ARTICLE Communicated by Pamela Reinagel

Analyzing Neural Responses to Natural Signals: Maximally Informative Dimensions

Tatyana Sharpee
sharpee@phy.ucsf.edu
Sloan–Swartz Center for Theoretical Neurobiology and Department of Physiology, University of California at San Francisco, San Francisco, CA 94143, U.S.A.

Nicole C. Rust
rust@cns.nyu.edu
Center for Neural Science, New York University, New York, NY 10003, U.S.A.

William Bialek
wbialek@princeton.edu
Department of Physics, Princeton University, Princeton, NJ 08544, U.S.A., and Sloan–Swartz Center for Theoretical Neurobiology and Department of Physiology, University of California at San Francisco, San Francisco, CA 94143, U.S.A.

We propose a method that allows for a rigorous statistical analysis of neural responses to natural stimuli that are nongaussian and exhibit strong correlations. We have in mind a model in which neurons are selective for a small number of stimulus dimensions out of a high-dimensional stimulus space, but within this subspace the responses can be arbitrarily nonlinear. Existing analysis methods are based on correlation functions between stimuli and responses, but these methods are guaranteed to work only in the case of gaussian stimulus ensembles. As an alternative to correlation functions, we maximize the mutual information between the neural responses and projections of the stimulus onto low-dimensional subspaces. The procedure can be done iteratively by increasing the dimensionality of this subspace. Those dimensions that allow the recovery of all of the information between spikes and the full unprojected stimuli describe the relevant subspace. If the dimensionality of the relevant subspace indeed is small, it becomes feasible to map the neuron's input-output function even under fully natural stimulus conditions. These ideas are illustrated in simulations on model visual and auditory neurons responding to natural scenes and sounds, respectively.

1 Introduction

From olfaction to vision and audition, a growing number of experiments are examining the responses of sensory neurons to natural stimuli (Creutzfeldt & Northdurft, 1978; Rieke, Bodnar, & Bialek, 1995; Baddeley et al., 1997; Stanley, Li, & Dan, 1999; Theunissen, Sen, & Doupe, 2000; Vinje & Gallant, 2000, 2002; Lewen, Bialek, & de Ruyter van Steveninck, 2001; Sen, Theunissen, & Doupe, 2001; Vickers, Christensen, Baker, & Hildebrand, 2001; Ringach, Hawken, & Shapley, 2002; Weliky, Fiser, Hunt, & Wagner, 2003; Rolls, Aggelopoulos, & Zheng, 2003; Smyth, Willmore, Baker, Thompson, & Tolhurst, 2003). Observing the full dynamic range of neural responses may require using stimulus ensembles that approximate those occurring in nature (Rieke, Warland, de Ruyter van Steveninck, & Bialek, 1997; Simoncelli & Olshausen, 2001), and it is an attractive hypothesis that the neural representation of these natural signals may be optimized in some way (Barlow, 1961, 2001; von der Twer & Macleod, 2001; Bialek, 2002). Many neurons exhibit strongly nonlinear and adaptive responses that are unlikely to be predicted from a combination of responses to simple stimuli; for example, neurons have been shown to adapt to the distribution of sensory inputs, so that any characterization of these responses will depend on context (Smirnakis, Berry, Warland, Bialek, & Meister, 1996; Brenner, Bialek, & de Ruyter van Steveninck, 2000; Fairhall, Lewen, Bialek, & de Ruyter van Steveninck, 2001). Finally, the variability of a neuron's responses decreases substantially when complex dynamical, rather than static, stimuli are used (Mainen & Sejnowski, 1995; de Ruyter van Steveninck, Lewen, Strong, Koberle, & Bialek, 1997; Kara, Reinagel, & Reid, 2000; de Ruyter van Steveninck, Borst, & Bialek, 2000). All of these arguments point to the need for general tools to analyze the neural responses to complex, naturalistic inputs.

The stimuli analyzed by sensory neurons are intrinsically high-dimensional, with dimensions $D \sim 10^2 - 10^3$. For example, in the case of visual neurons, the input is commonly specified as light intensity on a grid of at least 10×10 pixels. Each of the presented stimuli can be described as a vector $\mathbf{s}$ in this high-dimensional stimulus space (see Figure 1). The dimensionality becomes even larger if stimulus history has to be considered as well. For example, if we are interested in how the past N frames of the movie affect the probability of a spike, then the stimulus $\mathbf{s}$, being a concatenation of the past N samples, will have dimensionality N times that of a single frame. We also assume that the probability distribution $P(\mathbf{s})$ is sampled during an experiment ergodically, so that we can exchange averages over time with averages over the true distribution as needed.

Although direct exploration of a $D \sim 10^2 - 10^3$-dimensional stimulus space is beyond the constraints of experimental data collection, progress can be made provided we make certain assumptions about how the response has been generated. In the simplest model, the probability of response can be described by one receptive field (RF) or linear filter (Rieke et al., 1997).

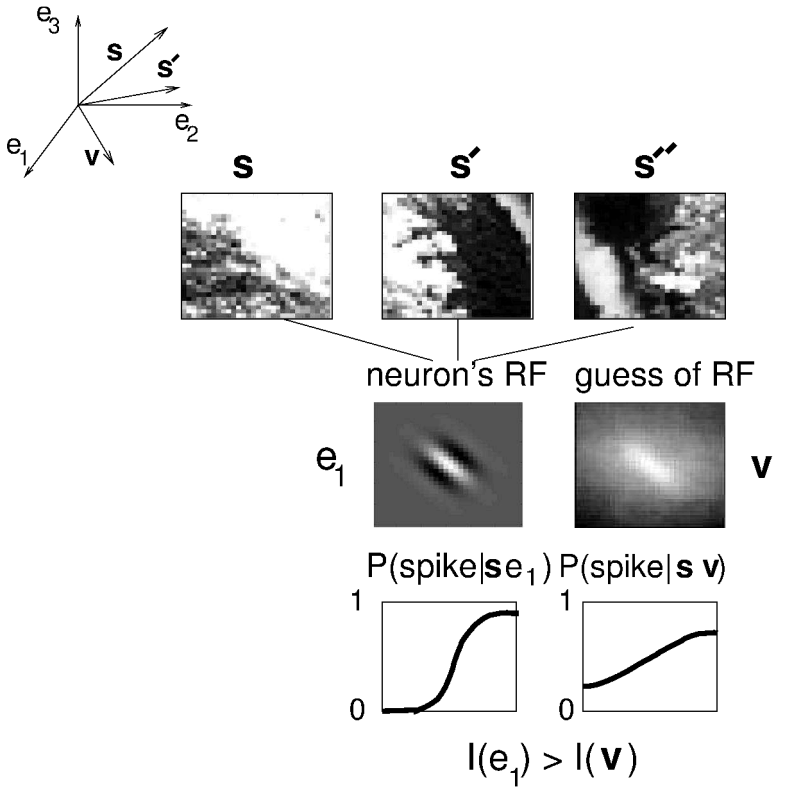

Figure 1: Schematic illustration of a model with a one–dimensional relevant subspace: $\hat{e}_1$ is the relevant dimension, and $\hat{e}_2$ and $\hat{e}_3$ are irrelevant ones. Shown are three example stimuli, $\mathbf{s}$, $\mathbf{s}'$, and $\mathbf{s}''$, the receptive field of a model neuron—the relevant dimension $\hat{e}_1$, and our guess $\mathbf{v}$ for the relevant dimension. Probabilities of a spike $P(\text{spike}|\mathbf{s}\cdot\hat{e}_1)$ and $P(\text{spike}|\mathbf{s}\cdot\mathbf{v})$ are calculated by first projecting all of the stimuli $\mathbf{s}$ onto each of the two vectors $\hat{e}_1$ and $\mathbf{v}$, respectively, and then applying equations (2.3, 1.2, 1.1) sequentially. Our guess $\mathbf{v}$ for the relevant dimension is adjusted during the progress of the algorithm in such a way as to maximize $I(\mathbf{v})$ of equation (2.5), which makes vector $\mathbf{v}$ approach the true relevant dimension $\hat{e}_1$.

The RF can be thought of as a template or special direction $\hat{e}_1$ in the stimulus space[1] such that the neuron's response depends on only a projection of a given stimulus $\mathbf{s}$ onto $\hat{e}_1$, although the dependence of the response on this

[1] The notation $\hat{e}$ denotes a unit vector, since we are interested only in the direction the vector specifies and not in its length.

projection can be strongly nonlinear (cf. Figure 1). In this simple model, the reverse correlation method (de Boer & Kuyper, 1968; Rieke et al., 1997; Chichilnisky, 2001) can be used to recover the vector $\hat{e}_1$ by analyzing the neuron's responses to gaussian white noise. In a more general case, the probability of the response depends on projections $s_i = \hat{e}_i \cdot \mathbf{s}$ of the stimulus $\mathbf{s}$ on a set of K vectors $\{\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_K\}$:

$$P(\text{spike}|\mathbf{s}) = P(\text{spike}) f(s_1, s_2, \ldots, s_K), \tag{1.1}$$

where $P(\text{spike}|\mathbf{s})$ is the probability of a spike given a stimulus $\mathbf{s}$ and $P(\text{spike})$ is the average firing rate. In what follows, we will call the subspace spanned by the set of vectors $\{\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_K\}$ the relevant subspace (RS).[2] We reiterate that vectors $\{\hat{e}_i\}, 1 \le i \le K$ may also describe how the time dependence of the stimulus $\mathbf{s}$ affects the probability of a spike. An example of such a relevant dimension would be a spatiotemporal RF of a visual neuron. Although the ideas developed below can be used to analyze input-output functions f with respect to different neural responses, such as patterns of spikes in time (de Ruyter van Steveninck & Bialek, 1988; Brenner, Strong, Koberle, Bialek, & de Ruyter van Steveninck, 2000; Reinagel & Reid, 2000), for illustration purposes we choose a single spike as the response of interest.[3]

Equation 1.1 in itself is not yet a simplification if the dimensionality K of the RS is equal to the dimensionality D of the stimulus space. In this article, we will assume that the neuron's firing is sensitive only to a small number of stimulus features ($K \ll D$). While the general idea of searching for low-dimensional structure in high-dimensional data is very old, our motivation here comes from work on the fly visual system, where it was shown explic-

[2] Since the analysis does not depend on a particular choice of a basis within the full D-dimensional stimulus space, for clarity we choose the basis in which the first K basis vectors span the relevant subspace and the remaining $D - K$ vectors span the irrelevant subspace.

