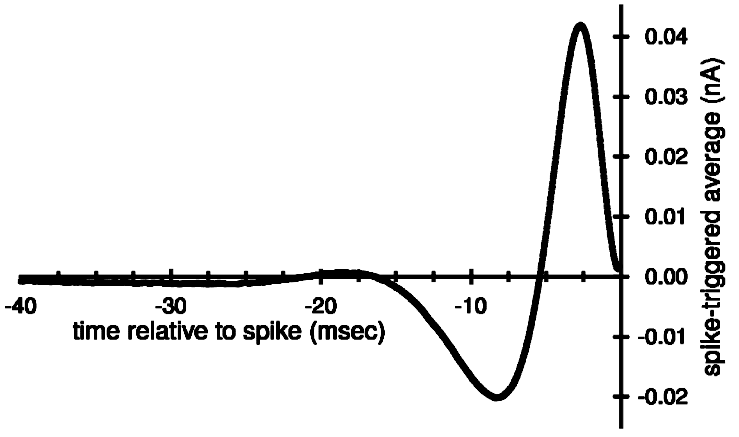

Figure 7: Spike-triggered average stimulus for isolated spikes.

Figure 7). We consider the response of the HH neuron to currents $I(t)$ with mean $I_0 = 0$ and spectral density of $S = 6.5 \times 10^{-4}$ nA^2 sec. Isolated spikes in this regime are defined by $t_{\text{silence}} = 60$ msec.

5.1 How Many Dimensions? As explained in section 2, our path to dimensionality reduction begins with the computation of covariance matrices for stimulus fluctuations surrounding a spike. The matrices are accumulated from stimulus segments 200 samples in length, roughly corresponding to sampling at the timescale sufficiently long to capture the relevant features. Thus, we begin in a 200-dimensional space. We emphasize that the theorem that connects eigenvalues of the matrix ΔC to the number of relevant dimensions is valid only for truly gaussian distributions of inputs and that by focusing on isolated spikes, we are essentially creating a nongaussian stimulus ensemble—namely, those stimuli that generate the silence out of which the isolated spike can appear. Thus, we expect that the covariance matrix approach will give us a heuristic guide to our search for lower-dimensional descriptions, but we should proceed with caution.

The "raw" isolated spike-triggered covariance $C_{\text{iso spike}}$ and the corresponding covariance difference ΔC, equation 2.7, are shown in Figure 8. The matrix shows the effect of the silence as an approximately translationally invariant band preceding the spike, the second-order analog of the constant negative bias in the isolated spike STA (see Figure 7). The spike itself is associated with features localized to ± 15 msec. In Figure 9, we show the spectrum of eigenvalues of ΔC computed using a sample of 80,000 spikes. Before calculating the spectrum, we multiply ΔC by C_{prior}^{-1}. This has the effect of giving us eigenvalues scaled in units of the input standard deviation

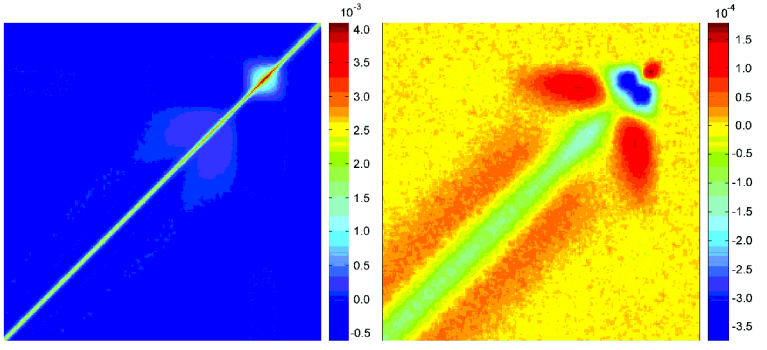

Figure 8: The isolated spike-triggered covariance $C_{\rm iso}$ (left) and covariance difference ΔC (right) for times $-30 < t < 5$ msec. The plots are in units of nA^2.

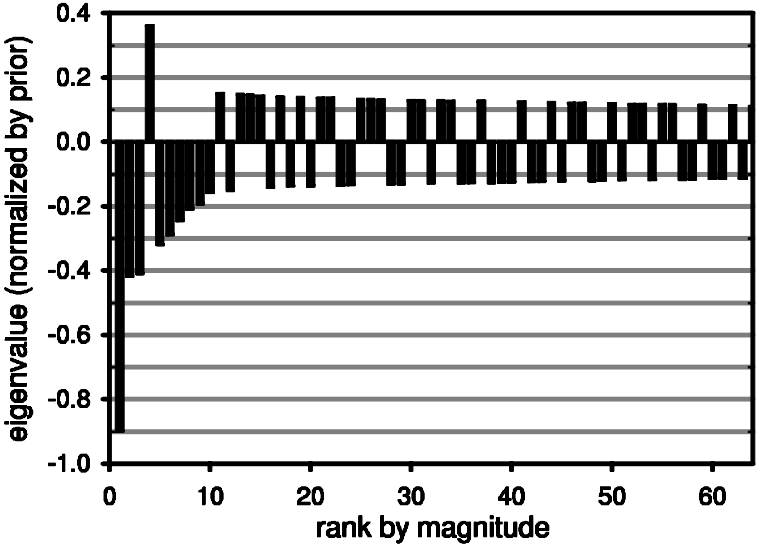

Figure 9: The leading 64 eigenvalues of the isolated spike-triggered covariance after accumulating 80,000 spikes.

along each dimension. Because the correlation time is short, $C_{\rm prior}$ is nearly diagonal.

