To appear in: R. Hecht-Neilsen (Ed.), *Theories of Cortical Processing*. Springer-Verlag.

Cortical Belief Networks

Richard S. Zemel
Departments of Psychology & Computer Science
University of Arizona
Tucson, AZ 85721
zemel@u.arizona.edu

Abstract

Most theoretical and empirical studies of cortical population codes make the assumption that underlying neuronal activities is a unique and unambiguous value of an encoded quantity. We propose an alternative hypothesis, that neural populations represent, and effectively compute probabilities. Under this hypothesis, population activities can contain additional information about such things as multiple values of or uncertainty about the quantity. We discuss methods for recovering this extra information, and show how this approach bears on psychophysical and neurophysiological studies. A natural extension of this probabilistic interpretation hypothesis casts interacting populations as a belief network, a structure which permits the analysis of information propagation from one population to another. This novel framework for population codes opens up new avenues for studying a diverse set of problems, including cue combination, decision-making, and visual attention.

1 Introduction

A central question in computational neuroscience concerns the information contained in the response of a population of neurons. In the highly constrained and controlled situations typically studied in neurophysiological experiments, the relevant information can often be described as a single underlying stimulus variable. For example, when a monkey is faced with a series of visual stimuli, each of which contains a set of randomly-placed dots all moving in a particular direction, the relevant information in a population of cells that respond selectively to motion direction is that single, unambiguous direction of motion.

In a more natural situation, however, the responses of a population of neurons are influenced by a variety of stimulus aspects, and the information about any particular variable is unlikely to be uniquely specified, with total certainty.

The aim of this chapter is to show that one needs to adopt a different view of the information contained in population in order to properly understand the neural code. The representational

Acknowledgements: Thanks to Peter Dayan, Geoffrey Hinton, Jonathon Pillow, Dave Towers, and Zhiyong Yang for useful discussions, and to Stefan Treue for providing the data plot and for informative discussions of his experiments. Supported by ONR N00014-98-1-0509.

capacity of a population must be extended, as including not just a single value, but instead preserving information about all possible values, and the degree to which each is consistent with the responses. The key hypothesis then that drives this approach is that neural populations represent and in effect compute probabilities.

2 An Example

Many investigators have examined neural and behavioral responses to stimuli composed of two patterns sliding across each other. These often create the impression of two separate surfaces moving in different directions. Neurophysiologists have focused on the responses of cells in the medial temporal (MT) area of monkeys to stimuli such as these, since these cells are selective for particular directions of motion, and are well activated by such random displays provided that some of the dots are moving in the directions that they prefer.

The general neurophysiological finding is that an MT cell's response to these stimuli can be characterized as the average of its responses to the individual components (van Wezel et al., 1996; Recanzone et al., 1997). As an example, Figure 1 shows data obtained from single-cell recordings in MT to random dot patterns consisting of two distinct motion directions (Treue, personal communication). Each plot is for a different relative angle ($\Delta\theta$) between the two directions. A plot can equivalently be viewed as the response of an population of MT cells having different preferred directions to a single presentation of a stimulus containing two directions. If $\Delta\theta$ is large, the activity profile is bimodal, but as the directional difference shrinks, the profile becomes unimodal. The population response to a $\Delta\theta = 30°$ motion stimulus is merely a wider version of the response to a stimulus containing a single direction of motion. However, this transition from a bimodal to unimodal profiles in MT does not apparently correspond to subjects' percepts; subjects can reliably perceive both motions in superimposed transparent random patterns down to an angle of $10°$ (Mather & Moulden, 1983). If these MT activities play a determining role in motion perception, the challenge is to understand how the visual system can extract both motions from such unimodal (and bimodal) response profiles.

Unlike the situation when the stimulus is a set of random dots all moving coherently in the same direction, so that the relevant information is that single direction of motion, in the transparency case the relevant information is *two* directions of motion. This seemingly simple change turns out to be crucial for methods of extracting information, or decoding the population response. As we show below, these methods fail to recover both motions from population responses such as those shown in Figure 1, essentially because they contain an inherent assumption that a single value underlies the response. In addition, these methods fail to recover the relevant information in other random-dot motion stimuli slightly more complicated than the coherent uni-directional stimulus, such as when the dot directions are chosen from a Gaussian distribution about a particular dominant direction.

The random dot motion stimuli are intended as simple examples that have been explored in neurophysiological studies. The idea that neural populations are best viewed as encoding probabilities over variables rather than single unambiguous values is relevant not just in these carefully constructed stimulus situations, but instead applies widely to many natural situations, particularly when stimuli become less constrained and more complex.