[3] We emphasize that our focus here on single spikes is not equivalent to assuming that the spike train is a Poisson process modulated by the stimulus. No matter what the statistical structure of the spike train is, we always can ask what features of the stimulus are relevant for setting the probability of generating a single spike at one moment in time. From an information-theoretic point of view, asking for stimulus features that capture the mutual information between the stimulus and the arrival times of single spikes is a well-posed question even if successive spikes do not carry independent information. Note also that spikes' carrying independent information is not the same as spikes' being generated as a Poisson process. On the other hand, if (for example) different temporal patterns of spikes carry information about different stimulus features, then analysis of single spikes will result in a relevant subspace of artifactually high dimensionality. Thus, it is important that the approach discussed here carries over without modification to the analysis of relevant dimensions for the generation of any discrete event, such as a pattern of spikes across time in one cell or synchronous spikes across a population of cells. For a related discussion of relevant dimensions and spike patterns using covariance matrix methods see de Ruyter van Steveninck and Bialek (1988) and Agüera y Arcas, Fairhall, and Bialek (2003).

itly that patterns of action potentials in identified motion-sensitive neurons are correlated with low-dimensional projections of the high-dimensional visual input (de Ruyter van Steveninck & Bialek, 1988; Brenner, Bialek, et al., 2000; Bialek & de Ruyter van Steveninck, 2003). The input-output function f in equation 1.1 can be strongly nonlinear, but it is presumed to depend on only a small number of projections. This assumption appears to be less stringent than that of approximate linearity, which one makes when characterizing a neuron's response in terms of Wiener kernels (see, e.g., the discussion in section 2.1.3 of Rieke et al., 1997). The most difficult part in reconstructing the input-output function is to find the RS. Note that for $K > 1$, a description in terms of any linear combination of vectors $\{\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_K\}$ is just as valid, since we did not make any assumptions as to a particular form of the nonlinear function f.

Once the relevant subspace is known, the probability $P(\text{spike}|\mathbf{s})$ becomes a function of only a few parameters, and it becomes feasible to map this function experimentally, inverting the probability distributions according to Bayes' rule:

$$f(s_1, s_2, \ldots, s_K) = \frac{P(s_1, s_2, \ldots, s_K|\text{spike})}{P(s_1, s_2, \ldots, s_K)}. \tag{1.2}$$

If stimuli are chosen from a correlated gaussian noise ensemble, then the neural response can be characterized by the spike-triggered covariance method (de Ruyter van Steveninck & Bialek, 1988; Brenner, Bialek, et al., 2000; Schwartz, Chichilnisky, & Simoncelli, 2002; Touryan, Lau, & Dan, 2002; Bialek & de Ruyter van Steveninck, 2003). It can be shown that the dimensionality of the RS is equal to the number of nonzero eigenvalues of a matrix given by a difference between covariance matrices of all presented stimuli and stimuli conditional on a spike. Moreover, the RS is spanned by the eigenvectors associated with the nonzero eigenvalues multiplied by the inverse of the a priori covariance matrix. Compared to the reverse correlation method, we are no longer limited to finding only one of the relevant dimensions $\{\hat{e}_i\}$, $1 \le i \le K$. Both the reverse correlation and the spike–triggered covariance method, however, give rigorously interpretable results only for gaussian distributions of inputs.

In this article, we investigate whether it is possible to lift the requirement for stimuli to be gaussian. When using natural stimuli, which certainly are nongaussian, the RS cannot be found by the spike-triggered covariance method. Similarly, the reverse correlation method does not give the correct RF, even in the simplest case where the input-output function in equation 1.1 depends on only one projection (see appendix A for a discussion of this point). However, vectors that span the RS are clearly special directions in the stimulus space independent of assumptions about $P(\mathbf{s})$. This notion can be quantified by the Shannon information. We note that the stimuli $\mathbf{s}$ do not have to lie on a low-dimensional manifold within the overall D-dimensional

space.[4] However, since we assume that the neuron's input-output function depends on a small number of relevant dimensions, the ensemble of stimuli conditional on a spike may exhibit clear clustering. This makes the proposed method of looking for the RS complementary to the clustering of stimuli conditional on a spike done in the information bottleneck method (Tishby, Pereira, & Bialek, 1999; see also Dimitrov & Miller, 2001). Noninformation-based measures of similarity between probability distributions $P(\mathbf{s})$ and $P(\mathbf{s}|\text{spike})$ have also been proposed to find the RS (Paninski, 2003a).

To summarize our assumptions:

- The sampling of the probability distribution of stimuli $P(\mathbf{s})$ is ergodic and stationary across repetitions. The probability distribution is not assumed to be gaussian. The ensemble of stimuli described by $P(\mathbf{s})$ does not have to lie on a low-dimensional manifold embedded in the overall D-dimensional space.
- We choose a single spike as the response of interest (for illustration purposes only). An identical scheme can be applied, for example, to particular interspike intervals or to synchronous spikes from a pair of neurons.
- The subspace relevant for generating a spike is low dimensional and Euclidean (cf. equation 1.1).
- The input-output function, which is defined within the low-dimensional RS, can be arbitrarily nonlinear. It is obtained experimentally by sampling the probability distributions $P(\mathbf{s})$ and $P(\mathbf{s}|\text{spike})$ within the RS.

The article is organized as follows. In section 2 we discuss how an optimization problem can be formulated to find the RS. A particular algorithm used to implement the optimization scheme is described in section 3. In section 4 we illustrate how the optimization scheme works with natural stimuli for model orientation-sensitive cells with one and two relevant dimensions, much like simple and complex cells found in primary visual cortex, as well as for a model auditory neuron responding to natural sounds. We also discuss the convergence of our estimates of the RS as a function of data set size. We emphasize that our optimization scheme does not rely on any specific statistical properties of the stimulus ensemble and thus can be used with natural stimuli.

[4] If one suspects that neurons are sensitive to low-dimensional features of their input, one might be tempted to analyze neural responses to stimuli that explore only the (putative) relevant subspace, perhaps along the line of the subspace reverse correlation method (Ringach et al., 1997). Our approach (like the spike-triggered covariance approach) is different because it allows the analysis of responses to stimuli that live in the full space, and instead we let the neuron "tell us" which low-dimensional subspace is relevant.

2 Information as an Objective Function

When analyzing neural responses, we compare the a priori probability distribution of all presented stimuli with the probability distribution of stimuli that lead to a spike (de Ruyter van Steveninck & Bialek, 1988). For gaussian signals, the probability distribution can be characterized by its second moment, the covariance matrix. However, an ensemble of natural stimuli is not gaussian, so that in a general case, neither second nor any finite number of moments is sufficient to describe the probability distribution. In this situation, Shannon information provides a rigorous way of comparing two probability distributions. The average information carried by the arrival time of one spike is given by Brenner, Strong, et al. (2000),

$$I_{\text{spike}} = \int d\mathbf{s} P(\mathbf{s}|\text{spike}) \log_2 \left[\frac{P(\mathbf{s}|\text{spike})}{P(\mathbf{s})} \right], \tag{2.1}$$

where $d\mathbf{s}$ denotes integration over full D–dimensional stimulus space. The information per spike as written in equation 2.1 is difficult to estimate experimentally, since it requires either sampling of the high-dimensional probability distribution $P(\mathbf{s}|\text{spike})$ or a model of how spikes were generated, that is, the knowledge of low-dimensional RS. However, it is possible to calculate I_{spike} in a model-independent way if stimuli are presented multiple times to estimate the probability distribution $P(\text{spike}|\mathbf{s})$. Then,

$$I_{\text{spike}} = \left\langle \frac{P(\text{spike}|\mathbf{s})}{P(\text{spike})} \log_2 \left[\frac{P(\text{spike}|\mathbf{s})}{P(\text{spike})} \right] \right\rangle_{\mathbf{s}}, \tag{2.2}$$

where the average is taken over all presented stimuli. This can be useful in practice (Brenner, Strong, et al., 2000), because we can replace the ensemble average $\langle\rangle_{\mathbf{s}}$ with a time average and $P(\text{spike}|\mathbf{s})$ with the time-dependent spike rate $r(t)$. Note that for a finite data set of N repetitions, the obtained value $I_{\text{spike}}(N)$ will be on average larger than $I_{\text{spike}}(\infty)$. The true value I_{spike} can be found by either subtracting an expected bias value, which is of the order of $\sim 1/(P(\text{spike})N 2\ln 2)$ (Treves & Panzeri, 1995; Panzeri & Treves, 1996; Pola, Schultz, Petersen, & Panzeri, 2002; Paninski, 2003b) or extrapolating to $N \to \infty$ (Brenner, Strong, et al., 2000; Strong, Koberle, de Ruyter van Steveninck, & Bialek, 1998). Measurement of I_{spike} in this way provides a model independent benchmark against which we can compare any description of the neuron's input-output relation.

Our assumption is that spikes are generated according to a projection onto a low-dimensional subspace. Therefore, to characterize relevance of a particular direction $\mathbf{v}$ in the stimulus space, we project all of the presented stimuli onto $\mathbf{v}$ and form probability distributions $P_{\mathbf{v}}(x)$ and $P_{\mathbf{v}}(x|\text{spike})$ of projection values x for the a priori stimulus ensemble and that conditional

on a spike, respectively:

$$P_{\mathbf{v}}(x) = \langle \delta(x - \mathbf{s}\cdot\mathbf{v})\rangle_{\mathbf{s}}, \tag{2.3}$$

$$P_{\mathbf{v}}(x|\text{spike}) = \langle \delta(x - \mathbf{s}\cdot\mathbf{v})|\text{spike}\rangle_{\mathbf{s}}, \tag{2.4}$$

where $\delta(x)$ is a delta function. In practice, both the average $\langle\cdots\rangle_{\mathbf{s}} \equiv \int d\mathbf{s}\cdots P(\mathbf{s})$ over the a priori stimulus ensemble and the average $\langle\cdots|\text{spike}\rangle_{\mathbf{s}} \equiv \int d\mathbf{s}\cdots P(\mathbf{s}|\text{spike})$ over the ensemble conditional on a spike are calculated by binning the range of projection values x. The probability distributions are then obtained as histograms, normalized in a such a way that the sum over all bins gives 1. The mutual information between spike arrival times and the projection x, by analogy with equation 2.1, is

$$I(\mathbf{v}) = \int dx P_{\mathbf{v}}(x|\text{spike}) \log_2\left[\frac{P_{\mathbf{v}}(x|\text{spike})}{P_{\mathbf{v}}(x)}\right], \tag{2.5}$$

which is also the Kullback-Leibler divergence $D[P_{\mathbf{v}}(x|\text{spike})||P_{\mathbf{v}}(x)]$. Notice that this information is a function of the direction $\mathbf{v}$. The information $I(\mathbf{v})$ provides an invariant measure of how much the occurrence of a spike is determined by projection on the direction $\mathbf{v}$. It is a function only of direction in the stimulus space and does not change when vector $\mathbf{v}$ is multiplied by a constant. This can be seen by noting that for any probability distribution and any constant c, $P_{c\mathbf{v}}(x) = c^{-1}P_{\mathbf{v}}(x/c)$ (see also theorem 9.6.4 of Cover & Thomas, 1991). When evaluated along any vector $\mathbf{v}$, $I(\mathbf{v}) \le I_{\text{spike}}$. The total information I_{spike} can be recovered along one particular direction only if $\mathbf{v} = \hat{e}_1$ and only if the RS is one-dimensional.