While the eigenvalues decay rapidly, there is no obvious set of outstanding eigenvalues. To verify that this is not an effect of finite sampling, Figure 10 shows the spectrum of eigenvalue magnitudes as a function of sample size N. Eigenvalues that are truly zero up to the noise floor determined by

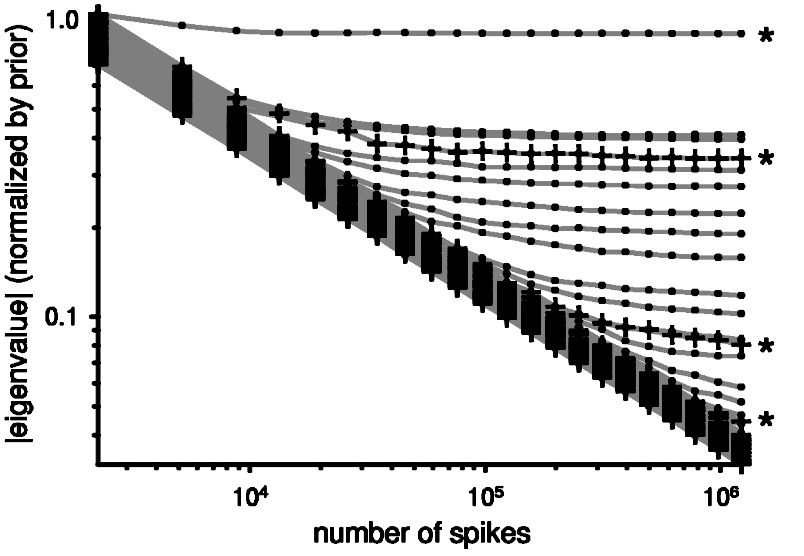

Figure 10: Convergence of the leading 64 eigenvalues of the isolated spike-triggered covariance with increasing sample size. The log slope of the diagonal is $1/\sqrt{n_{\text{spikes}}}$. Positive eigenvalues are indicated by crosses and negative by dots. The spike-associated modes are labeled with an asterisk.

sampling decrease like $\sqrt{N}$. We find that a sequence of eigenvalues emerges stably from the noise.

These results do not, however, imply that a low-dimensional approximation cannot be identified. The extended structure in the covariance matrix induced by the silence requirement is responsible for the apparent high dimensionality. In fact, as has been shown in Agüera y Arcas and Fairhall (2003), the covariance eigensystem includes modes that are local and spike associated, and others that are extended and silence associated, and thus irrelevant to a causal model of spike timing prediction. Fortunately, because extended silences and spikes are (by definition) statistically independent, there is no mixing between the two types of modes. To identify the spike-associated modes, we follow the diagnostic of Agüera y Arcas and Fairhall (2003), computing the fraction of the energy of each mode concentrated in the period of silence, which we take to be $-60 \leq t \leq -40$ msec. The energy of a spike-associated mode in the silent period is due entirely to noise and will therefore decrease like $1/n_{\text{spikes}}$ with increasing sample size, while this energy remains of order unity for silence modes. Carrying out the test on the covariance modes, we obtain Figure 11, which shows that the first and fourth modes rapidly emerge as spike associated. Two further spike-associated modes appear over the sample shown, with the suggestion

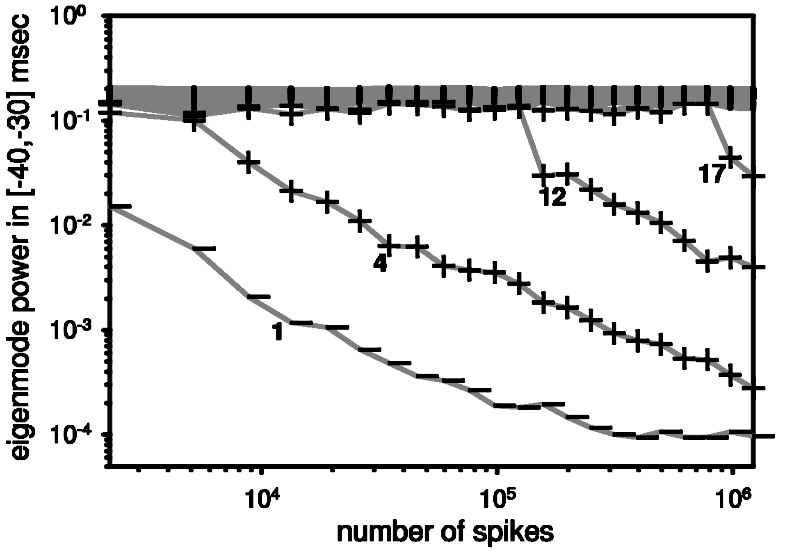

Figure 11: For the leading 64 modes, fraction of the mode energy over the interval $-40 < t < -30$ msec as a function of increasing sample size. Modes emerging with low energy are spike associated. The symbols indicate the sign of the eigenvalue.

of other, weaker modes yet to emerge. The two leading silence modes are shown in Figure 12. Those shown are typical; most modes resemble Fourier modes, as the silence condition is close to time translationally invariant.

Examining the eigenvectors corresponding to the two leading spike-associated eigenvalues, which for convenience we will denote s_1 and s_2 (although they are not the leading modes of the matrix), we find (see Figure 13) that the first mode closely resembles the isolated spike STA and the second is close to the derivative of the first. Both modes approximate differentiating operators; there is no linear combination of these modes that would produce an integrator.

