published in: R. Rao, B. Olshausen and M. Lewicki (eds.)
Probabilistic Models of the Brain. MIT Press, 2002.

A Probabilistic Network Model of Population Responses

Richard S. Zemel
Department of Computer Science
University of Toronto
Toronto, ON M5S 1A4
zemel@cs.toronto.edu

Jonathan Pillow
Center for Neural Science
New York University
New York, NY 10013
pillow@cns.nyu.edu

Abstract

A central question in computational neuroscience concerns how the response properties of neural populations depend on the activity of neurons both within and outside the population. Various models contain different types and combinations of feedforward and recurrent connections. Each model can be characterized by the particular set of assumptions about what information underlies the population response, and these in turn are reflected in the model's behavior. We propose that the population response is designed to preserve full information about the relevant dimension in the stimulus, which could be a single unambiguous value, a single ambiguous value, or more than one value. We design an objective based on preserving this information, and use it in training the weights in a model. Our results demonstrate that a combination of feedforward and recurrent connections can generate broadly information-preserving population responses in a network.

1 Introduction

An important debate in computational neuroscience centers on the origin of selectivities in populations of cortical cells. A focus of this debate has been an extensive set of empirical data on orientation selectivity in primary visual cortex. A central question concerns how the observed sharp tuning of striate cells arises from broadly-tuned LGN input that is purely excitatory and grows monotonically with contrast. One class of network models posits that the sharp tuning is primarily due to the effects of recurrent connections within the striate population. These *recurrent* models (e.g., Ben-Yishai, Bar-Or, & Sompolinsky, 1995; Somers, Nelson, & Douglas, 1995) can account for a wide range of the empirical data (reviewed in Sompolinsky and Shapley, 1997). Other models, such as *feedforward* models (Troyer, Krukowski, & Miller, 1998) and *gain-control* models (Carandini, Heeger, & Movshon, 1997), emphasize a variety of other mechanisms.

We suggest that some insight into these models can be gained by comparing their answers to the following question: What information is presumed to be contained in the population response? A model that presumes that the population provides a noisy encoding of a single value will focus on methods of removing the noise that would suggest an incorrect value. A model that assumes that the input signal is confounded by irrelevant dimensions, such as contrast variation when the encoded dimension is orientation, will focus on filtering out those other dimensions.

In this paper we propose a different answer to this question. Our hypothesis is that the population response is designed to preserve full information about relevant dimensions in the stimulus. This information could

be a single unambiguous value, an ambiguous value (where the ambiguity in orientation could be due to low contrast, fuzzy edges, curved edges, etc.), or more than one value.

We have shown previously that population responses can represent this additional information, by interpreting the population response as a probability distribution over the underlying stimulus dimension (Zemel, Dayan, & Pouget, 1998). This work did not say anything about how this information could arise in the population, but instead simply assumed that it was there. In this chapter we show how to generate population responses that contain this additional information, in a neural network model using a combination of feedforward and recurrent connections. A novel computational aspect of this approach is that we use the preservation of this information as an objective in training the weights in the model.

The proposal that a model population can faithfully encode more than one value is not without controversy. Carandini and Ringach (1997) studied a simple recurrent model of a hypercolumn of striate cells. Their model successfully and succinctly captures a range of experimental results, including sharp orientation tuning and contrast-invariant tuning width. However, their model makes several peculiar predictions when the input contains more than a single orientation. If the input contains two orientations differing by less than 45°, the model responds as to a single orientation at the mean of the two values; if the two orientations differ by more than 45°, the model responds as if they were nearly orthogonal. The model also cannot signal the presence of three orientations, and generates a spurious orthogonal response to noisy single-orientation inputs. These authors analyzed their model and showed that these effects, termed *attraction* and *repulsion* between orientations, are unavoidable in a broad class of recurrent models of orientation selectivity.

The aim of the model presented here is to explore if a broader range of activity patterns can be maintained within a population, to convey additional relevant stimulus information. We first describe the probabilistic formulation of the information that underlies the population response, then provide details of the model. We then compare the results of our model and several others to a variety of stimuli, and analyze the key differences.

2 Our Approach

The heart of our model is a probabilistic formulation of population responses. We have previously used this formulation to interpret the responses of a population of units in the model. Here we apply this same formulation to adjust the weights in the model to produce the desired responses.

2.1 Probabilistic formulation

We recently developed a *distribution population coding* (DPC) model (Zemel et al., 1998) that generalizes the standard population coding model to the situation in which a probability distribution underlies the population activity. Decoding in DPC, like the standard statistical decoding methods, is based on a probabilistic model of observed responses, which allows the formulation of a decoding method that is optimal in a statistical sense. The key contribution of our model concerns the fact that most models of population codes assume that the population response encodes a single value, and therefore discard information about multiple values and/or ambiguity. By recovering a full distribution, our model preserves this information.

The DPC model begins from the same starting point as the standard model: the neurophysiological finding that in many cases responses of cells within a cortical area can be predicted based on their *tuning curves*. Cell i's tuning curve, $f_i(\mathbf{x})$, describes its expected firing rate, typically defined as a spike count, as a function of the relevant dimension(s) $\mathbf{x}$. These tuning curves are estimated by observing the cell's response on many trials using a set of stimuli that vary along $\mathbf{x}$. A *population code* is defined to be a population of cells whose