By analogy with equation 2.5, one could also calculate information $I(\mathbf{v}_1, \ldots, \mathbf{v}_n)$ along a set of several directions $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ based on the multipoint probability distributions of projection values $x_1, x_2, \ldots, x_n$ along vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ of interest:

$$P_{\mathbf{v}_1,\ldots,\mathbf{v}_n}(\{x_i\}|\text{spike}) = \left\langle \prod_{i=1}^{n} \delta(x_i - \mathbf{s}\cdot\mathbf{v}_i)|\text{spike}\right\rangle_{\mathbf{s}}, \tag{2.6}$$

$$P_{\mathbf{v}_1,\ldots,\mathbf{v}_n}(\{x_i\}) = \left\langle \prod_{i=1}^{n} \delta(x_i - \mathbf{s}\cdot\mathbf{v}_i)\right\rangle_{\mathbf{s}}. \tag{2.7}$$

If we are successful in finding all of the directions $\{\hat{e}_i\}$, $1 \le i \le K$ contributing to the input-output relation, equation 1.1, then the information evaluated in this subspace will be equal to the total information I_{spike}. When we calculate information along a set of K vectors that are slightly off from the RS, the answer, of course, is smaller than I_{spike} and is initially quadratic in small deviations $\delta\mathbf{v}_i$. One can therefore hope to find the RS by maximizing information with respect to K vectors simultaneously. The information does

not increase if more vectors outside the RS are included. For uncorrelated stimuli, any vector or a set of vectors that maximizes $I(\mathbf{v})$ belongs to the RS. On the other hand, as discussed in appendix B, the result of optimization with respect to a number of vectors $k < K$ may deviate from the RS if stimuli are correlated. To find the RS, we first maximize $I(\mathbf{v})$ and compare this maximum with I_{spike}, which is estimated according to equation 2.2. If the difference exceeds that expected from finite sampling corrections, we increment the number of directions with respect to which information is simultaneously maximized.

3 Optimization Algorithm

In this section, we describe a particular algorithm we used to look for the most informative dimensions in order to find the relevant subspace. We make no claim that our choice of the algorithm is most efficient. However, it does give reproducible results for different starting points and spike trains with differences taken to simulate neural noise. Overall, choices for an algorithm are broader because the information $I(\mathbf{v})$ as defined by equation 2.5 is a continuous function, whose gradient can be computed. We find (see appendix C for a derivation)

$$\nabla_{\mathbf{v}} I = \int dx P_{\mathbf{v}}(x) \left[\langle \mathbf{s}|x, \text{spike} \rangle - \langle \mathbf{s}|x \rangle \right] \cdot \left[\frac{d}{dx} \frac{P_{\mathbf{v}}(x|\text{spike})}{P_{\mathbf{v}}(x)} \right], \tag{3.1}$$

where

$$\langle \mathbf{s}|x, \text{spike} \rangle = \frac{1}{P(x|\text{spike})} \int d\mathbf{s}\, \mathbf{s} \delta(x - \mathbf{s} \cdot \mathbf{v}) P(\mathbf{s}|\text{spike}), \tag{3.2}$$

and similarly for $\langle \mathbf{s}|x \rangle$. Since information does not change with the length of the vector, we have $\mathbf{v} \cdot \nabla_{\mathbf{v}} I = 0$, as also can be seen directly from equation 3.1.

As an optimization algorithm, we have used a combination of gradient ascent and simulated annealing algorithms: successive line maximizations were done along the direction of the gradient (Press, Teukolsky, Vetterling, & Flannery, 1992). During line maximizations, a point with a smaller value of information was accepted according to Boltzmann statistics, with probability $\propto \exp[(I(\mathbf{v}_{i+1}) - I(\mathbf{v}_i))/T]$. The effective temperature T is reduced by factor of $1 - \epsilon_T$ upon completion of each line maximization. Parameters of the simulated annealing algorithm to be adjusted are the starting temperature T_0 and the cooling rate ϵ_T, $\Delta T = -\epsilon_T T$. When maximizing with respect to one vector, we used values $T_0 = 1$ and $\epsilon_T = 0.05$. When maximizing with respect to two vectors, we either used the cooling schedule with $\epsilon_T = 0.005$ and repeated it several times (four times in our case) or allowed the effective temperature T to increase by a factor of 10 upon convergence to a local maximum (keeping $T \leq T_0$ always), while limiting the total number of line maximizations.

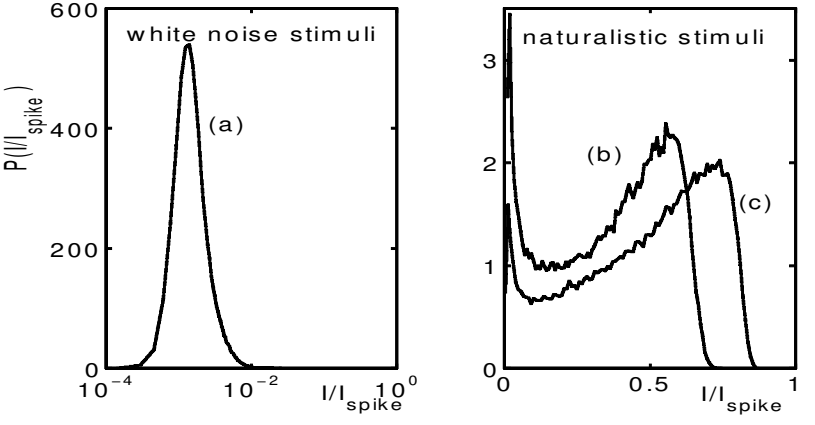

Figure 2: The probability distribution of information values in units of the total information per spike in the case of (a) uncorrelated binary noise stimuli, (b) correlated gaussian noise with power spectrum of natural scenes, and (c) stimuli derived from natural scenes (patches of photos). The distribution was obtained by calculating information along 10^5 random vectors for a model cell with one relevant dimension. Note the different scales in the two panels.

The problem of maximizing a function often is related to the problem of making a good initial guess. It turns out, however, that the choice of a starting point is much less crucial in cases where the stimuli are correlated. To illustrate this point, we plot in Figure 2 the probability distribution of information along random directions $\mathbf{v}$ for both white noise and naturalistic stimuli in a model with one relevant dimension. For uncorrelated stimuli, not only is information equal to zero for a vector that is perpendicular to the relevant subspace, but in addition, the derivative is equal to zero. Since a randomly chosen vector has on average a small projection on the relevant subspace (compared to its length) $v_r/|\mathbf{v}| \sim \sqrt{n/d}$, the corresponding information can be found by expanding in $v_r/|\mathbf{v}|$:

$$I \approx \frac{v_r^2}{2|\mathbf{v}|^2} \int dx P_{\hat{e}_{ir}}(x) \left(\frac{P'_{\hat{e}_{ir}}(x)}{P_{\hat{e}_{ir}}(x)} \right)^2 [\langle \mathbf{s}\hat{e}_r|\text{spike}\rangle - \langle \mathbf{s}\hat{e}_r\rangle]^2, \tag{3.3}$$

where vector $v = v_r\hat{e}_r + v_{ir}\hat{e}_{ir}$ is decomposed in its components inside and outside the RS, respectively. The average information for a random vector is therefore $\sim (\langle v_r^2\rangle/|\mathbf{v}|^2) = K/D$.

In cases where stimuli are drawn from a gaussian ensemble with correlations, an expression for the information values has a similar structure to equation 3.3. To see this, we transform to Fourier space and normalize each Fourier component by the square root of the power spectrum $S(\mathbf{k})$.

In this new basis, both the vectors $\{e_i\}$, $1 \le i \le K$, forming the RS and the randomly chosen vector $\mathbf{v}$ along which information is being evaluated, are to be multiplied by $\sqrt{S(\mathbf{k})}$. Thus, if we now substitute for the dot product v_r^2 the convolution weighted by the power spectrum, $\sum_i^K (\mathbf{v} * \hat{e}_i)^2$, where

$$\mathbf{v} * \hat{e}_i = \frac{\sum_{\mathbf{k}} v(\mathbf{k}) \hat{e}_i(\mathbf{k}) S(\mathbf{k})}{\sqrt{\sum_{\mathbf{k}} v^2(\mathbf{k}) S(\mathbf{k})} \sqrt{\sum_{\mathbf{k}} \hat{e}_i^2(\mathbf{k}) S(\mathbf{k})}}, \tag{3.4}$$

then equation 3.3 will describe information values along randomly chosen vectors $\mathbf{v}$ for correlated gaussian stimuli with the power spectrum $S(\mathbf{k})$. Although both v_r and $v(\mathbf{k})$ are gaussian variables with variance $\sim 1/D$, the weighted convolution has not only a much larger variance but is also strongly nongaussian (the nongaussian character is due to the normalizing factor $\sum_{\mathbf{k}} v^2(\mathbf{k}) S(\mathbf{k})$ in the denominator of equation 3.4). As for the variance, it can be estimated as $< (\mathbf{v} * \hat{e}_1)^2 >= 4\pi / \ln^2 D$, in cases where stimuli are taken as patches of correlated gaussian noise with the two-dimensional power spectrum $S(\mathbf{k}) = A/k^2$. The large values of the weighted dot product $\mathbf{v} * \hat{e}_i$, $1 \le i \le K$ result not only in significant information values along a randomly chosen vector, but also in large magnitudes of the derivative ∇I, which is no longer dominated by noise, contrary to the case of uncorrelated stimuli. In the end, we find that randomly choosing one of the presented frames as a starting guess is sufficient.