If the neuron filtered its input and generated a spike when the output of the filter crosses threshold, we would find two significant dimensions associated with a spike. The first dimension would correspond simply to the filter, as the variance in this dimension is reduced to zero (for a noiseless system) at the occurrence of a spike. As the threshold is always crossed from below, the stimulus projection onto the filter's derivative must be positive, again resulting in a reduced variance. It is tempting to suggest, then, that filtered threshold crossing is a good approximation to the HH model, but we will see that this is not correct.

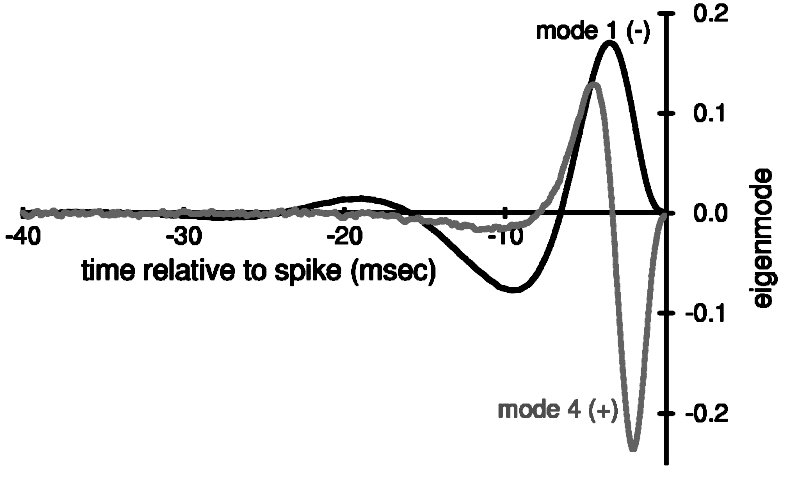

Figure 12: Modes 2 and 3 of the spike-triggered covariance (silence associated).

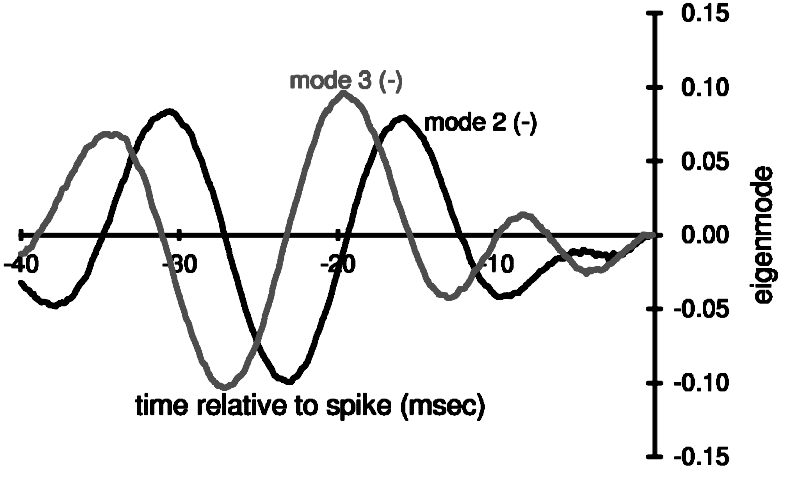

Figure 13: Modes 1 and 4 of the spike-triggered covariance, which are the leading spike-associated modes.

5.2 Evaluating the Nonlinearity. At each instant of time, we can find the projections of the stimulus along the leading spike-associated dimensions s_1 and s_2. By construction, the distribution of these signals over the whole experiment, $P(s_1, s_2)$, is gaussian. The appropriate prior for the isolation condition, $P(s_1, s_2 \mid \text{silence})$, differs only subtly from the gaussian prior. On the other hand, for each spike, we obtain a sample from the dis-

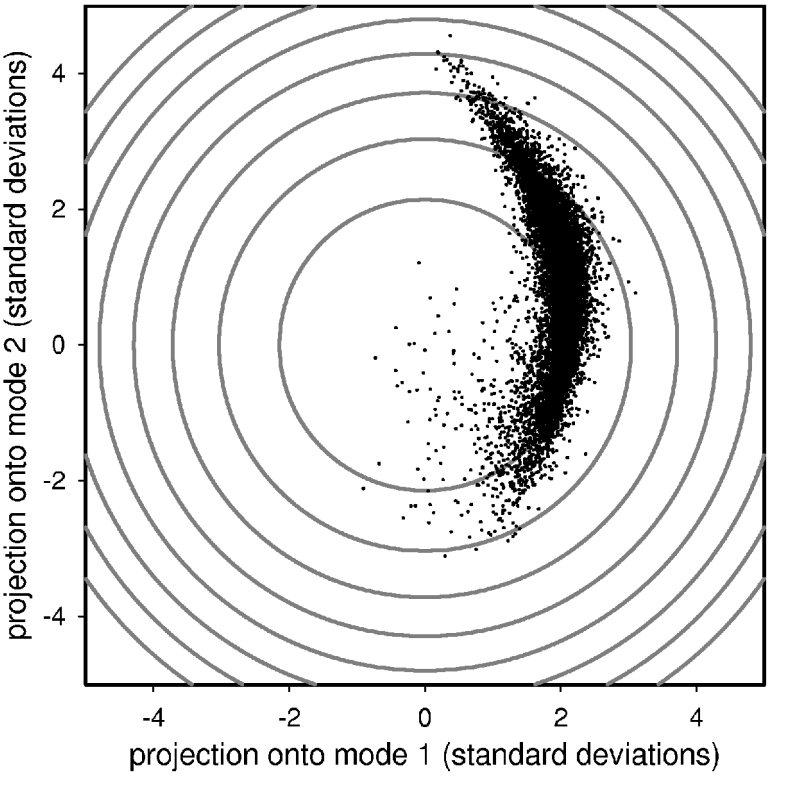

Figure 14: 10^4 spike-conditional stimuli (or "spike histories") projected along the first two covariance modes. The axes are in units of standard deviation on the prior gaussian distribution. The circles, from the inside out, enclose all but $10^{-1}, 10^{-2}, \ldots, 10^{-8}$ of the prior.