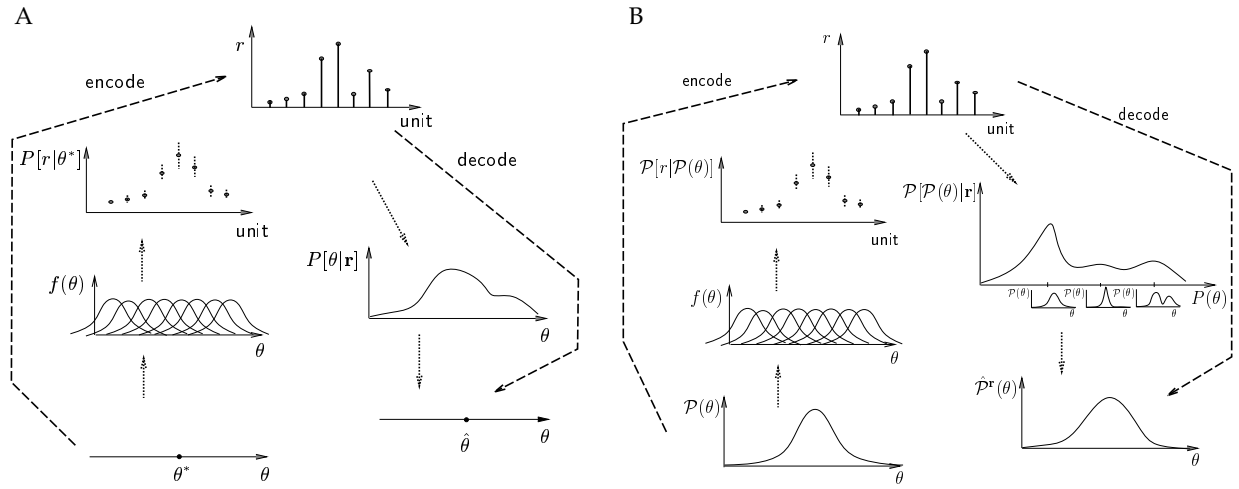

Figure 1: (A) The standard Bayesian population coding framework assumes that a single value is encoded in a set of noisy neural activities. (B) The distribution population coding framework shows how a distribution over θ can be encoded and then decoded from noisy population activities. From Zemel and Dayan (1999).

tuning curves span the space of $\mathbf{x}$ values; this population can thus be considered as a basis for encoding the dimensions $\mathbf{x}$ underlying a set of stimuli. Here we will focus on the case where $\mathbf{x}$ is a single dimension θ, such as the orientation of a grating.

The standard model of population responses (e.g., Seung & Sompolinsky, 1993) assumes that firing rates vary, even for a fixed input. The recorded activity r_i of each unit i is characterized as a stochastic function of its tuning curve $f_i(\theta)$, where the noise n_i is typically zero-mean Gaussian or Poisson, and the responses of the different units are conditionally independent given θ. This leads to the simple *encoding* model, describing how θ is coded in $\mathbf{r}$:

$$r_i = f_i(\theta) + n_i \tag{1}$$

Bayesian *decoding* inverts the encoding model to find the posterior distribution $\mathcal{P}(\theta|\mathbf{r})$, which describes how likely each direction is given the observed responses (Földiák, 1994; Salinas & Abbott, 1994; Sanger, 1996). For example, under an independent Poisson noise assumption,

$$\mathcal{P}[\theta|\mathbf{r}] \sim \log\left\{\mathcal{P}[\theta]\prod_i \mathcal{P}[r_i|\theta]\right\} \sim \sum_i r_i \log f_i(\theta) \tag{2}$$

This method thus provides a multiplicative kernel density estimate, tending to produce a sharp distribution for a single value of θ. A single estimate $\hat{\theta}$ can then be extracted from $\mathcal{P}[\theta|\mathbf{r}]$ using some criterion (e.g., the orientation with the highest probability, or the value that maximizes the data likelihood), and interpreted as the most likely single value in the stimulus. Figure 1A illustrates the framework used for standard decoding.

We are interested in cases where the activities $\mathbf{r}$ code a whole distribution $\mathcal{P}[\theta]$ over the underlying variable, as opposed to a single value θ. A simple associated *encoding* model extends the standard model, positing that the expected response of the cell is the average of its response to the set of values described by $\mathcal{P}[\theta]$:

$$\langle r_i \rangle = \int_\theta \mathcal{P}[\theta] f_i(\theta) d\theta \tag{3}$$

Evidence for this encoding model comes from studies of cells in area MT using stimuli containing multiple motions within the cell's receptive field. The general neurophysiological finding is that an MT cell's response to these stimuli can be characterized as a scaled sum of its responses to the individual components (van

Wezel, Lankheet, Verstrate, Maree, & van de Grind, 1996; Recanzone, Wurtz, & Schwarz, 1997; Treue, Hol, & Rauber, 1999).

Decoding in this model is more complicated than in the standard model. Bayesian *decoding* takes the observed activities and produces a probability distribution over probability distributions over θ, i.e., $\mathcal{P}[\mathcal{P}[\theta]|\mathbf{r}]$. Figure 1B illustrates the framework used for decoding in DPC. In Zemel et al. (1998), we proposed decoding using an approximate form of maximum likelihood in distributions over θ, finding the $\hat{\mathcal{P}}^{\mathbf{r}}(\theta)$ that maximizes

$$L\left[\mathcal{P}(\theta)|\mathbf{r}\right] \sim \sum_i r_i \log\left[f_i(\theta) * \mathcal{P}(\theta)\right] - \alpha g\left[\mathcal{P}(\theta)\right] \tag{4}$$

where $\mathcal{P}[\theta]$ is the prior probability distribution over θ, and the smoothness term $g[]$ acts as a regularizer.

This decoding method is not meant to model neural processing, as it involves complicated and non-neural operations. Simpler decoding methods, such as linear decoding from a set of basis functions, may be used. However, the primary focus in this paper is to understand what information in principle may be conveyed by a population code under the interpretation that populations are encoding distributions. This leads to the interpretation that the extracted distribution $\hat{\mathcal{P}}^{\mathbf{r}}(\mathbf{x})$ describes the information about $\mathcal{P}[\theta]$ available in the population response $\mathbf{r}$. A sensible objective for the population then is that its response to a stimulus characterized by $\mathcal{P}[\theta]$ should faithfully encode this true distribution, i.e., $\hat{\mathcal{P}}^{\mathbf{r}}(\mathbf{x})$ should provide a good approximation to $\mathcal{P}[\theta]$. Indeed, a range of simulations with this model have shown that if activities are generated according to the encoding model (Equation 3), then the decoded distribution $\hat{\mathcal{P}}^{\mathbf{r}}(\mathbf{x})$ does preserve most of the information in the original $\mathcal{P}[\theta]$ (Zemel et al., 1998; Zemel & Dayan, 1999).

In this paper, rather than directly using the encoding model to specify a response pattern $\mathbf{r}$, we use a more realistic network to generate the population activities. The key question of interest if whether activities in a network can faithfully encode the true distribution $\mathcal{P}[\theta]$ in the stimulus.