4 Results

We tested the scheme of looking for the most informative dimensions on model neurons that respond to stimuli derived from natural scenes and sounds. As visual stimuli, we used scans across natural scenes, which were taken as black and white photos digitized to 8 bits with no corrections made for the camera's light-intensity transformation function. Some statistical properties of the stimulus set are shown in Figure 3. Qualitatively, they reproduce the known results on the statistics of natural scenes (Ruderman & Bialek, 1994; Ruderman, 1994; Dong & Atick, 1995; Simoncelli & Olshausen, 2001). Most important properties for this study are strong spatial correlations, as evident from the power spectrum S(k) plotted in Figure 3b, and deviations of the probability distribution from a gaussian one. The nongaussian character can be seen in Figure 3c, where the probability distribution of intensities is shown, and in Figure 3d, which shows the distribution of projections on a Gabor filter (in what follows, the units of projections, such as s_1, will be given in units of the corresponding standard deviations). Our goal is to demonstrate that although the correlations present in the ensemble are nongaussian, they can be removed successfully from the estimate of vectors defining the RS.

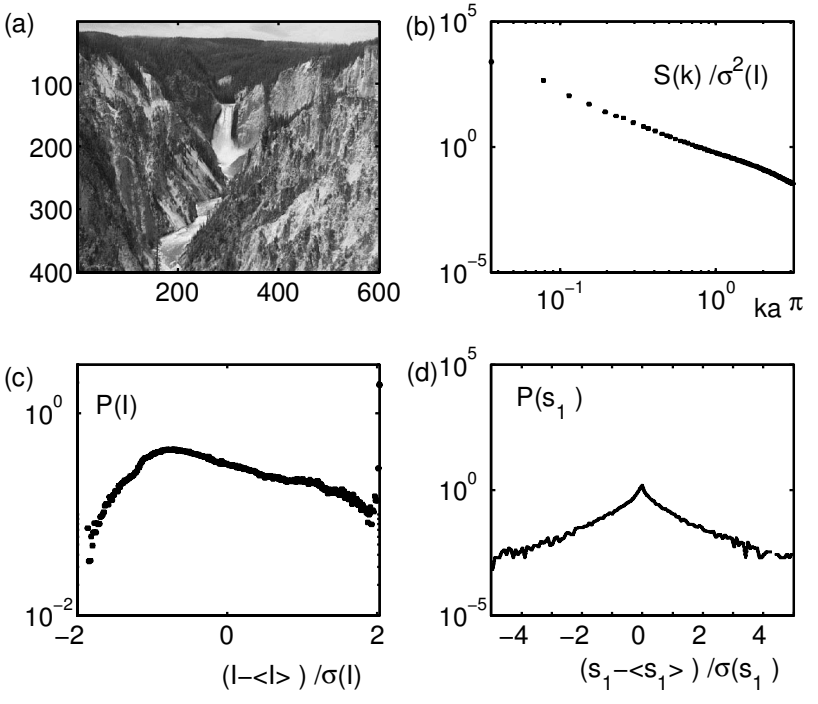

Figure 3: Statistical properties of the visual stimulus ensemble. (a) One of the photos. Stimuli would be 30×30 patches taken from the overall photograph. (b) The power spectrum, in units of light intensity variance $\sigma^2(I)$, averaged over orientation as a function of dimensionless wave vector ka, where a is the pixel size. (c) The probability distribution of light intensity in units of $\sigma(I)$. (d) The probability distribution of projections between stimuli and a Gabor filter, also in units of the corresponding standard deviation $\sigma(s_1)$.

4.1 A Model Simple Cell. Our first example is based on the properties of simple cells found in the primary visual cortex. A model phase- and orientation-sensitive cell has a single relevant dimension $\hat{e}_1$, shown in Figure 4a. A given stimulus $\mathbf{s}$ leads to a spike if the projection $s_1 = \mathbf{s} \cdot \hat{e}_1$ reaches a threshold value θ in the presence of noise:

$$\frac{P(\text{spike}|\mathbf{s})}{P(\text{spike})} \equiv f(s_1) = \langle H(s_1 - \theta + \xi) \rangle, \tag{4.1}$$

where a gaussian random variable ξ of variance σ^2 models additive noise, and the function $H(x) = 1$ for $x > 0$, and zero otherwise. Together with the RF $\hat{e}_1$, the parameters θ for threshold and the noise variance σ^2 determine the input-output function.

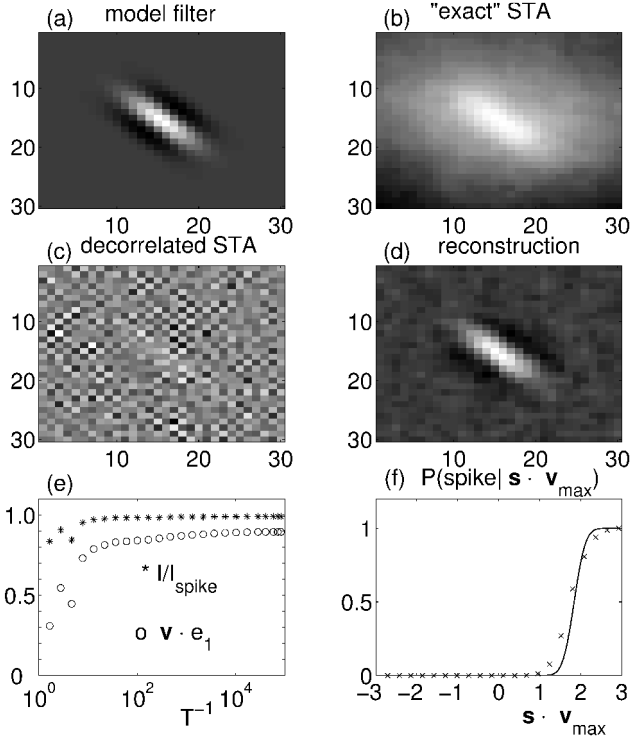

Figure 4: Analysis of a model simple cell with RF shown in (a). (b) The "exact" spike-triggered average $\mathbf{v}_{\rm sta}$. (c) An attempt to remove correlations according to the reverse correlation method, $C^{-1}_{a\,priori}\mathbf{v}_{\rm sta}$. (d) The normalized vector $\hat{v}_{\rm max}$ found by maximizing information. (e) Convergence of the algorithm according to information $I(\mathbf{v})$ and projection $\hat{v}\cdot\hat{e}_1$ between normalized vectors as a function of the inverse effective temperature T^{-1}. (f) The probability of a spike $P(\text{spike}|\mathbf{s}\cdot\hat{v}_{\rm max})$ (crosses) is compared to $P(\text{spike}|s_1)$ used in generating spikes (solid line). Parameters of the model are $\sigma = 0.31$ and $\theta = 1.84$, both given in units of standard deviation of s_1, which is also the units for the x-axis in f.

When the spike-triggered average (STA), or reverse correlation function, is computed from the responses to correlated stimuli, the resulting vector will be broadened due to spatial correlations present in the stimuli (see Figure 4b). For stimuli that are drawn from a gaussian probability distribution, the effects of correlations could be removed by multiplying $\mathbf{v}_{\rm sta}$ by the inverse of the a priori covariance matrix, according to the reverse correlation method, $\hat{v}_{Gaussian\,est} \propto C^{-1}_{a\,priori}\mathbf{v}_{\rm sta}$, equation A.2. However, this procedure tends to amplify noise. To separate errors due to neural noise from those

due to the nongaussian character of correlations, note that in a model, the effect of neural noise on our estimate of the STA can be eliminated by averaging the presented stimuli weighted with the exact firing rate, as opposed to using a histogram of responses to estimate $P(\text{spike}|\mathbf{s})$ from a finite set of trials. We have used this "exact" STA,

$$\mathbf{v}_{\text{sta}} = \int d\mathbf{s}\, \mathbf{s} P(\mathbf{s}|\text{spike}) = \frac{1}{P(\text{spike})} \int d\mathbf{s} P(\mathbf{s})\, \mathbf{s} P(\text{spike}|\mathbf{s}), \tag{4.2}$$

in calculations presented in Figures 4b and 4c. Even with this noiseless STA (the equivalent of collecting an infinite data set), the standard decorrelation procedure is not valid for nongaussian stimuli and nonlinear input-output functions, as discussed in detail in appendix A. The result of such a decorrelation in our example is shown in Figure 4c. It clearly is missing some of the structure in the model filter, with projection $\hat{e}_1 \cdot \hat{v}_{Gaussian\,est} \approx 0.14$. The discrepancy is not due to neural noise or finite sampling, since the "exact" STA was decorrelated; the absence of noise in the exact STA also means that there would be no justification for smoothing the results of the decorrelation. The discrepancy between the true RF and the decorrelated STA increases with the strength of nonlinearity in the input-output function.

In contrast, it is possible to obtain a good estimate of the relevant dimension $\hat{e}_1$ by maximizing information directly (see Figure 4d). A typical progress of the simulated annealing algorithm with decreasing temperature T is shown in Figure 4e. There we plot both the information along the vector and its projection on $\hat{e}_1$. We note that while information I remains almost constant, the value of projection continues to improve. Qualitatively, this is because the probability distributions depend exponentially on information. The final value of projection depends on the size of the data set, as discussed below. In the example shown in Figure 4, there were $\approx 50,000$ spikes with average probability of spike ≈ 0.05 per frame, and the reconstructed vector has a projection $\hat{v}_{\max} \cdot \hat{e}_1 = 0.920 \pm 0.006$. Having estimated the RF, one can proceed to sample the nonlinear input-output function. This is done by constructing histograms for $P(\mathbf{s} \cdot \hat{v}_{\max})$ and $P(\mathbf{s} \cdot \hat{v}_{\max}|\text{spike})$ of projections onto vector $\hat{v}_{\max}$ found by maximizing information and taking their ratio, as in equation 1.2. In Figure 4f, we compare $P(\text{spike}|\mathbf{s} \cdot \hat{v}_{\max})$ (crosses) with the probability $P(\text{spike}|s_1)$ used in the model (solid line).