tribution $P(s_1, s_2 \mid \text{iso spike at } t_0)$, leading to the picture in Figure 14. The prior and spike-conditional distributions are clearly better separated in two dimensions than in one, which means that the two-dimensional description captures more information than projection onto the spike-triggered average alone. Surprisingly, the spike-conditional distribution is curved, unlike what we would expect for a simple thresholding device. Furthermore, the eigenvalue of ΔC, which we associate with the direction of threshold crossing (plotted on the y-axis in Figure 14), is positive, indicating increased rather than decreased variance in this direction. As we see, projections onto this mode are almost equally likely to be positive or negative, ruling out the threshold crossing interpretation.

Combining equations 2.2 and 2.3 for isolated spikes, we have

$$g(s_1, s_2) = \frac{P(s_1, s_2 \mid \text{iso spike at } t_0)}{P(s_1, s_2 \mid \text{silence})}, \tag{5.1}$$

so that these two distributions determine the input-output relation of the neuron in this 2D space (Brenner, Bialek, et al., 2000). Recall that although the subspace is linear, g can have arbitrary nonlinearity. Figure 14 shows that this input-output relation has clear structure, but also some fuzziness. As the HH model is deterministic, the input-output relation should be a singular function in the continuous space of inputs—spikes occur only when certain exact conditions are met. Of course, finite time resolution introduces some blurring, and so we need to understand whether the blurring of the input-output relation in Figure 14 is an effect of finite time resolution or a real limitation of the 2D description.

5.3 Information Captured in Two Dimensions. We will measure the effectiveness of our description by computing the information in the 2D approximation, according to the methods described in section 3. If the two-dimensional approximation were exact, we would find that $I^{s_1,s_2}_{\text{isospike}} = I_{\text{isospike}}$; more generally, one finds $I^{s_1,s_2}_{\text{isospike}} \leq I_{\text{isospike}}$, and the fraction of the information captured measures the quality of the approximation. This fraction is plotted in Figure 15 as a function of time resolution. For comparison, we also show the information captured in the one-dimensional case, considering only the stimulus projection along the STA.

We find that our low-dimensional model captures a substantial fraction of the total information available in spike timing in an HH neuron over a range of time resolutions. The approximation is best near $\Delta t = 3$ msec,

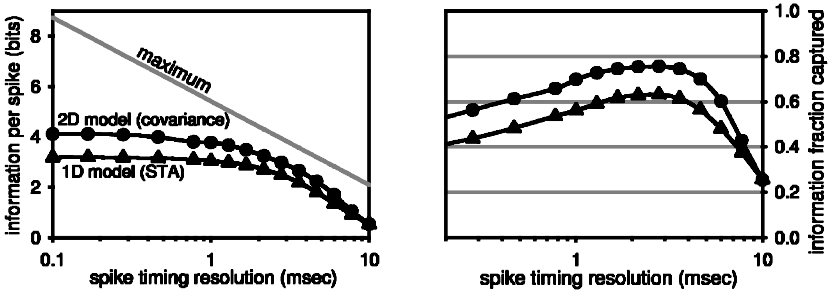

Figure 15: Bits per spike (left) and fraction of the theoretical limit (right) of timing information in a single spike at a given temporal resolution captured by projection onto the STA alone (triangles) and projection onto ΔC covariance modes 1 and 2 (circles).

reaching 75%. Thus, the complex nonlinear dynamics of the HH model can be approximated by saying that the neuron is sensitive to a 2D linear subspace in the high-dimensional space of input signals, and this approximate description captures up to 75% of the mutual information between input currents and (isolated) spike arrival times.

The dependence of information on time resolution (see Figure 15) shows that the absolute information captured saturates for both the 1D and 2D cases, at ≈ 3.2 and 4.1 bits respectively. Hence, for smaller Δt, the information fraction captured drops. The model provides, at its best, a time resolution of 3 msec, so that information carried by more precise spike timing is lost in our low-dimensional projection. Might this missing information be important for a real neuron? Stochastic HH simulations with realistic channel densities suggest that the timing of spikes in response to white noise stimuli is reproducible to within 1 to 2 msec (Schneidman, Freedman, & Segev, 1998), a figure that is comparable to what is observed for pyramidal cells in vitro (Mainen & Sejnowski, 1995), as well in vivo in the fly's visual system (de Ruyter van Steveninck, Lewen, Strong, Koberle, & Bialek, 1997; Lewen, Bialek, & de Ruyter van Steveninck, 2001), the vertebrate retina (Berry, Warland, & Meister, 1997), the cat lateral geniculate nucleus (LGN) (Reinagel & Reid, 2000), and the bat auditory cortex (Dear, Simmons, & Fritz, 1993). This suggests that such timing details may indeed be important. We must therefore ask why our approximation seems to carry an inherent time resolution limitation and why, even at its optimal resolution, the full information in the spike is not recovered.

For many purposes, recovering 75% of the information at $\sim$ 3 msec resolution might be considered a resounding success. On the other hand, with such a simple underlying model, we would hope for a more compelling conclusion. From a methodological point of view, it behooves us to ask what we are missing in our 2D model, and perhaps the methods we use in finding the missing information in the present case will prove applicable more generally.