This aim is analogous to the work of Pouget et al. (1998), in which lateral connections within a population were formulated such that the response of a population to a noisy input was a stable hill of activity, where its peak position corresponded to the most likely direction estimate based on the decoding model (Equation 2). That work showed that the population could actively remove noise in its response to faithfully encode a single estimate. In the work presented here, we aim to do the same thing in the case where the population encodes a distribution rather than a single value.

2.2 Model architecture and dynamics

The basic set-up in our model resembles others, such as the Carandini and Ringach model. The population is a group of rate-coded units identical except in preferred value, where the response of units with preferred value ϕ to a stimulus containing a value θ depends only on $(\theta - \phi)$. The model is governed by the following mean-field equation:

$$\tau \frac{dV}{dt} + V = V^{FF} + V^{LAT} \tag{5}$$

where $V(\theta, t)$ is the membrane potential of all units sharing preferred value θ at time t, τ is the time constant, V^{FF} is the feedforward input from the layer below, and V^{LAT} represents lateral (intra-population) input, which can be resolved into excitatory and inhibitory components: $V^{LAT} = V^{LAT\_E} - V^{LAT\_I}$.

As in many models, the lateral input is obtained by convolving the responses $\mathbf{r}$ with a narrowly tuned set of excitatory weights and a broader profile of inhibitory weights. The combination of these two sets of weights produces a center-surround profile. Thus $V^{LAT} = w * \mathbf{r}$, where the net weights w can be expressed as the difference between excitatory and inhibitory weights. Finally, unit firing rates $\mathbf{r}$ are computed as a function of the unit membrane potential, $\mathbf{r} = g(V)$. We use an approximation to the rectified linear function:

$$g(V) = a \log(1 + \exp(b(V + c))), \tag{6}$$

where $a, b,$ and c are constants (Zhang, 1996). Rectification is a nonlinear operation necessary to prevent negative firing rates; this function $g(V)$ has the added desirable property that it is easy to invert analytically. We can therefore express the lateral input as

$$V^{LAT} = w * g(V) \tag{7}$$

The primary differences in our model from the standard ones arise from the probabilistic formulation described above. We apply this probabilistic formulation in two ways. First, we use it to optimize the lateral weights in the model, so that the population activity patterns provide a faithful encoding of the true information about θ contained in the stimulus. Second, we use the probabilistic model to read out, or interpret the information conveyed about the underlying variable θ by the population response. This is accomplished via the distribution decoding method described by Equation 4. These points are discussed in more detail below.

2.3 Weight optimization

In order for the network response to encode information about θ, we need to be able to specify target response patterns as a function of the evidence for orientation θ in the input. We use the distribution encoding model described by Equation 3 to determine the targets. We defined $\mathcal{P}[\theta]$ as the information about θ in the input. If this is known for a given input, then an activity pattern

$$\mathbf{r} = \{r_i\} = \left\{ \int_\theta \mathcal{P}[\theta] f_i(\theta) d\theta \right\} \tag{8}$$

is the appropriate target, where f_i is the unit tuning function. Each cell's tuning function (describing its expected response to a singular stimulus, containing only one value of the relevant variable) is a circular normal distribution about its preferred value:

$$f_i(\theta) = A + B \exp(K \cos(2 * (\theta - \theta_i))), \tag{9}$$

where θ_i is the i'th unit's preferred orientation, A is baseline (stimulus-independent) firing rate, and B and K determine the amplitude and sharpness of tuning, respectively. We used the values $A = 1$, $B = .013$, and $K = 8$ during weight optimization, giving the cell a maximal firing rate of 40 spikes per second and a relatively sharp tuning profile.

Two important aspects of this approach bear comment here. First, in our model, these tuning functions are used only to generate the target responses; for decoding purposes we use a function based on the model's actual response to a single-orientation stimulus $\mathcal{P}[\theta] = \delta(\theta)$. These two activity patterns may not be the same, as the true stable states of the network may not exactly match the shape of tuning functions used to train it. In this case, decoding must be based on the information available in the observed population responses. Second, any decoding model based on inverting the encoding model should obtain a $\hat{\mathcal{P}}^{\mathbf{r}}(\mathbf{x})$ that approximates the true $\mathcal{P}[\theta]$. Optimization in our model is thus only a function of the encoding method, and does not depend on the particular decoding method.

The recurrent weights, both excitatory and inhibitory, are optimized based on a set of examples. Each example consists of: (a) the feedforward input for a stimulus containing 1, 2, or 3 values, and (b) target output values for the model units, obtained from Equation 8 based on the units' tuning curves (Equation 9) and the values in the stimulus. In our simulations, $\mathcal{P}[\theta]$ is a scaled sum of sharp normal distributions centered on the set of values present in the stimulus. The training set contained 100 examples each of single, double, and triple orientation stimuli, with the orientations selected randomly.

The weights are modified such that the targets are the stable activity profiles of the differential state update equation (Equation 5). This can be implemented in the Fourier domain as

$$\hat{w} = \left\langle \frac{(\hat{\mathbf{V}}(j) - \hat{\mathbf{V}}^{FF})\hat{\mathbf{r}}(j)}{\|\lambda + \hat{\mathbf{r}}(j)\|^2} \right\rangle_{j=1}^{m} \tag{10}$$

where $\mathbf{V}(j)$ is the target membrane voltage (computed as $\mathbf{V} = g^{-1}(\mathbf{r})$), λ is a regularization parameter, and $\langle \ldots \rangle$ denotes the mean over the set of examples.

2.4 Read-out

A second important facet of our model involves a different way of interpreting the activities of the population. In order to determine what information about orientation is present in the unit responses, we apply the statistical decoding approach in the framework outlined above. The method we use here is described by Equation 4. This method takes into account the units' tuning functions in finding the distribution over θ most likely to have generated the observed responses.

The decoding procedure produces a distribution over orientations, $\hat{\mathcal{P}}^{\mathbf{r}}(\mathbf{x})$, that is meant to describe the evidence across the range of orientations in an input stimulus. An additional step is required to match this to the true orientations present in the stimulus. We adopt a simple procedure of picking out the modes of the extracted distribution, and assuming that each mode corresponds to a separate orientation in the input.