4.2 Estimated Deviation from the Optimal Dimension. When information is calculated from a finite data set, the (normalized) vector $\hat{v}$, which maximizes I, will deviate from the true RF $\hat{e}_1$. The deviation $\delta\mathbf{v} = \hat{v} - \hat{e}_1$ arises because the probability distributions are estimated from experimental histograms and differ from the distributions found in the limit of infinite data size. For a simple cell, the quality of reconstruction can be characterized by the projection $\hat{v} \cdot \hat{e}_1 = 1 - \frac{1}{2}\delta\mathbf{v}^2$, where both $\hat{v}$ and $\hat{e}_1$ are normalized and $\delta\mathbf{v}$ is by definition orthogonal to $\hat{e}_1$. The deviation $\delta\mathbf{v} \sim A^{-1}\nabla I$, where A is

the Hessian of information. Its structure is similar to that of a covariance matrix:

$$A_{ij} = \frac{1}{\ln 2} \int dx P(x|\text{spike}) \left(\frac{d}{dx} \ln \frac{P(x|\text{spike})}{P(x)} \right)^2 \times (\langle s_i s_j | x \rangle - \langle s_i | x \rangle \langle s_j | x \rangle). \tag{4.3}$$

When averaged over possible outcomes of N trials, the gradient of information is zero for the optimal direction. Here, in order to evaluate $\langle \delta \mathbf{v}^2 \rangle = \mathrm{Tr}[A^{-1} \langle \nabla I \nabla I^T \rangle A^{-1}]$, we need to know the variance of the gradient of I. Assuming that the probability of generating a spike is independent for different bins, we can estimate $\langle \nabla I_i \nabla I_j \rangle \sim A_{ij}/(N_{\text{spike}} \ln 2)$. Therefore, an expected error in the reconstruction of the optimal filter is inversely proportional to the number of spikes. The corresponding expected value of the projection between the reconstructed vector and the relevant direction $\hat{e}_1$ is given by

$$\hat{v} \cdot \hat{e}_1 \approx 1 - \frac{1}{2} \langle \delta \mathbf{v}^2 \rangle = 1 - \frac{\mathrm{Tr}'[A^{-1}]}{2 N_{\text{spike}} \ln 2}, \tag{4.4}$$

where Tr′ means that the trace is taken in the subspace orthogonal to the model filter.[5] The estimate, equation 4.4, can be calculated without knowledge of the underlying model; it is $\sim D/(2N_{\text{spike}})$. This behavior should also hold in cases where the stimulus dimensions are expanded to include time. The errors are expected to increase in proportion to the increased dimensionality. In the case of a complex cell with two relevant dimensions, the error is expected to be twice that for a cell with single relevant dimension, also discussed in section 4.3.

We emphasize that the error estimate according to equation 4.4 is of the same order as errors of the reverse correlation method when it is applied for gaussian ensembles. The latter are given by $(\mathrm{Tr}[C^{-1}] - C_{11}^{-1})/[2N_{\text{spike}} \langle f'^2(s_1) \rangle]$. Of course, if the reverse correlation method were to be applied to the nongaussian ensemble, the errors would be larger. In Figure 5, we show the result of simulations for various numbers of trials, and therefore N_{spike}. The average projection of the normalized reconstructed vector $\hat{v}$ on the RF $\hat{e}_1$ behaves initially as $1/N_{\text{spike}}$ (dashed line). For smaller data sets, in this case, $N_{\text{spikes}} \lesssim 30,000$, corrections $\sim N_{\text{spikes}}^{-2}$ become important for estimating the expected errors of the algorithm. Happily, these corrections have a sign such that smaller data sets are more effective than one might have expected from the asymptotic calculation. This can be verified from the expansion $\hat{v} \cdot \hat{e}_1 = [1 - \delta \mathbf{v}^2]^{-1/2} \approx 1 - \frac{1}{2} \langle \delta \mathbf{v}^2 \rangle + \frac{3}{8} \langle \delta \mathbf{v}^4 \rangle$, were only the first two terms were taken into account in equation 4.4.

[5] By definition, $\delta v_1 = \delta \mathbf{v} \cdot \hat{e}_1 = 0$, and therefore $\langle \delta v_1^2 \rangle \propto A_{11}^{-1}$ is to be subtracted from $\langle \delta \mathbf{v}^2 \rangle \propto \mathrm{Tr}[A^{-1}]$. Because $\hat{e}_1$ is an eigenvector of A with zero eigenvalue, A_{11}^{-1} is infinite. Therefore, the proper treatment is to take the trace in the subspace orthogonal to $\hat{e}_1$.

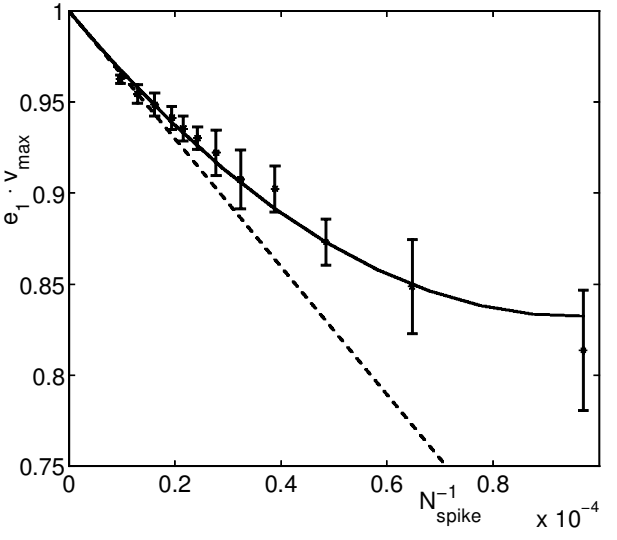

Figure 5: Projection of vector $\hat{v}_{\max}$ that maximizes information on RF $\hat{e}_1$ is plotted as a function of the number of spikes. The solid line is a quadratic fit in $1/N_{\rm spike}$, and the dashed line is the leading linear term in $1/N_{\rm spike}$. This set of simulations was carried out for a model visual neuron with one relevant dimension from Figure 4a and the input-output function (see equation 4.1), with parameter values $\sigma \approx 0.61\sigma(s_1), \theta \approx 0.61\sigma(s_1)$. For this model neuron, the linear approximation for the expected error is applicable for $N_{\rm spike} \gtrsim 30,000$.

4.3 A Model Complex Cell. A sequence of spikes from a model cell with two relevant dimensions was simulated by projecting each of the stimuli on vectors that differ by $\pi/2$ in their spatial phase, taken to mimic properties of complex cells, as in Figure 6. A particular frame leads to a spike according to a logical OR, that is, if either $|s_1|$ or $|s_2|$ exceeds a threshold value θ in the presence of noise, where $s_1 = \mathbf{s} \cdot \hat{e}_1$, $s_2 = \mathbf{s} \cdot \hat{e}_2$. Similarly to equation 4.1,

$$\frac{P(\text{spike}|\mathbf{s})}{P(\text{spike})} = f(s_1, s_2) = \langle H(|s_1| - \theta - \xi_1) \ \vee \ H(|s_2| - \theta - \xi_2) \rangle, \qquad (4.5)$$

where ξ_1 and ξ_2 are independent gaussian variables. The sampling of this input-output function by our particular set of natural stimuli is shown in Figure 6c. As is well known, reverse correlation fails in this case because the spike–triggered average stimulus is zero, although with gaussian stimuli, the spike-triggered covariance method would recover the relevant dimensions (Touryan et al., 2002). Here we show that searching for maximally informative dimensions allows us to recover the relevant subspace even under more natural stimulus conditions.

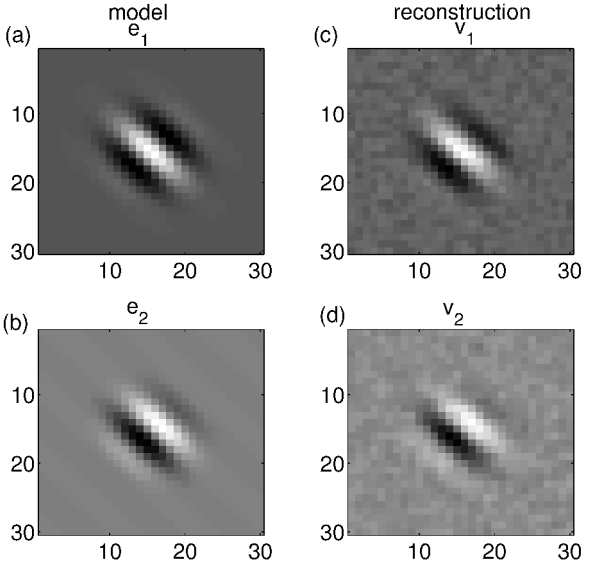

Figure 6: Analysis of a model complex cell with relevant dimensions $\hat{e}_1$ and $\hat{e}_2$ shown in (a) and (b), respectively. Spikes are generated according to an "OR" input-output function $f(s_1, s_2)$ with the threshold $\theta \approx 0.61\sigma(s_1)$ and noise standard deviation $\sigma = 0.31\sigma(s_1)$. (c,d) Vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ found by maximizing information $I(\mathbf{v}_1, \mathbf{v}_2)$.

We start by maximizing information with respect to one vector. Contrary to the result in Figure 4e for a simple cell, one optimal dimension recovers only about 60% of the total information per spike (see equation 2.2). Perhaps surprisingly, because of the strong correlations in natural scenes, even a projection onto a random vector in the $D \sim 10^3$-dimensional stimulus space has a high probability of explaining 60% of total information per spike, as can be seen in Figure 2. We therefore go on to maximize information with respect to two vectors. As a result of maximization, we obtain two vectors $\mathbf{v}_1$ and $\mathbf{v}_2$, shown in Figure 6. The information along them is $I(\mathbf{v}_1, \mathbf{v}_2) \approx 0.90$, which is within the range of information values obtained along different linear combinations of the two model vectors $I(\hat{e}_1, \hat{e}_2)/I_{\text{spike}} = 0.89 \pm 0.11$. Therefore, the description of neuron's firing in terms of vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ is complete up to the noise level, and we do not have to look for extra relevant dimensions. Practically, the number of relevant dimensions can be determined by comparing $I(\mathbf{v}_1, \mathbf{v}_2)$ to either I_{spike} or $I(\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3)$, the latter being the result of maximization with respect to three vectors simultaneously. As mentioned in section 1, information along set a of vectors does not increase when extra dimensions are added to the relevant subspace. Therefore, if $I(\mathbf{v}_1, \mathbf{v}_2) = I(\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3)$ (again, up to the noise level), this means that

there are only two relevant dimensions. Using $I_{\rm spike}$ for comparison with $I(\mathbf{v}_1, \mathbf{v}_2)$ has the advantage of not having to look for an extra dimension, which can be computationally intensive. However, $I_{\rm spike}$ might be subject to larger systematic bias errors than $I(\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3)$.

Vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ obtained by maximizing $I(\mathbf{v}_1, \mathbf{v}_2)$ are not exactly orthogonal and are also rotated with respect to $\hat{e}_1$ and $\hat{e}_2$. However, the quality of reconstruction, as well as the value of information $I(\mathbf{v}_1, \mathbf{v}_2)$, is independent of a particular choice of basis with the RS. The appropriate measure of similarity between the two planes is the dot product of their normals. In the example of Figure 6, $\hat{n}_{(\hat{e}_1,\hat{e}_2)} \cdot \hat{n}_{(\mathbf{v}_1,\mathbf{v}_2)} = 0.82 \pm 0.07$, where $\hat{n}_{(\hat{e}_1,\hat{e}_2)}$ is a normal to the plane passing through vectors $\hat{e}_1$ and $\hat{e}_2$. Maximizing information with respect to two dimensions requires a significantly slower cooling rate and, consequently, longer computational times. However, the expected error in the reconstruction, $1 - \hat{n}_{(\hat{e}_1,\hat{e}_2)} \cdot \hat{n}_{(\mathbf{v}_1,\mathbf{v}_2)}$, scales as $1/N_{\rm spike}$ behavior, similar to equation 4.4, and is roughly twice that for a simple cell given the same number of spikes. We make vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ orthogonal to each others upon completion of the algorithm.

4.4 A Model Auditory Neuron with One Relevant Dimension. Because stimuli $\mathbf{s}$ are treated as vectors in an abstract space, the method of looking for the most informative dimensions can be applied equally well to auditory as well as to visual neurons. Here we illustrate the method by considering a model auditory neuron with one relevant dimension, which is shown in Figure 7c and is taken to mimic the properties of cochlear neurons. The model neuron is probed by two ensembles of naturalistic stimuli: one is a recording of a native Russian speaker reading a piece of Russian prose, and the other is a recording of a piece of English prose read by a native English speaker. Both ensembles are nongaussian and exhibit amplitude distributions with long, nearly exponential tails (see Figure 7a) which are qualitatively similar to those of light intensities in natural scenes (Voss & Clarke, 1975; Ruderman, 1994). However, the power spectrum is different in the two cases, as can be seen in Figure 7b. The differences in the correlation structure in particular lead to different STAs across the two ensembles (cf. Figure 7d). Both of the STAs also deviate from the model filter shown in Figure 7c.

Despite differences in the probability distributions $P(\mathbf{s})$, it is possible to recover the relevant dimension of the model neuron by maximizing information. In Figure 7c we show the two most informative vectors found by running the algorithm for the two ensembles and replot the model filter from Figure 7c to show that the three vectors overlap almost perfectly. Thus, different nongaussian correlations can be successfully removed to obtain an estimate of the relevant dimension. If the most informative vector changes with the stimulus ensemble, this can be interpreted as caused by adaptation to the probability distribution.

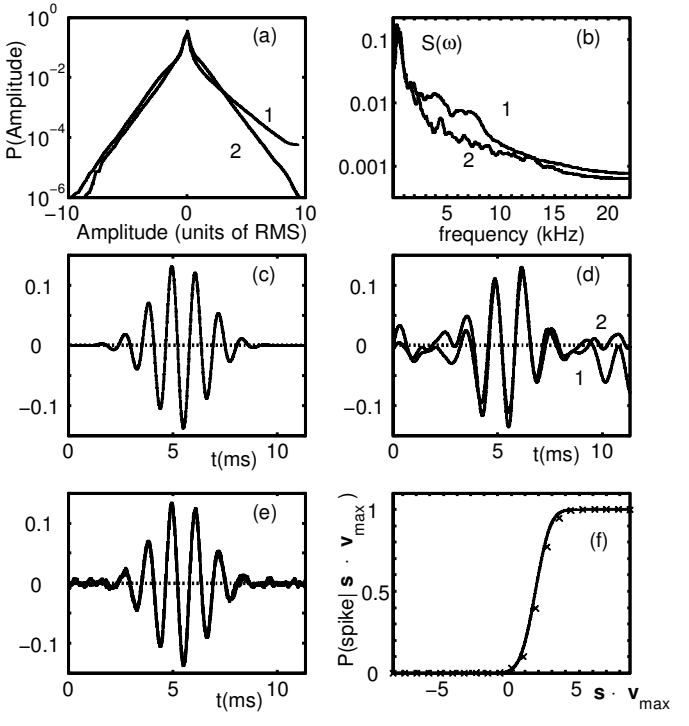

Figure 7: A model auditory neuron is probed by two natural ensembles of stimuli: a piece of English prose (1) and a piece of of Russian prose (2) . The size of the stimulus ensemble was the same in both cases, and the sampling rate was 44.1 kHz. (a) The probability distribution of the sound pressure amplitude in units of standard deviation for both ensembles is strongly nongaussian. (b) The power spectra for the two ensembles. (c) The relevant vector of the model neuron, of dimensionality $D = 500$. (d) The STA is broadened in both cases, but differs between the two cases due to differences in the power spectra of the two ensembles. (e) Vectors that maximize information for either of the ensembles overlap almost perfectly with each other and with the model filter, which is also replotted here from *c*. (f) The probability of a spike $P(\text{spike}|\mathbf{s}\cdot\hat{v}_{\max})$ (crosses) is compared to $P(\text{spike}|s_1)$ used in generating spikes (solid line). The input-output function had parameter values $\sigma \approx 0.9\sigma(s_1)$ and $\theta \approx 1.8\sigma(s_1)$.

5 Summary

Features of the stimulus that are most relevant for generating the response of a neuron can be found by maximizing information between the sequence of responses and the projection of stimuli on trial vectors within the stimulus space. Calculated in this manner, information becomes a function of

direction in stimulus space. Those vectors that maximize the information and account for the total information per response of interest span the relevant subspace. The method allows multiple dimensions to be found. The reconstruction of the relevant subspace is done without assuming a particular form of the input-output function. It can be strongly nonlinear within the relevant subspace and is estimated from experimental histograms for each trial direction independently. Most important, this method can be used with any stimulus ensemble, even those that are strongly nongaussian, as in the case of natural signals. We have illustrated the method on model neurons responding to natural scenes and sounds. We expect the current implementation of the method to be most useful when several most informative vectors (≤ 10, depending on their dimensionality) are to be analyzed for neurons probed by natural scenes. This technique could be particularly useful in describing sensory processing in poorly understood regions of higher-level sensory cortex (such as visual areas V2, V4, and IT and auditory cortex beyond A1) where white noise stimulation is known to be less effective than naturalistic stimuli.

Appendix A: Limitations of the Reverse Correlation Method

Here we examine what sort of deviations one can expect when applying the reverse correlation method to natural stimuli even in the model with just one relevant dimension. There are two factors that, when combined, invalidate the reverse correlation method: the nongaussian character of correlations and the nonlinearity of the input-output function (Ringach, Sapiro, & Shapley, 1997). In its original formulation (de Boer & Kuyper, 1968), the neuron is probed by white noise, and the relevant dimension $\hat{e}_1$ is given by the STA $\hat{e}_1 \propto \langle \mathbf{s}r(\mathbf{s})\rangle$. If the signals are not white, that is, the covariance matrix $C_{ij} = \langle s_i s_j\rangle$ is not a unit matrix, then the STA is a broadened version of the original filter $\hat{e}_1$. This can be seen by noting that for any function $F(\mathbf{s})$ of gaussian variables $\{s_i\}$, the identity holds:

$$\langle s_i F(\mathbf{s})\rangle = \langle s_i s_j\rangle \langle \partial_{s_j} F(\mathbf{s})\rangle, \quad \partial_j \equiv \partial_{s_j}. \tag{A.1}$$

When property A.1 is applied to the vector components of the STA, $\langle s_i r(\mathbf{s})\rangle = C_{ij}\langle \partial_j r(\mathbf{s})\rangle$. Since we work within the assumption that the firing rate is a (nonlinear) function of projection onto one filter $\hat{e}_1$, $r(\mathbf{s}) = r(s_1)$, the latter average is proportional to the model filter itself, $\langle \partial_j r\rangle = \hat{e}_{1j}\langle r'(s_1)\rangle$. Therefore, we arrive at the prescription of the reverse correlation method,

$$\hat{e}_{1i} \propto [C^{-1}]_{ij}\langle s_j r(\mathbf{s})\rangle. \tag{A.2}$$

The gaussian property is necessary in order to represent the STA as a convolution of the covariance matrix C_{ij} of the stimulus ensemble and the model

filter. To understand how the reconstructed vector obtained according to equation A.2 deviates from the relevant one, we consider weakly nongaussian stimuli, with the probability distribution

$$P_{nG}(\mathbf{s}) = \frac{1}{Z} P_0(\mathbf{s}) e^{\epsilon H_1(\mathbf{s})}, \tag{A.3}$$

where $P_0(\mathbf{s})$ is the gaussian probability distribution with covariance matrix C and the normalization factor $Z = \langle e^{\epsilon H_1(\mathbf{s})} \rangle$. The function H_1 describes deviations of the probability distribution from gaussian, and therefore we will set $\langle s_i H_1 \rangle = 0$ and $\langle s_i s_j H_1 \rangle = 0$, since these averages can be accounted for in the gaussian ensemble. In what follows, we will keep only the first-order terms in perturbation parameter ϵ. Using property A.1, we find the STA to be given by

$$\langle s_i r \rangle_{nG} = \langle s_i s_j \rangle \left[\langle \partial_j r \rangle + \epsilon \langle r \partial_j (H_1) \rangle \right], \tag{A.4}$$

where averages are taken with respect to the gaussian distribution. Similarly, the covariance matrix C_{ij} evaluated with respect to the nongaussian ensemble is given by

$$C_{ij} = \frac{1}{Z} \langle s_i s_j e^{\epsilon H_1} \rangle = \langle s_i s_j \rangle + \epsilon \langle s_i s_k \rangle \langle s_j \partial_k (H_1) \rangle, \tag{A.5}$$

so that to the first order in ϵ, $\langle s_i s_j \rangle = C_{ij} - \epsilon C_{ik} \langle s_j \partial_k (H_1) \rangle$. Combining this with equation A.4, we get

$$\langle s_i r \rangle_{nG} = \text{const} \times C_{ij} \hat{e}_{1j} + \epsilon C_{ij} \langle \left(r - s_1 \langle r' \rangle \right) \partial_j (H_1) \rangle. \tag{A.6}$$

The second term in equation A.6 prevents the application of the reverse correlation method for nongaussian signals. Indeed, if we multiply the STA, equation A.6, with the inverse of the a priori covariance matrix C_{ij} according to the reverse correlation method, equation A.2, we no longer obtain the RF $\hat{e}_1$. The deviation of the obtained answer from the true RF increases with ϵ, which measures the deviation of the probability distribution from gaussian. Since natural stimuli are known to be strongly nongaussian, this makes the use of the reverse correlation problematic when analyzing neural responses to natural stimuli.