6 What Is Missing?

The obvious first approach to improving the 2D approximation is to add more dimensions. Let us consider the neglected modes. We recall from Figure 11 that in simulations with very large numbers of spikes, we can isolate at least two more modes that have significant eigenvalues and are associated with the isolated spike rather than the preceding silence; these are shown in Figure 16. We see that these modes look like higher-order derivatives, which makes some sense since we are missing information at high time resolution. On the other hand, if all we are doing is expanding in a basis of higher-order derivatives it is not clear that we will do qualitatively better by including one or two more terms, particularly given that sampling higher-dimensional distributions becomes very difficult.

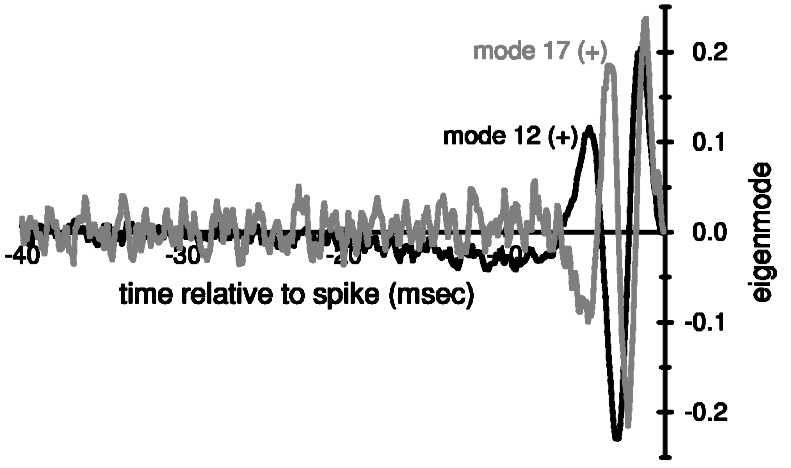

Figure 16: The two next spike-associated modes. These resemble higher-order derivatives.

Our original model attempted to approximate the set of relevant features as lying within a K-dimensional linear subspace of the original D-dimensional input space. The covariance spectrum indicates that additional dimensions play a small but significant role. We suggest that a reasonable next step is to consider the relevant feature space to be low dimensional but not flat. The first two covariance modes define a plane; we will consider next a 2D geometric construction that curves into additional dimensions.

Several methods have been proposed for the general problem of identifying low-dimensional nonlinear manifolds (Cottrell, Munro, & Zipser, 1988; Boser, Guyon, & Vapnik, 1992; Guyon, Boser, & Vapnik, 1993; Oja & Karhunen, 1995; Roweis & Saul, 2000), but these various approaches share the disadvantage that the manifold or, equivalently, the relevant set of features remains implicit. Our hope is to understand the behavior of the neuron explicitly; we therefore wish to obtain an explicit representation of this curved feature space in terms of the basis vectors (features) that span it.

A first approach to determining the curved subspace is to approximate it as a set of locally linear "tiles." At any place on the surface, we wish to find two orthogonal directions that form the surface's local linear approximation. Curvature of the subspace means that these two dimensions will vary across the surface. We apply a simple algorithm to construct a locally linear tiling with the advantage that it requires only first-order statistics. First, we take one of the directions to be globally defined; it is natural to take this direction to be the overall spike-triggered average. Allowing for curvature in the feature subspace means that along the direction of the STA, we allow the second, orthogonal direction to vary. In principle, this vari-

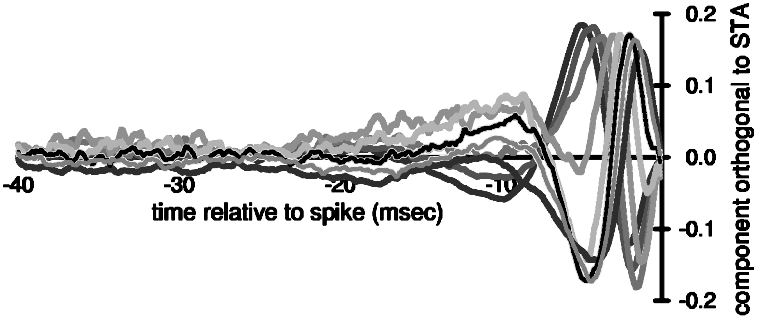

Figure 17: The orthonormal components of spike-triggered averages from 80,000 spikes conditioned on their projection onto the overall spike-triggered average (eight conditional averages shown).

ation is continuous, but we will not be able to sample sufficiently to find the complete description of the second direction, so we sort the stimulus histories according to their projection along this direction and bin the sorted histories into a small number of bins. This defines the number of tiles used to cover the surface. We determine the conditional average of the stimulus histories in each bin and compute the (normalized) component orthogonal to the overall STA. This provides a second locally meaningful basis vector for the subspace in that bin. The resulting family of curves orthogonal to the STA is shown in Figure 17. Applying singular value decomposition to the family of curves shows that there are at least four significant independent directions in stimulus space apart from the STA. This gives us an estimate of the embedding dimension of the feature subspace.

Geometrically, this construction is equivalent to approximating the surface as a twisting ribbon, with the leading direction that of the STA, but where the surface is allowed to rotate about the STA axis. Further, we have discretized the twisting direction into a small number of tiles. The discretization is fixed by the size of the data. There is a trade-off between the precision of estimating the conditional average and the fidelity with which one follows the twist. Here we have restricted ourselves to an experimentally realistic number of isolated spikes (80,000), using an equal number of spikes per bin. Varying over the number of bins, we found that eight bins gave the best result. Note that our model is discontinuous; it might be possible to improve it by interpolating smoothly between successive tiles.