3 Results

3.1 Other models

Many different models for orientation selectivity in primary visual cortex have been proposed (Ben-Yishai et al.1995; Somers et al.1995; Carandini & Ringach, 1997; Troyer et al., 1999; Adorjan et al., 2000). We have selected three of these models in order to analyze the significance of different assumptions about the population response and to compare the performance of our model with several others.

The Carandini-Ringach model is the first model for comparison. It contains an architecture and dynamics similar to our model, and its performance on orientation stimuli containing multiple values and noisy values has already been analyzed. A second model we explore here is the Ben-Yishai et al. (1995) model. This model ssumes that a single orientation underlies the input, and endeavors to generate a population response whose tuning is nearly independent of both the contrast and the strength of the orientation bias (anisotropy) in the stimulus. The third and final comparison model is the Pouget et al. (1998) model; as stated above, the primary aim of this model closely resembles ours. Both models attempt to use the recurrent weights in a population to settle on an activity pattern that is an expected pattern under the encoding model. We discuss the main features of these three models and our implementations of them below.

The populations of all models are set up to contain identical, independent units, which differ only in their preferred values θ_i, ranging from $-\pi/2$ to $\pi/2$. Each model contains a dynamic update equation, external input specified as a function of stimulus orientation, a set of recurrent weights, and an activation function for converting membrane potential V into a spike rate r. In all models, feedforward input to a unit depends only on the difference between its preferred value and the orientations in the input, and the recurrent weights are a symmetric function of the difference in preferred values between units.

(1). *Carandini-Ringach model*: This model uses the same update equation as ours (Equation 5), but has a semilinear activation function. The recurrent weights have a center-surround profile given by a weighted difference of Gaussians, clipped at $\pm 60°$. Input to the model is a broadly-tuned Gaussian, and is kept constant during the dynamic updates.

(2). *Ben-Yishai et al.model*: This model uses the update equation (Equation 5) and a semilinear activation function. The recurrent weights have a cosine profile of the form $w(\theta) = -J_0 + J_2 \cos(2\theta)$, where $-J_0$ represents

a uniform level of recurrent inhibition and J_2 determines the strength of orientation-dependent feedback. Input to the model is given by another cosine function, $c(1-\epsilon+\epsilon\cos(2\theta))$, where c represents contrast, and ϵ gives the strength of the orientation bias in the input. This input is constant over time.

(3). *Pouget et al. model*: This model uses an invertible nonlinear activation function like ours (Equation 6). It has no constant external input—the input specifies only the initial state of the network, and has the shape of the unit tuning function. Dynamic updates are performed according to $V_t = (1-\gamma)V_{t-1} + \gamma V_{t-1}^{LAT}$, where γ specifies the rate of change between time steps. Weights in this model are determined by an optimization procedure that seeks to make the tuning functions into stable states of the network. As in our model, tuning functions have the shape of a circular normal distribution.

For the Ben-Yishai et al. model, we performed simulations with the values $J_0 = 86$ and $J_2 = 112$. These parameters specify a regime in which the model's tuning is dominated by cortical interactions and (as desired) is relatively independent of contrast and orientation anisotropy. The results described below all use model populations containing $N = 512$ units, except for the Pouget et al. model, which contained $N = 64$ units. We have since tested this model with 512 units, and found no significant difference in the results.

3.2 Test stimuli

Comparing these models is somewhat complicated for the fact that they make different assumptions about the form of the input they receive. All of the models are capable of generating a narrowly tuned response from a stimulus with a weak single orientation signal. However, the models vary greatly in the tuning width and amplitude of the input they expect to receive. Under the Pouget et al model, expected input has the narrow profile of the unit tuning function, while the Ben-Yishai et al model expects broadly tuned cosine input (in fact, this input is unimodal even when multiple orientations are present). In order to make comparisons fair, for the simulations reported in this paper, each model is given the inputs it expects. For our model, we used input with a circular normal shape, but matched the amplitude and tuning to that of the Gaussian input used in the Carandini-Ringach model. (Parameters: $A = 0$, $B = .182$, $K = 1.7$). This input was at least twice as broad as our unit tuning functions, and is intermediate between the narrow input profile of the Pouget et al model and the broad input of the Ben-Yishai et al model.

We sought to compare the models using test stimuli that simulate basic properties of LGN responses to various complex stimuli, including stimuli containing two or three orientations, or a single, noisy orientation. Although little is known about the actual LGN input to a V1 cell for such stimuli, a reasonable first guess is to use the sum of the responses to each orientation contained in the stimulus. To this end, we generated test input for comparing models by using each model's expected input profile convolved with the orientations present in the stimulus. For noisy input, we used each model's expected input profile corrupted with Gaussian noise.

3.3 Responses to two-value stimuli

The response of our model to stimuli containing two values, of varying angular difference, is shown in Figure 2. Our model forms different stable activity patterns to all three pairs, and the decoding method accurately recovers the original values present in the stimulus. In contrast, none of the other three models is able to veridically encode any of these stimuli, as demonstrated in Figure 3.

The qualitative performance of the different models is highlighted in Figure 4, which represents their responses to all possible two-orientation stimuli. Each horizontal slice represents the response to a particular stimulus, and the angular spread between the two orientations increases from 0° to 90° along the y-axis. Note that our model contains a smooth transition from uni- to bimodal response patterns. The response profile