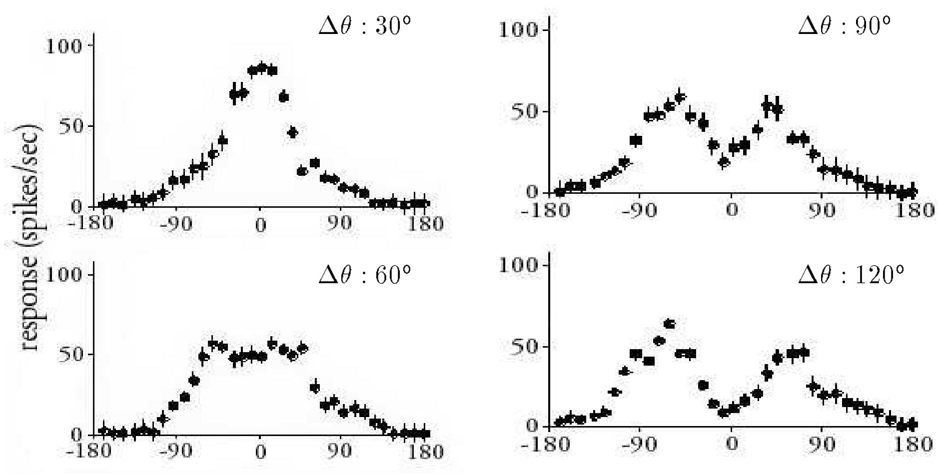

Figure 1: Each of the four plots depicts a single MT cell response (spikes per second) to a transparent motion stimulus of a fixed directional difference ($\Delta\theta$) between the two motion directions. The x-axis gives the average direction of stimulus motion relative to the cell's preferred direction ($0°$). From Treue, personal communication.

For example, uncertainty must be taken into account when one considers the general case of image motion, and velocity encoding in local regions. The response of a population of MT cells, when only a single straight edge is moving in their common receptive field, contains an inherent ambiguity—several two-dimensional velocities are consistent with the velocity of any moving one-dimensional pattern (e.g., a long edge or grating). This ambiguity is known as the "aperture problem", since the motion of a one-dimensional stimulus viewed through a small circular aperture is ambiguous. In such cases where the velocity is inherently ambiguous, extracting a single velocity from the MT population code is impossible. In locations where the stimulus contains more spatial structure, the motion may be disambiguated. However, when noise is considered, all locations will have some degree of ambiguity. The system then needs to quantify the degree to which the data is consistent with a particular velocity. In addition, uncertainty about local velocity stems from sources other than local geometry, including contrast and duration. At high contrast, the image information may unambiguously support the true velocity. But at lower contrasts, many velocities may be consistent with the image information, and distinguishing the true velocity from others will be increasingly difficult as contrast is reduced. Similarly, at short durations distinguishing the true velocity will also be difficult, and a range of velocities will be consistent with the image information.

The second aspect of image information that motivates the interpretation of a population response as a full distribution is the presence of multiple values within a receptive field. Transparency is just one example of this. Another example concerns higher levels of the visual system. For example, in inferotemporal (IT) cortex, cells have large receptive fields, and in a natural scene containing multiple objects, more than one object is likely to lie within the shared receptive fields of a single IT population. Any description of the objects that can account for that population response would thus need to include support for each of the objects present. In this case, as with the case of uncertainty, a more natural description of the information contained in the population response is not a single best value of the relevant variable but rather the probability that various values are each consistent with the observed response.

The rest of this chapter is organized as follows: we first describe methods by which a population of cells can represent a probability distribution over an underlying value. We then show how this approach is useful for elucidating cortical function in a number of different ways, including analyzing responses of a single cortical population, understanding lateral interactions within a population, and characterizing information processing from one population to another.

3 Representing Distributions in Populations

From the standpoint of information processing, representing probabilities rather than just single values can lead to significant improvements. Consider an abstract view of cortex as a multiple-level processor where information converges and diverges from several populations at one level to several at the next. If at one level the information in the population response is reduced to a single value, then downstream populations will not have access to the uncertainty of the estimate, and to the likelihood of other values. This uncertainty may be important, particularly as information from other populations is included, to ensure that potentially better values are not ignored. Representing the degree of support across the range of values prevents premature commitment, and enables better judgements to be made downstream.

A key point with respect to cortical representations of probabilities is that the probabilities need not be explicitly decoded from a population response. If the probabilistic information contained in one population can give rise to probabilistic information in another population, no explicit decoding ever needs to take place. Thus the focus in this chapter is not on neural mechanisms of extracting distributions from a population response (i.e., how subsequent stages of neural processing might actually extract information from the population), but rather on methods of establishing in a more abstract way the information that can be shown by any means to be present (a similar approach is pursued for single neuron in decoding in Rieke, Warland, De Ruyter, & Bialek, 1997).