Computing the information as a function of Δt using this locally linear model, we obtain the curve shown in Figure 18, where the results can be compared against the information found from the STA alone and from the covariance modes. The information from the new model captures a maxi-

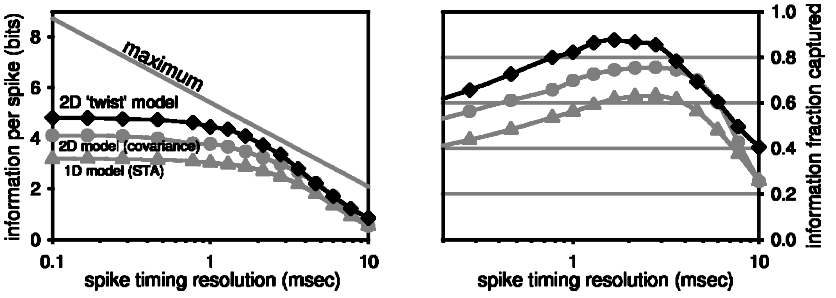

Figure 18: Bits per spike (left) and fraction of the theoretical limit (right) of timing information in a single spike at a given temporal resolution captured by the locally linear tiling "twist" model (diamonds), compared to models using the STA alone (triangles), and projection onto ΔC covariance modes 1 and 2 (circles).

mum of 4.8 bits, recovering $\sim$ 90% of the information at a time resolution of approximately 1 msec.

One of the main strengths of this simple approach is that we have succeeded in extracting additional geometrical information about the feature subspace using very limited data, as we compute only averages. Note that a similar number of spikes cannot resolve more than two spike-associated covariance modes in the covariance matrix analysis.

7 Discussion

The HH equations describe the dynamics of four degrees of freedom, and almost since these equations were first written down, there have been attempts to find simplifications or reductions. FitzHugh and Nagumo proposed a 2D system of equations that approximate the HH model (Fitzhugh, 1961; Nagumo, Arimoto, & Yoshikawa, 1962), and this has the advantage that one can visualize the trajectories directly in the plane and thus achieve an intuitive graphical understanding of the dynamics and its dependence on parameters. The need for reduction in the sense pioneered by FitzHugh and by Nagumo et al. has become only more urgent with the growing use of increasingly complex HH-style model neurons with many different channel types. With this problem in mind, Kepler, Abbott, and Marder have introduced reduction methods that are more systematic, making use of the difference in timescales among the gating variables (Kepler, Abbott, & Marder, 1992; Abbott & Kepler, 1990).

In the presence of constant current inputs, it makes sense to describe the HH equations as a 4D autonomous dynamical system; by well-known methods in dynamical systems theory, one could consider periodic input

currents by adding an extra dimension. The question asked by FitzHugh and Nagumo was whether this 4D or 5D description could be reduced to two or three dimensions.

Closer in spirit to our approach is the work by Kistler, Gerstner, and van Hemmen (1997), who focused in particular on the interaction among successive action potentials. They argued that one could approximate the HH model by a nearly linear dynamical system with a threshold, identifying threshold crossing with spiking, provided that each spike generated either a change in threshold or an effective input current that influences the generation of subsequent spikes.

The notion of model dimensionality considered here is distinct from the dynamical systems perspective in which one simply counts the system's degrees of freedom. Here we are attempting to find a description of the dynamics, which is essentially functional or computational. We have identified the output of the system as spike times, and our aim is to construct as complete a description as possible of the mapping between input and output. The dimensionality of our model is that of the space of inputs relevant for this mapping. There is no necessary relationship between these two notions of dimensionality. For example, in a neural network with two attractors, a system described by a potentially large number of variables, there might be a simple rule (perhaps even a linear filter) that allows us to look at the inputs to the network and determine the times at which the switching events will occur. Conversely, once we leave the simplified world of constant or periodic inputs, even the small number of differential equations describing a neuron's channel dynamics could in principle be equivalent to a very complicated set of rules for mapping inputs into spike times.

In our context, simplicity is (roughly) feature selectivity: the mapping is simple if spiking is determined by a small number of features in the complex history of inputs. Following the ideas that emerged in the analysis of motion-sensitive neurons in the fly (de Ruyter van Steveninck & Bialek, 1988; Bialek & de Ruyter van Steveninck, 2003), we have identified "features" with "dimensions" and searched for low-dimensional descriptions of the input history that preserve the mutual information between inputs and outputs (spike times). We have considered only the generation of isolated spikes, leaving aside the question of how spikes interact with one another, as considered by Kistler et al. (1997). For these isolated spikes, we began by searching for projections onto a low-dimensional linear subspace of the originally $\sim$ 200-dimensional stimulus space, and we found that a substantial fraction of the mutual information could be preserved in a model with just two dimensions. Searching for the information that is missing from this model, we found that rather than adding more (Euclidean) dimensions, we could capture approximately 90% of the information at high time resolution by keeping a 2D description but allowing these dimensions to vary over the surface, so that the neuron is sensitive to stimulus features that lie in a curved 2D subspace.

The geometrical picture of neurons as being sensitive to features that are defined by a low-dimensional stimulus subspace is attractive and, as noted in section 1, corresponds to a widely shared intuition about the nature of neuronal feature selectivity. While curved subspaces often appear as the targets for learning in complex neural computations such as invariant object recognition, the idea that such subspaces appear already in the description of single-neuron computation we believe to be novel.