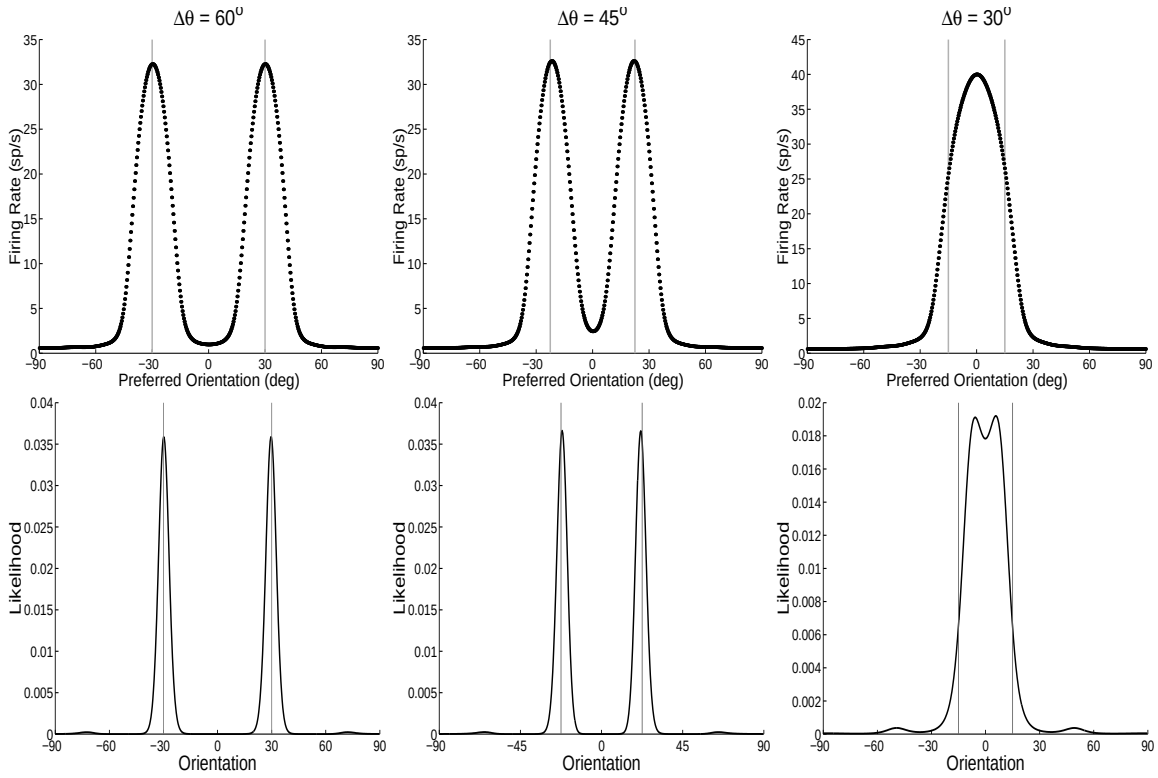

Figure 2: The top row of plots depicts stable response profiles of our model population to stimuli containing two values. The three columns correspond to stimuli with differing angular spread between the pair of orientations. Note that the model can stably maintain bimodal activity patterns even when the modes are far from orthogonal. The bottom row of plots depicts the information about the underlying variable decoded from those response profiles. In both rows of plots, the vertical gray lines are the values of the true orientations contained in the stimulus. The lower right plot demonstrates that information about multiple values may be preserved, even when the population response is itself unimodal.

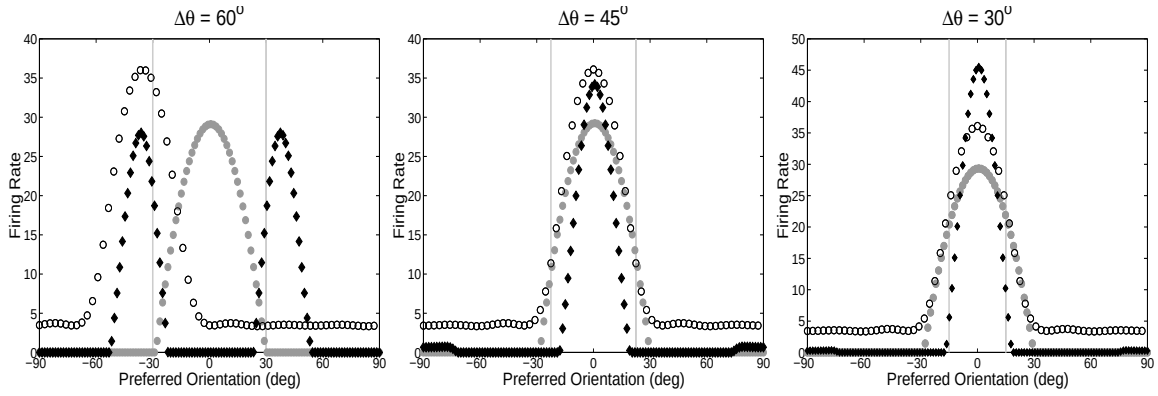

Figure 3: These plots show the response profiles of the comparison models when presented the same stimuli used to generate the responses shown in Figure 2. The open circles correspond to unit responses of the Pouget et al. (1998) model; the gray circles are for the Ben-Yishai et al. (1995) model; and black triangles are for the Carandini and Ringach (1997) model. The vertical gray lines again represent the orientations the models are attempting to encode.