The difference in applying the reverse correlation to stimuli drawn from a correlated gaussian ensemble versus a nongaussian one is illustrated in Figures 8b and 8c. In the first case, shown in Figure 8b, stimuli are drawn from a correlated gaussian ensemble with the covariance matrix equal to that of natural images. In the second case, shown in Figure 8c, the patches of photos are taken as stimuli. The STA is broadened in both cases. Although the two-point correlations are just as strong in the case of gaussian stimuli

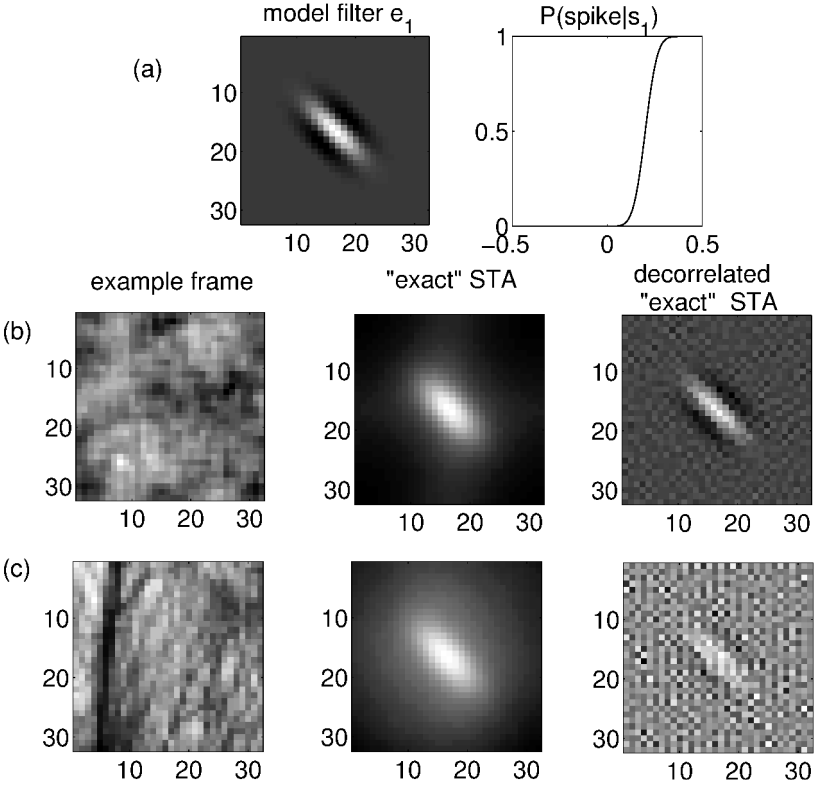

Figure 8: The nongaussian character of correlations present in natural scenes invalidates the reverse correlation method for neurons with a nonlinear input-output function. (a) A model visual neuron has one relevant dimension $\hat{e}_1$ and the nonlinear input-output function. The "exact" STA is used (see equation 4.2) to separate effects of neural noise from alterations introduced by the method. The decorrelated "exact" STA is obtained by multiplying the "exact" STA by the inverse of the covariance matrix, according to equation A.2. (b) Stimuli are taken from a correlated gaussian noise ensemble. The effect of correlations in STA can be removed according to equation A.2. (c) When patches of photos are taken as stimuli for the same model neuron as in *b*, the decorrelation procedure gives an altered version of the model filter. The two stimulus ensembles have the same covariance matrix.

as they are in the natural stimuli ensemble, gaussian correlations can be successfully removed from the STA according to equation A.2 to obtain the model filter. On the contrary, an attempt to use reverse correlation with natural stimuli results in an altered version of the model filter. We reiterate that for this example, the apparent noise in the decorrelated vector is not

due to neural noise or finite data sets, since the "exact" STA has been used (see equation 4.2) in all calculations presented in Figures 8 and 9.

The reverse correlation method gives the correct answer for any distribution of signals if the probability of generating a spike is a linear function of s_i, since then the second term in equation A.6 is zero. In particular, a linear input-output relation could arise due to a neural noise whose variance is much larger than the variance of the signal itself. This point is illustrated in Figures 9a, 9b, and 9c, where the reverse correlation method is applied to a threshold input-output function at low, moderate, and high signal-to-noise ratios. For small signal-to-noise ratios where the noise standard deviation is similar to that of projections s_1, the threshold nonlinearity in the input-output function is masked by noise and is effectively linear. In this limit, the reverse correlation can be applied with the exact STA. However, for experimentally calculated STA at low signal-to-noise ratios, the decorrelation procedure results in strong noise amplification. At higher signal-to-noise ratios, decorrelation fails due to the nonlinearity of the input-output function in accordance with equation A.6.

Appendix B: Maxima of $I(v)$: What Do They Mean?

The relevant subspace of dimensionality K can be found by maximizing information simultaneously with respect to K vectors. The result of maximization with respect to a number of vectors that is less than the true dimensionality of the relevant subspace may produce vectors that have components in the irrelevant subspace. This happens only in the presence of correlations in stimuli. As an illustration, we consider the situation where the dimensionality of the relevant subspace $K = 2$, and vector $\hat{e}_1$ describes the most informative direction within the relative subspace. We show here that although the gradient of information is perpendicular to both $\hat{e}_1$ and $\hat{e}_2$, it may have components outside the relevant subspace. Therefore, the vector $v_{\max}$ that corresponds to the maximum of $I(v)$ will lie outside the relevant subspace. We recall from equation 3.1 that

$$\nabla I(\hat{e}_1) = \int ds_1 P(s_1) \frac{d}{ds_1} \frac{P(s_1|\text{spike})}{P(s_1)} (\langle \mathbf{s}|s_1, \text{spike}\rangle - \langle \mathbf{s}|s_1\rangle), \tag{B.1}$$

We can rewrite the conditional averages $\langle \mathbf{s}|s_1\rangle = \int ds_2 P(s_1, s_2)\langle \mathbf{s}|s_1, s_2\rangle / P(s_1)$ and $\langle \mathbf{s}|s_1, \text{spike}\rangle = \int ds_2 f(s_1, s_2) P(s_1, s_2)\langle \mathbf{s}|s_1, s_2\rangle / P(s_1|\text{spike})$, so that

$$\nabla I(\hat{e}_1) = \int ds_1 ds_2 P(s_1, s_2)\langle \mathbf{s}|s_1, s_2\rangle \frac{P(\text{spike}|s_1, s_2) - P(\text{spike}|s_1)}{P(\text{spike})} \times \frac{d}{ds_1} \ln \frac{P(s_1|\text{spike})}{P(s_1)}. \tag{B.2}$$

Because we assume that the vector $\hat{e}_1$ is the most informative within the relevant subspace, $\hat{e}_1 \nabla I = \hat{e}_2 \nabla I = 0$, so that the integral in equation B.2

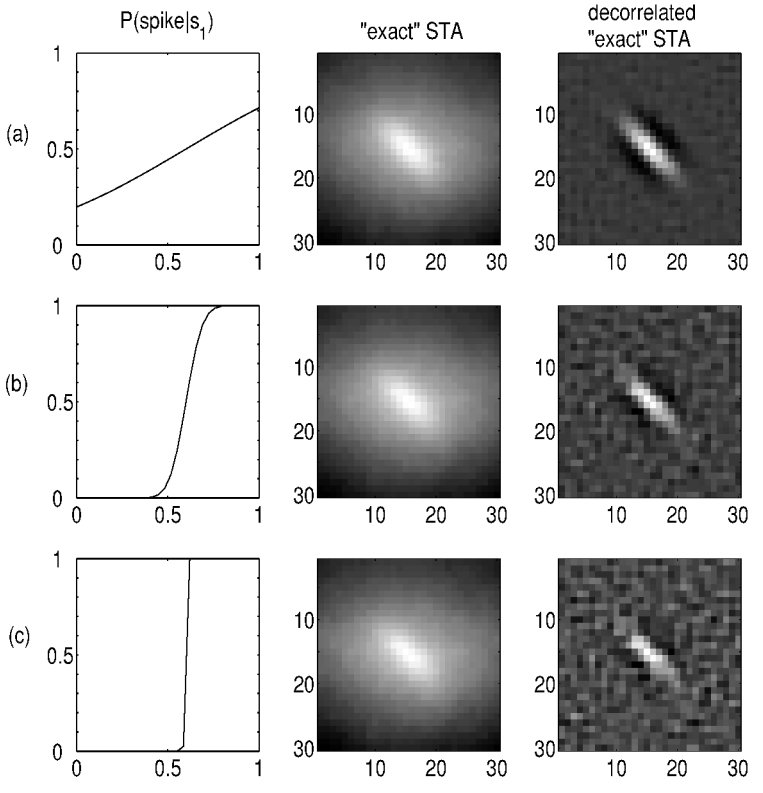

Figure 9: Application of the reverse correlation method to a model visual neuron with one relevant dimension $\hat{e}_1$ and a threshold input-output function of decreasing values of noise variance $\sigma/\sigma(s_1)s \approx 6.1,\ 0.61,\ 0.06$ in a, b, and c, respectively. The model $P(\text{spike}|s_1)$ becomes effectively linear when the signal-to-noise ratio is small. The reverse correlation can be used together with natural stimuli if the input-output function is linear. Otherwise, the deviations between the decorrelated STA and the model filter increase with the nonlinearity of $P(\text{spike}|s_1)$.

is zero for those directions in which the component of the vector $\langle \mathbf{s}|s_1, s_2\rangle$ changes linearly with s_1 and s_2. For uncorrelated stimuli, this is true for all directions, so that the most informative vector within the relevant subspace is also the most informative in the overall stimulus space. In the presence of correlations, the gradient may have nonzero components along some irrelevant directions if projection of the vector $\langle \mathbf{s}|s_1, s_2\rangle$ on those directions is not a linear function of s_1 and s_2. By looking for a maximum of information, we will therefore be driven outside the relevant subspace. The deviation of $v_{\max}$ from the relevant subspace is also proportional to the

strength of the dependence on the second parameter s_2 because of the factor $[P(s_1, s_2|\text{spike})/P(s_1, s_2) - P(s_1|\text{spike})/P(s_1)]$ in the integrand.