While we have exploited the fact that long simulations of the HH model are quite tractable to generate large amounts of "data" for our analysis, it is important that, in the end, our construction of a curved, relevant, stimulus subspace involves a series of computations that are just simple generalizations of the conventional reverse correlation or spike-triggered average. This suggests that our approach can be applied to real neurons without requiring qualitatively larger data sets than might have been needed for a careful reverse correlation analysis. In the same spirit, recent work has shown how covariance matrix analysis of the fly's motion-sensitive neurons can reveal nonlinear computations in a 4D subspace using data sets of fewer than 10^4 spikes (Bialek & de Ruyter van Steveninck, 2003). Low-dimensional linear subspaces can be found even in the response of model neurons to naturalistic inputs if one searches directly for dimensions that capture the largest fraction of the mutual information between inputs and spikes (Sharpee et al., in press), and again the errors involved in identifying the relevant dimensions are comparable to the errors in reverse correlation (Sharpee, Rust, & Bialek, 2003). All of these results point to the practical feasibility of describing real neurons in terms of nonlinear computation on low-dimensional relevant subspaces in a high-dimensional stimulus space.

Our reduced model of the HH neuron both illustrates a novel approach to dimensional reduction and gives new insight into the computation performed by the neuron. The reduced model is essentially that of an edge detector for current trajectories but is sensitive to a further stimulus parameter, producing a curved manifold. An interpretation of this curvature will be presented in a forthcoming manuscript. This curved representation is able to capture almost all information that isolated spikes convey about the stimulus, or conversely, allow us to predict isolated spike times with high temporal precision from the stimulus. The emergence of a low-dimensional curved manifold in a model as simple as the HH neuron suggests that such a description may also be appropriate for biological neurons.

Our approach is limited in that we address only isolated spikes. This restricted class of spikes nonetheless has biological relevance; for example, in vertebrate retinal ganglion cells (Berry & Meister, 1999), in rat somatosensory cortex (Panzeri, Petersen, Schultz, Lebedev, & Diamond, 2001), and in LGN (Reinagel, Godwin, Sherman, & Koch, 1999), the first spike of a burst has been shown to convey distinct (and the majority of the) information.

However, a clear next step in this program is to extend our formalism to take into account interspike interaction. For neurons or models with explicit long timescales, adaptation induces very long-range history dependence, which complicates the issue of spike interactions considerably. A full understanding of the interaction between stimulus and spike history will therefore in general involve understanding the meanings of spike patterns (de Ruyter van Steveninck & Bialek, 1988; Brenner, Strong, et al., 2000) and the influence of the larger statistical context (Fairhall et al., 2001). Our results point to the need for a more parsimonious description of self-excitation, even for the simple case of dependence on only the last spike time.

We close by reminding readers of the more ambitious goal of building bridges between the burgeoning molecular-level description of neurons and the functional or computational level. Armed with a description of spike generation as a nonlinear operation on a low-dimensional, curved manifold in the space of inputs, it is natural to ask how the details of this computational picture are related to molecular mechanisms. Are neurons with more different types of ion channels sensitive to more stimulus dimensions, or do they implement more complex nonlinearities in a low-dimensional space? Are adaptation and modulation mechanisms that change the nonlinearity separable from those that change the dimensions to which the cell is sensitive? Finally, while we have shown how a low-dimensional description can be constructed numerically from observations of the input-output properties of the neuron, one would like to understand analytically why such a description emerges and whether it emerges universally from the combinations of channel dynamics selected by real neurons.

Acknowledgments

We thank N. Brenner for discussions at the start of this work, and M. Berry for comments on the manuscript.

References

Abbott, L. F., & Kepler, T. (1990). Model neurons: From Hodgkin-Huxley to Hopfield. In L. Garrido (Ed.), *Statistical mechanisms of neural networks* (pp. 5–18). Berlin: Springer-Verlag.

Agüera y Arcas, B. (1998). *Reducing the neuron: A computational approach.* Unpublished master's thesis, Princeton University.

Agüera y Arcas, B., Bialek, W., & Fairhall, A. L. (2001). What can a single neuron compute? In T. Leen, T. Dietterich, & V. Tresp (Eds.), *Advances in neural information processing systems, 13* (pp. 75–81). Cambridge, MA: MIT Press.

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

Barlow, H. B. (1953). Summation and inhibition in the frog's retina. *J. Physiol., 119,* 69–88.

Barlow, H. B., Hill, R. M., & Levick, W. R. (1964). Retinal ganglion cells responding selectively to direction and speed of image motion in the rabbit. *J. Physiol., 173*, 377–407.

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

Berry, M. J. II, Warland, D., & Meister, M. (1997). The structure and precision of retinal spike trains. *Proc. Natl. Acad. Sci. U.S.A., 94*, 5411–5416.

Bialek, W., & de Ruyter van Steveninck, R. R. (2003). *Features and dimensions: Motion estimation in fly vision*. Unpublished manuscript.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In D. Haussler (Ed.), *5th Annual ACM Workshop on COLT* (pp. 144–152). Pittsburgh, PA: ACM Press.

Bray, D. (1995). Protein molecules as computational elements in living cells. *Nature, 376*, 307–312.

Brenner, N., Agam, O., Bialek, W., & de Ruyter van Steveninck, R. R. (1998). Universal statistical behavior of neural spike trains. *Phys. Rev. Lett., 81*, 4000–4003.