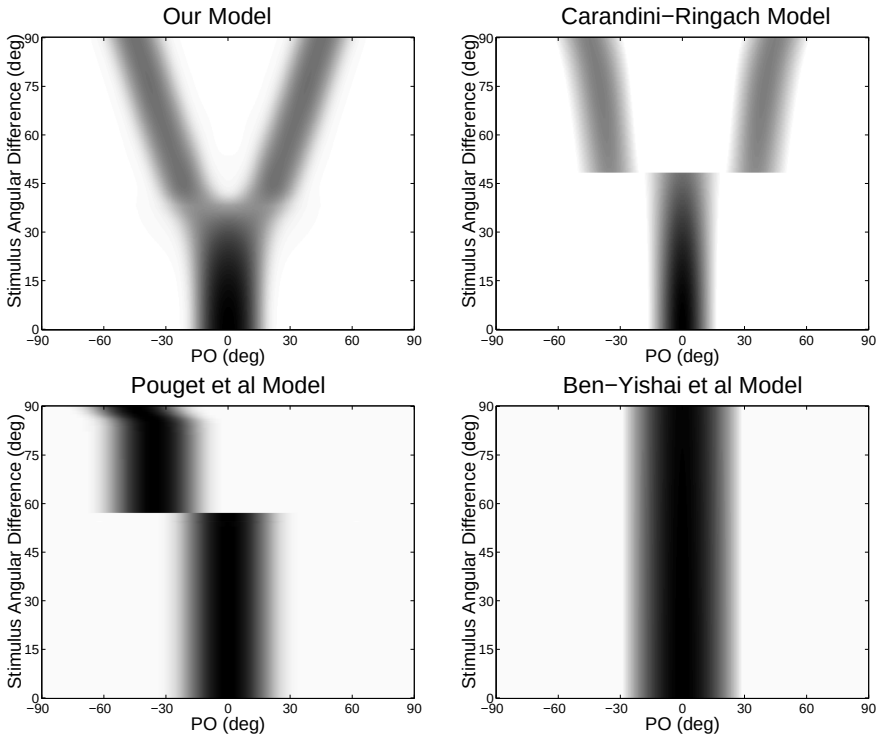

Figure 4: Summary of responses to all 2-value stimuli for all four models. In each plot, the model units are arranged according to their preferred orientation value (PO). Each horizontal slice shows the unit activities to a particular 2-value stimulus. The vertical axis represents the difference between the two values in the stimulus, ranging from identical to orthogonal orientations. The response at the top of the plots is therefore to a stimulus with two orientations present at $\pm 45°$. The "correct" encoding of the stimuli corresponds to a pair of lines running diagonally from $0°$ at the bottom of a plot to $\pm 45°$ at the top. Note that the Carandini-Ringach model response is discontinuous near 45°, while our model contains a smooth transition. The Pouget et al. and Ben-Yishai et al. models both give unimodal responses to all stimuli.

broadens and flattens as the two orientations in the stimulus become more distinct, and then separate into two hills of activity as the spread grows beyond 30°. The Carandini-Ringach model is the only other model capable of generating a stable bimodal response pattern, but note that it contains a significant discontinuity near 45°, where the response jumps from being uni- to bimodal. This shows that the model successfully encodes two orientations when they are nearly orthogonal, but has difficulty with intermediate values.

The behavior of the Ben-Yishai et al. model is determined by the cosine shape of its recurrent weights, together with the fact that recurrent input so strongly dominates its response. A cosine weight profile ensures that recurrent input is really just a measurement of the F1 component of the current population activity, leading to a response centered at the phase of the F1 component of the input. The F1 phase is zero for all stimuli shown here, so the model's response is approximately invariant.

The Pouget et al. model contains a discontinuity in its response and is also incapable of maintaining a bimodal activity profile. For separations smaller than 45° the orientations "attract" and the response is centered at the mean of the two orientations in the stimulus (0°). For larger separations, the two modes in the original input "repel" before one finally wins out, resulting in a response that is centered near either +45° or -45° (these two outcomes are equally likely, though Figure 4 only shows those resulting in a final response centered near -45°).

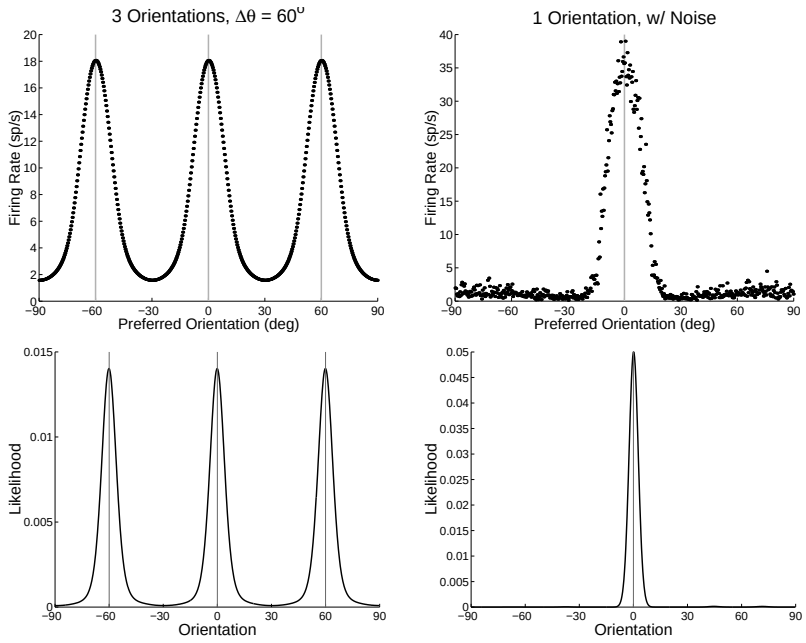

Figure 5: The top row of plots show the response of our model to stimuli containing three values (left), and a single value plus noise (right). The noise added was Gaussian with mean and standard deviation 33% and 20% of the amplitude of the input. The decoding of these responses is shown below. Vertical gray lines show the location of the orientations in the stimulus.

3.4 Responses to three-value and noise stimuli

Our model is also able to maintain a stable response profile for stimuli containing three values, as well as a clean unimodal response to noisy single-value stimuli (see Figure 5). In both cases, the decoding method reconstructs the values present in the stimulus.

None of the comparison models is able to encode three values (Figure 6). The Carandini-Ringach model's response is bimodal, with a peak at one of the stimulus value and at the orientation orthogonal to this value. Because the three orientations used in Figure 6 were evenly spaced (60° apart), noise alone determined which stimulus orientation dominated the model response. The Ben-Yishai et al. model and Pouget et al. models both give unimodal responses to all multi-orientation stimuli. For three evenly-spaced orientations, the input to the Ben-Yishai et al. model is completely isotropic (it is the weighted sum of 3 cosines), so the location of the response peak is entirely random, as seen in Figure 6. The Pouget et al. model response is centered at one of the three input orientations, though also determined arbitrarily. For all three models, the exact location of the response peak(s), for unevenly spaced orientations, is a complicated function of the orientations present in the stimulus, which depends in part on the response discontinuities revealed in Figure 4. However, responses to these stimuli never convey more than the mean, median, or mode of the stimulus values present.

The models fare better when exposed to a single orientation corrupted by input noise (Figure 6). The Ben-Yishai et al. and Pouget et al. models both generate a very smooth population response, resulting from the fact that both models are dominated by recurrent activity and have only weak external input. The Carandini-Ringach model response is somewhat noisier, but the most important difference is that added noise leads to a spurious response at the orthogonal orientation.