Appendix C: The Gradient of Information

According to expression 2.5, the information $I(\mathbf{v})$ depends on the vector $\mathbf{v}$ only through the probability distributions $P_{\mathbf{v}}(x)$ and $P_{\mathbf{v}}(x|\text{spike})$. Therefore, we can express the gradient of information in terms of gradients of those probability distributions:

$$\nabla_{\mathbf{v}} I = \frac{1}{\ln 2} \int dx \left[\ln \frac{P_{\mathbf{v}}(x|\text{spike})}{P_{\mathbf{v}}(x)} \nabla_{\mathbf{v}} (P_{\mathbf{v}}(x|\text{spike})) - \frac{P_{\mathbf{v}}(x|\text{spike})}{P_{\mathbf{v}}(x)} \nabla_{\mathbf{v}} (P_{\mathbf{v}}(x)) \right], \tag{C.1}$$

where we took into account that $\int dx P_{\mathbf{v}}(x|\text{spike}) = 1$ and does not change with $\mathbf{v}$. To find gradients of the probability distributions, we note that

$$\begin{aligned} \nabla_{\mathbf{v}} P_{\mathbf{v}}(x) &= \nabla_{\mathbf{v}} \left[\int d\mathbf{s} P(\mathbf{s}) \delta(x - \mathbf{s} \cdot \mathbf{v}) \right] = - \int d\mathbf{s} P(\mathbf{s}) \mathbf{s} \delta'(x - \mathbf{s} \cdot \mathbf{v}) \\ &= -\frac{d}{dx} \left[p(x) \langle \mathbf{s}|x \rangle \right], \end{aligned} \tag{C.2}$$

and analogously for $P_{\mathbf{v}}(x|\text{spike})$:

$$\nabla_{\mathbf{v}} P_{\mathbf{v}}(x|\text{spike}) = -\frac{d}{dx} \left[p(x|\text{spike}) \langle \mathbf{s}|x, \text{spike} \rangle \right]. \tag{C.3}$$

Substituting expressions C.2 and C.3 into equation C.1 and integrating once by parts, we obtain

$$\nabla_{\mathbf{v}} I = \int dx P_{\mathbf{v}}(x) \left[\langle \mathbf{s}|x, \text{spike} \rangle - \langle \mathbf{s}|x \rangle \right] \cdot \left[\frac{d}{dx} \frac{P_{\mathbf{v}}(x|\text{spike})}{P_{\mathbf{v}}(x)} \right],$$

which is expression 3.1 of the main text.

Acknowledgments

We thank K.D. Miller for many helpful discussions. Work at UCSF was supported in part by the Sloan and Swartz foundations and by a training grant from the NIH. Our collaboration began at the Marine Biological Laboratory in a course supported by grants from NIMH and the Howard Hughes Medical Institute.

References

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

Baddeley, R., Abbott, L. F., Booth, M.C.A., Sengpiel, F., Freeman, T., Wakeman, E. A., & Rolls, E. T. (1997). Responses of neurons in primary and inferior temporal visual cortices to natural scenes. *Proc. R. Soc. Lond. B, 264,* 1775–1783.

Barlow, H. (1961). Possible principles underlying the transformations of sensory images. In W. Rosenblith (Ed.), *Sensory communication* (pp. 217–234). Cambridge, MA: MIT Press.

Barlow, H. (2001). Redundancy reduction revisited. *Network: Comput. Neural Syst., 12,* 241–253.

Bialek, W. (2002). Thinking about the brain. In H. Flyvbjerg, F. Jülicher, P. Ormos, & F. David (Eds.), *Physics of Biomolecules and Cells* (pp. 485–577). Berlin: Springer-Verlag. See also physics/0205030.[6]

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

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

Brenner, N., Strong, S. P., Koberle, R., Bialek, W., & de Ruyter van Steveninck, R. R. (2000). Synergy in a neural code. *Neural Computation, 12,* 1531–1552. See also physics/9902067.

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

Creutzfeldt, O. D., & Northdurft H. C. (1978). Representation of complex visual stimuli in the brain. *Naturwissenschaften, 65,* 307–318.

Cover, T. M., & Thomas, J. A. (1991). *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 neuron in the blowfly visual system: Coding and information transfer in short spike sequences. *Proc. R. Soc. Lond. B, 265,* 259–265.

de Ruyter van Steveninck, R. R., Borst, A., & Bialek, W. (2000). Real time encoding of motion: Answerable questions and questionable answers from the fly's visual system. In J. M. Zanker, & J. Zeil (Eds.), *Motion vision: Computational, neural and ecological constraints* (pp. 279–306). New York: Springer-Verlag. See also physics/0004060.

de Ruyter van Steveninck, R. R., Lewen, G. D., Strong, S. P., Koberle, R., & Bialek, W. (1997). Reproducibility and variability in neural spike trains. *Science, 275,* 1805–1808. See also cond-mat/9603127.

[6] Where available we give references to the physics e–print archive, which may be found on-line at http://arxiv.org/abs/*/*; thus, Bialek (2002) is available on-line at http://arxiv.org/abs/physics/0205030. Published papers may differ from the versions posted to the archive.

Dimitrov, A. G., & Miller, J. P. (2001). Neural coding and decoding: Communication channels and quantization. *Network: Comput. Neural Syst., 12*, 441–472.
Dong, D. W., & Atick, J. J. (1995). Statistics of natural time-varying images. *Network: Comput. Neural Syst., 6*, 345–358.
Fairhall, A. L., Lewen, G. D., Bialek, W., & de Ruyter van Steveninck, R. R. (2001). Efficiency and ambiguity in an adaptive neural code. *Nature, 412*, 787–792.
Kara, P., Reinagel, P., & Reid, R. C. (2000). Low response variability in simultaneously recorded retinal, thalamic, and cortical neurons. *Neuron, 27*, 635–646.
Lewen, G. D., Bialek, W., & de Ruyter van Steveninck, R. R. (2001). Neural coding of naturalistic motion stimuli. *Network: Comput. Neural Syst., 12*, 317–329. See also physics/0103088.
Mainen, Z. F., & Sejnowski, T. J. (1995). Reliability of spike timing in neocortical neurons. *Science, 268*, 1503–1506.
Paninski, L. (2003a). Convergence properties of three spike–triggered analysis techniques. *Network: Compt. in Neural Systems, 14*, 437–464.
Paninski, L. (2003b). Estimation of entropy and mutual information. *Neural Computation, 15*, 1191–1253.
Panzeri, S., & Treves, A. (1996). Analytical estimates of limited sampling biases in different information measures. *Network: Comput. Neural Syst., 7*, 87–107.
Pola, G., Schultz, S. R., Petersen, R., & Panzeri, S. (2002). A practical guide to information analysis of spike trains. In R. Kotter (Ed.), *Neuroscience databases: A practical guide*, (pp. 137–152). Norwell, MA: Kluwer.
Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (1992). *Numerical recipes in C: The art of scientific computing*. Cambridge: Cambridge University Press.
Reinagel, P., & Reid, R. C. (2000). Temporal coding of visual information in the thalamus. *J. Neurosci., 20*, 5392–5400.
Rieke, F., Bodnar, D. A., & Bialek, W. (1995). Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory afferents. *Proc. R. Soc. Lond. B, 262*, 259–265.
Rieke, F., Warland, D., de Ruyter van Steveninck, R. R., & Bialek, W. (1997). *Spikes: Exploring the neural code*. Cambridge, MA: MIT Press.
Ringach, D. L., Hawken, M. J., & Shapley, R. (2002). Receptive field structure of neurons in monkey visual cortex revealed by stimulation with natural image sequences. *Journal of Vision, 2*, 12–24.
Ringach, D. L., Sapiro, G., & Shapley, R. (1997). A subspace reverse-correlation technique for the study of visual neurons. *Vision Res., 37*, 2455–2464.
Rolls, E. T., Aggelopoulos, N. C., & Zheng, F. (2003). The receptive fields of inferior temporal cortex neurons in natural scenes. *J. Neurosci., 23*, 339–348.
Ruderman, D. L. (1994). The statistics of natural images. *Network: Compt. Neural Syst., 5*, 517–548.
Ruderman, D. L., & Bialek, W. (1994). Statistics of natural images: Scaling in the woods. *Phys. Rev. Lett., 73*, 814–817.
Schwartz, O., Chichilnisky, E. J., & Simoncelli, E. (2002). Characterizing neural gain control using spike-triggered covariance. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in Neural Information Processing* (pp. 269–276). Cambridge, MA: MIT Press.

Sen, K., Theunissen, F. E., & Doupe, A. J. (2001). Feature analysis of natural sounds in the songbird auditory forebrain. *J. Neurophysiol., 86,* 1445–1458.

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

Smirnakis, S. M., Berry, M. J., Warland, D. K., Bialek, W., & Meister, M. (1996). Adaptation of retinal processing to image contrast and spatial scale. *Nature, 386,* 69–73.

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

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

Strong, S. P., Koberle, R., de Ruyter van Steveninck, R. R., & Bialek, W. (1998). Entropy and information in neural spike trains. *Phys. Rev. Lett., 80,* 197–200.

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

Tishby, N., Pereira, F. C., & Bialek, W. (1999). The information bottleneck method. In B. Hajek & R. S. Sreenivas (Eds.), *Proceedings of the 37th Allerton Conference on Communication, Control and Computing* (pp. 368–377). Urbana: University of Illinois. See also physics/0004057.

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

Treves, A., & Panzeri, S. (1995). The upward bias in measures of information derived from limited data samples. *Neural Comp., 7,* 399–407.

von der Twer, T., & Macleod, D. I. A. (2001). Optimal nonlinear codes for the perception of natural colours. *Network: Comput. Neural Syst., 12,* 395–407.

Vickers, N. J., Christensen, T. A., Baker, T., & Hildebrand, J. G. (2001). Odour-plume dynamics influence the brain's olfactory code. *Nature, 410,* 466–470.

Vinje, W. E., & Gallant, J. L. (2000). Sparse coding and decorrelation in primary visual cortex during natural vision. *Science, 287,* 1273–1276.

Vinje, W. E., & Gallant, J. L. (2002). Natural stimulation of the nonclassical receptive field increases information transmission efficiency in V1. *J. Neurosci., 22,* 2904–2915.

Voss, R. F., & Clarke, J. (1975). "1/f noise" in music and speech. *Nature,* 317–318.

Weliky, M., Fiser, J., Hunt, R., & Wagner, D. N. (2003). Coding of natural scenes in primary visual cortex. *Neuron, 37,* 703–718.

Received January 21, 2003; accepted August 6, 2003.