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

Brenner, N., Strong, S., Koberle, R., Bialek, W., & de Ruyter van Steveninck, R. R. (2000). Synergy in a neural code. *Neural Comp., 12*, 1531–1552. Available on-line: http://xxx.lanl.gov/abs/physics/9902067.

Cottrell, G. W., Munro, P., & Zipser, D. (1988). Image compression by back propagation: A demonstration of extensional programming. In N. Sharkey (Ed.), *Models of cognition: A review of cognitive science* (Vol. 2, pp. 208–240). Norwood, NJ: Ablex.

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. New York: Wiley.

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

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

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

Dear, S. P., Simmons, J. A., & Fritz, J. (1993). A possible neuronal basis for representation of acoustic scenes in auditory cortex of the big brown bat. *Nature, 364*, 620–623.

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

Fitzhugh, R. (1961). Impulse and physiological states in models of nerve membrane. *Biophysics J., 1*, 445–466.

Guyon, I. M., Boser, B. E., & Vapnik, V. N. (1993). Automatic capacity tuning of very large VC-dimension classifiers. In S. J. Hanson, J. D. Cowan, & C. Giles (Eds.), *Advances in neural information processing systems, 5* (pp. 147–155). San Mateo, CA: Morgan Kaufmann.

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

Hille, B. (1992). *Ionic channels of excitable membranes*. Sunderland, MA: Sinauer.

Hodgkin, A. L., & Huxley, A. F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physiol., 463*, 391–407.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol. (Lond.), 160*, 106–154.

Iverson, L., & Zucker, S. W. (1995). Logical/linear operators for image curves. *IEEE Trans. Pattern Analysis and Machine Intelligence, 17*, 982–996.

Keat, J., Reinagel, P., Reid, R. C., & Meister, M. (2001). Predicting every spike: A model for the responses of visual neurons. *Neuron, 30*(3), 803–817.

Kepler, T., Abbott, L. F., & Marder, E. (1992). Reduction of conductance-based neuron models. *Biological Cybernetics, 66*, 381–387.

Kistler, W., Gerstner, W., & van Hemmen, J. L. (1997). Reduction of the Hodgkin-Huxley equations to a single-variable threshold model. *Neural Computation, 9*, 1015–1045.

Koch, C. (1999). *Biophysics of computation: Information processing in single neurons*. New York: Oxford University Press.

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

Lewen, G. D., Bialek, W., & de Ruyter van Steveninck, R. R. (2001). Neural coding of naturalistic motion stimuli. *Network, 12*, 317–329.

Mainen, Z. F., & Sejnowski, T. J. (1995). Reliability of spike timing in neocortical neurons. *Science, 268*, 1503–1506.

Nagumo, J., Arimoto, S., & Yoshikawa, Z. (1962). An active pulse transmission line simulating nerve axon. *Proc. IRE, 50*, 2061–2071.

Oja, E., & Karhunen, J. (1995). Signal separation by nonlinear Hebbian learning. In M. Palaniswami, Y. Attikiouzel, R. J. Marks II, D. Fogel, & T. Fukuda (Eds.), *Computational intelligence—a dynamic system perspective* (pp. 83–97). New York: IEEE Press.

Panzeri, S., Petersen, R., Schultz, S., Lebedev, M., & Diamond, M. (2001). The role of spike timing in the coding of stimulus location in rat somatosensory cortex. *Neuron, 29*, 769–777.

Reinagel, P., Godwin, D., Sherman, S. M., & Koch, C. (1999). Encoding of visual information by LGN bursts. *J. Neurophys., 81*, 2558–2569.

Reinagel, P., & Reid, R. C. (2000). Temporal coding of visual information in the thalamus. *J. Neuroscience, 20*(14), 5392–5400.

Rieke, F., Warland, D., Bialek, W., & de Ruyter van Steveninck, R. R. (1997). *Spikes: Exploring the neural code*. Cambridge, MA: MIT Press.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review, 65*, 386–408.

Rosenblatt, F. (1962). *Principles of neurodynamics*. New York: Spartan Books.

Roweis, S., & Saul, L. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science, 290*, 2323–2326.

Schneidman, E., Freedman, B., & Segev, I. (1998). Ion channel stochasticity may be critical in determining the reliability and precision of spike timing. *Neural Comp., 10*, 1679–1703.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Sys. Tech. Journal, 27*, 379–423, 623–656.

Sharpee, T., Rust, N. C., & Bialek, W. (in press). Maximally informative dimensions: analysing neural responses to natural signals. *Neural Information Processing Systems 2002.* Available on-line: http://xxx.lanl.gov/abs/physics/0208057.

Sharpee, T., Rust, N. C., & Bialek, W. (2003). *Maximally informative dimensions: Analysing neural responses to natural signals.* Unpublished manuscript.

Stanley, G. B., Lei, F. F., & Dan, Y. (1999). Reconstruction of natural scenes from ensemble responses in the lateral geniculate nucleus. *J. Neurosci., 19*(18), 8036–8042.

Theunissen, F., Sen, K., & Doupe, A. (2000). Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. *J. Neurosci., 20*, 2315–2331.

Tiesinga, P. H. E., José, J., & Sejnowski, T. (2000). Comparison of current-driven and conductance-driven neocortical model neurons with Hodgkin-Huxley voltage-gated channels. *Physical Review E, 62*, 8413–8419.

Tishby, N., Pereira, F., & Bialek, W. (1999). The information bottleneck method. In B. Hajek & R. S. Sreenivas (Eds.), *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing* (pp. 368–377). Champaign, IL. Available on-line: http://xxx.lanl.gov/abs/physics/0004057.

Received January 9, 2003; accepted January 28, 2003.