The optimized recurrent weights in our model (see Figure 7) have a similar shape to the center-surround weights used by most recurrent models. The wiggles in the surround of the weights gives them some power

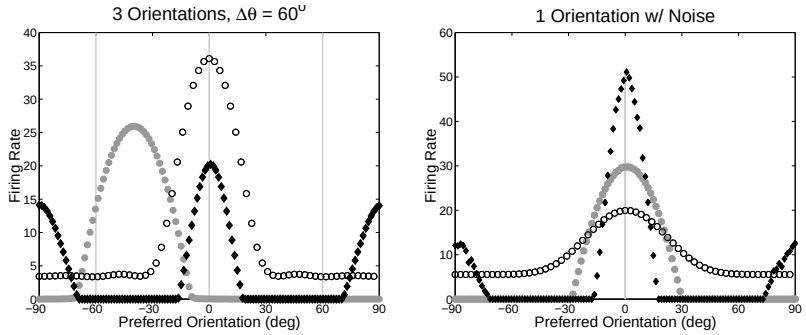

Figure 6: The response of comparison models to inputs containing three values (left), and a single value plus noise (right) (lines labeled as in Figure 3). The three orientations in the left plot are located at 0 and $\pm 60°$. The noise added in the right plot was Gaussian with mean and standard deviation (33%, 20%) relative to the magnitude of the input to each model. Note the spurious response to the orthogonal value in the Carandini-Ringach model.

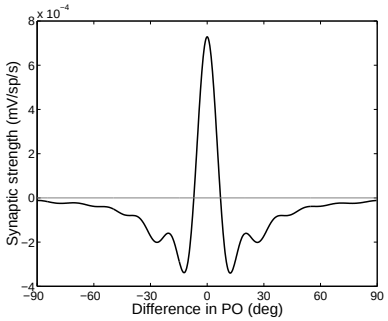

Figure 7: The optimized recurrent connection strengths between pairs of units in the population are plotted as a function of the difference in their preferred values (POs). Recurrent input is computed by convolving this weight profile with the current firing rates. The weights' center-surround profile sharpens the weak bias present in the feedforward input without destroying information about multiple values.

at higher frequencies. We are currently analyzing these differences in greater detail. A crucial difference not apparent in this figure is that the relative strength of the recurrent versus feedforward weights is smaller in our model than the Carandini-Ringach model. This difference likely underlies our model's reduced ability to remove noise and remain strongly contrast invariant, but enhanced ability to encode multiple orientations.

4 Discussion

We have proposed a novel network model of population responses that can support a variety of response profiles, and veridically encodes multiple values. The significant contribution of this model lies in the proposal that the role of the population is to represent the full range of information about orientation present in the input. Previous work established that a population can represent information including uncertainty and multiple values about a variable like orienation, under the hypothesis that a distribution over the variable gives rise to the population activity. Whereas this earlier work took such population activities as given, here we show how these activities can be produced in a standard network model. In this paper, the probabilistic formulation leads both to an objective used to adapt the recurrent weights in the population, and a method for reading out the full information about orientation contained in the population activity.

We have applied this model to the issue of orientation selectivity in striate cortex. We have not yet considered even a fraction of the extensive studies on this issue, but instead our focus is on questions that have received little attention, such as responses to multi-orientation or ambiguous-orientation stimuli. Our model can replicate the cross-orientation effects observed in one of the only empirical investigations into primary visual cortex responses to multiple orientations (DeAngelis, Robson, Ohzawa, & Freeman, 1992). It also makes a number of new predictions concerning responses of these cells to stimuli containing multiple orientations: (1) a range of stable activity patterns will be produced for stimuli containing two orientations with differing angular spread; (2) inputs containing three orientations separated by 60° will produce trimodal activity patterns; and (3) noisy stimuli containing single orientations will only rarely give rise to spurious bimodal response profiles.

Each of these points differs from predictions derived from other models of orientation selectivity. The results presented in this paper show that a range of other models, such as those proposed by Carandini and Ringach (1997), Ben-Yishai et al. (1995), and Pouget et al. (1998) behave differently than our model when the input contains multiple orientation values or noisy single orientations. The brief synopsis of these results is that our model can veridically encode this information while the others cannot. The differences in behavior are somewhat surprising, since at first glance our model closely resembles th eothers. Indeed, the main elements of our models are identical to these models: the activation and tuning functions are the same as in Pouget et al. (1998), while the update equations are the same as in the other two models.

These differences in behavior may primarily be traced to two feature of the models: the relative roles of feedforward and recurrent inputs, and the recurrent weight profile. The models considered here differ greatly along these crucial dimensions. At one extreme of the first dimension lies a model such as Pouget et al. (1998), in which the input is shut off after the initial step. This allows the recurrent lateral connections to force the population activity profile to fit a specified template. At the opposite extreme lies a purely feedforward model. The other models considered here lie at different points along this dimension. The Carandini-Ringach model maintains the input at a constant level throughout the simulation duration, and gives this input a relatively strong weighting, which allows the model to encode multiple orientations when they are sufficiently separated. The Ben-Yishai et al. model also maintains a constant input, but weights it much less than the recurrent inputs, leading to an almost input-invariant response. Other models not considered here may also be partially understood by characterizing their balance of feedforward and recurrent inputs. The model proposed by Troyer et al. (1998) is primarily a feedforward model, which makes it sensitive to variations in the input, but a different formulation of recurrent connections that allows it to filter out large changes in the DC component of the inputs. Finally, a recent model by Obermayer and colleagues (Adorjan, Schwabe, Piepenbrock, & Obermayer, 2000) proposes a mixed answer to this issue. In their model, the roles change over time, as the recurrent weights dominate during the first phase of processing, but the feedforward weights dominate later.

With respect to the second crucial dimension, the recurrent weight profile, the models also differ greatly. Of particular note is the Ben-Yishai et al. model, in which the cosine tuning of the recurrent weights makes the model's response insensitive to the angular difference between orientations in multiple-orientation stimuli.

In our model, we specifically tune the recurrent weights to support multiple orientations and even ambiguity in orientations. We derive an objective function that measured the degree to which this information in the input was preserved in the population, and adjust the weights to maximize this objective. Thus the weight profile is tailored to the underlying hypothesis about the role of the population. Note that this also allows the model to optimize the balance between recurrent and feedforward inputs by adjusting the overall strength of the recurrent weights.

The main conclusions from this study then are that: (1) various orientation selectivity models can be characterized by their assumptions about what information underlies the population responses in V1, which is reflected in their balance between feedforward and recurrent inputs, and in the recurrent weight profile; (2) it is possible to devise a network model that optimizes the recurrent weights and balances these two forces to preserve information about multiple values in the input; and (3) the distribution population coding pro-

vides a natural framework for understanding the information encoded in a population response, and for formulating objectives for preserving and manipulating this information.

We conclude by discussing three final issues and current directions. First, an important and largely neglected aspect of processing in population codes that we have also ignored here is the role of time. It is known that the population response significantly changes during the time-course of a stimulus presentation. The aforementioned model by Adorjan et al. (2000) suggests one method by which the population activity may change over time, initially encoding a single value and then multiple values in a later processing phase. A promising direction for future studies is to incorporate dynamics in the encoded information into the distribution population coding framework studied here. This would considerably extend the applicability o fthe framework. For example, it would permit a natural formulation of ambiguity resolution over time in the population.

Second, an important unresolved issue in the current model concerns the read-out of orientations from the decoded distribution. We adopted a simple approach of assigning a different orientation to each mode of the decoded distribution. However, other interpretations are possible: a distribution with two modes may not represent two orientations but may instead represent a single orientation that has two potential values. This ambiguity between multiplicity and uncertainty may be resolved by using a different encoding (and decoding) model in the DPC framework (Sahani & Dayan, personal communication).

A third and final point concerns the consequences of preserving the full range of orientation information, such as uncertainty and multiplicity. Our model has established that a population can potentially encode this information, but it remains to be seen how this information can be utilized in the next stage of processing, particularly without invoking any complicated non-neural decoding method. We are currently developing a model to investigate how preserving this orientation information may affect processing downstream from V1. It is known that V2 cells respond to illusory contours and figure-ground information; the underlying hypothesis of this new model is that preserving information about multiple orientations within individual V1 populations plays an important role in V2 responses.

Acknowledgements: This work was funded by ONR Young Investigator Award N00014-98-1-0509 to RZ. We thank Peter Dayan and Alexandre Pouget for many useful discussions of statistical population coding.

References

[1] Adorjan, P., Schwabe, L., Piepenbrock, C., & Obermayer, K. (2000). Recurrent cortical competition: Strengthen or weaken? In *Advances in Neural Information Processing Systems 12*, pp. 89–95. Cambridge, MA: MIT Press.

[2] Ben-Yishai, R., Bar-Or, R. L., & Sompolinsky, H. (1995). Theory of orientation tuning in visual cortex. *Proceedings of the National Academy of Sciences, USA*, *92*, 3844–3848.

[3] Carandini, M. & Ringach, D. L. (1997). Predictions of a recurrent model of orientation selectivity. *Vision Research, 37:21*, 3061–3071.

[4] Carandini, M., Heeger, D. J., & Movshon, J.A. (1997) Linearity and normalization in simple cells of the macaque primary visual cortex. *Journal of Neuroscience*, *17:21*, 8621–8644.

[5] DeAngelis, G. C., Robson, J. G., Ohzawa, I., & Freeman, R. D. (1992). Organization of suppression in receptive fields of neurons in cat visual cortex. *Journal of Neurophysiology*, *68:1*, 144–163.

[6] Földiák, P. (1993). The 'ideal homunculus': statistical inference from neural population responses. In Eeckman, F. H. and Bower, J., editors, *Computation and Neural Systems 1992*, pp. 55–60. Norwell, MA: Kluwer Academic Publishers.

[7] Pouget, A., Zhang, K., Deneve, S., & Latham, P.E. (1998). Statistically efficient estimation using population codes. *Neural Computation*, *10*, 373–401.

[8] Recanzone, G. H., Wurtz, R. H., & Schwarz, U. (1997). Responses of MT and MST neurons to one and two moving objects in the receptive field. *Journal of Neurophysiology*, *78:6*, 2904–2915.

[9] Salinas, E. and Abbott, L. F. (1996). A model of multiplicative neural responses in parietal cortex. *Proceedings of the National Academy of Sciences, USA*, *93*, 11956–11961.

[10] Sanger, T. D. (1996). Probability density estimation for the interpretation of neural population codes. *Journal of Neurophysiology*, *76:4*, 2790–2793.

[11] Seung, H. S. & Sompolinsky, H. (1993). Simple models for reading neuronal population codes. *Proceedings of the National Academy of Sciences, USA*, *90*, 10749–10753.

[12] Somers, D. C., Nelson, S. B., & Douglas, R. J. (1995). An emergent model of orientation selectivity in cat visual cortical simple cells. *Journal of Neuroscience*, *15*, 6700–6719.

[13] Sompolinsky, H. & Shapley, R. (1997). New perspectives on the mechanisms for orientation selectivity. *Current Opinion in Neurobiology*, *7*, 514–522.

[14] Treue, S., Hol, K., & Rauber, H-J. (1999). Seeing multiple directions of motion—physiology and psychophysics. *Nature Neuroscience*, *3:3*, 270–276.

[15] Troyer, T. W., Krukowski, A. E., & Miller, K. D. (1998). Contrast-invariant orientation tuning in cat visual cortex: Thalamocortical input tuning and correlation-based intracortical connectivity. *Journal of Neuroscience*, *18:15*, 5908–5927.

[16] Van Wezel, R. J., Lankheet, M. J., Verstraten, F. A., Maree, A. F., & van de Grind, W. A. (1996). Responses of complex cells in area 17 of the cat to bi-vectorial transparent motion. *Vision Research*, *36:18*, 2805–13.

[17] Zemel, R. S., Dayan, P., & Pouget, A. (1998). Probabilistic interpretation of population codes. *Neural Computation*, *10*, 403–430.

[18] Zemel, R. S. & Dayan, P. (1999). Distributional population codes and multiple motion models. In *Advances in Neural Information Processing Systems 11*, pp. 174–180. Cambridge, MA: MIT Press.

[19] Zhang, K. (1996). Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: A theory. *Journal of Neuroscience*, *16*, 2112–2126.