Methods for representing a probability distribution in a population can be divided into two classes: basis functions, and generative methods. Before describing these two classes, we introduce some notation. Associated with each cell in the population is a *tuning function* $f_i(\mathbf{x})$, which describes the expected response of that cell as a function of the value of the relevant variable. These tuning functions are typically determined using simple stimuli (e.g., ones containing uni-directional motion). An important point is that the methods discussed here make no assumptions about the form of these tuning functions; rather, all that they require is some description of how the cell response varies as function of the variable value. The aim then is to form a distribution $\hat{\mathcal{P}}^{\mathbf{r}}(\mathbf{x})$ over these variable values based on the response vector $\mathbf{r}$. The elements in the response vector, r_i could, for example, be the number of spikes that cell i emits in a fixed time interval following the stimulus onset. Equivalently, r_i could could be considered the rate parameter of a Poisson distribution that describes the neuron's probability of emitting a spike. The goal is to decode the collection over values that are consistent with these responses on that trial.

3.1 Basis function representations

Heeger and Simoncelli (1991) proposed a model in which MT cell responses were viewed as spatiotemporal filters that could be used to derive not a single local estimation of motion but instead a local likelihood map, i.e., a distribution over velocities. Weiss and Adelson (1998) have recently adopted a similar approach to model a wide range of results in visual motion psychophysics. Basis function methods have also been proposed as a way of understanding population responses in parietal cortex (Pouget & Sejnowski, 1997). A general theory of using basis functions as a means to interpret population codes as representing probability distributions was developed by Anderson (1994) and Anderson and Van Essen (1994).

A basis function approach represents the distribution $\hat{\mathcal{P}}^{\mathbf{r}}(\mathbf{x})$ as a linear combination of simple kernel or basis functions $\psi_i(\mathbf{x})$ associated with each cell, weighted by a normalized function of its activity r_i:

$$\hat{\mathcal{P}}^{\mathbf{r}}(\mathbf{x}) = \sum_i r_i' \psi_i(\mathbf{x}) \quad (1)$$

Here the r_i' are normalized such that $\hat{\mathcal{P}}^{\mathbf{r}}(\mathbf{x})$ is a probability distribution. If the $\psi_i(\mathbf{x})$ are probability distributions themselves, and r_i are all positive, a natural choice is to have

$$r_i' = \frac{r_i}{\sum_j r_j}. \quad (2)$$

Note that the kernel functions $\psi_i(\mathbf{x})$ are *not* the tuning functions $f_i(\mathbf{x})$ of the cells that would commonly be measured in an experiment. They need have *no* neural instantiation; instead, they form part of the interpretive structure for the population code.

3.2 Generative representations

The basis function approach can be viewed as approximating a probability distribution by adding together distributions associated with each neuron. The second class of models also builds a distribution by adding pieces together, but the addition is done in the domain of log probabilities (Hinton, personal communication). This overcomes an important limitation of the basis function approach, which cannot represent distributions that are sharper than the component distributions.

This second class of models adopts a *generative* approach, where the goal is to find a value or set of values of the underlying variable $\mathbf{x}$ that could have generated the observed population response $\mathbf{r}$. These methods begin by hypothesizing an *encoding* model, that describes how $\mathbf{x}$ is encoded in the noisy elements of the population response. This encoding model specifies $\mathcal{P}[\mathbf{r}|\mathbf{x}]$, a distribution over response patterns given the variable. They then apply a *decoding* algorithm to estimate the distribution over $\mathbf{x}$, typically using Bayes Rule to invert the relationship between $\mathbf{r}$ and $\mathbf{x}$.

3.2.1 Standard Bayesian approach

A standard approach to decoding population responses turns out to be a type of generative representation. This approach has enjoyed some success in applications to decoding actual population responses. For example, Zhang, Ginzburg, McNaughton, and Sejnowski (1998) applied the Bayesian method to estimate the position of a rat on a maze based on spike counts from approximately 30 hippocampal place cells; the estimation accuracy was within the accuracy range of the position sensor on the rat.

Standard Bayesian decoding starts with the activities $\mathbf{r} = \{r_i\}$ and generates a distribution $\mathcal{P}[\mathbf{x}|\mathbf{r}]$. The simplest model assumes that the response of each cell i in the population is Poisson where $f_i(\mathbf{x})$ is the mean rate, and that the noise in these responses is independent. Under this model:

$$L\left[\hat{\mathcal{P}}^{\mathbf{r}}(\mathbf{x})\right] = \log \mathcal{P}[\mathbf{x}|\mathbf{r}] \sim \log \left\{\mathcal{P}[\mathbf{x}] \prod_i \mathcal{P}[r_i|\mathbf{x}]\right\} \sim \sum_i r_i \log f_i(\mathbf{x}) \tag{3}$$

This method is also a form of basis function estimate, but it is *multiplicative*, because it is additive in the log domain. It thus tends to produce a sharp distribution for a single value of $\mathbf{x}$. In the standard approach, a single estimate $\hat{\mathbf{x}}$ is extracted from $\mathcal{P}[\mathbf{x}|\mathbf{r}]$. However, this final step is not necessary, and the information about all possible values of $\mathbf{x}$ can be maintained.

By comparing Equations 1 and 3, one sees that the decoding in both cases involves a weighted sum of each cell's tuning function (or log thereof), weighted by its observed spike count. In the generative case, however, this quantity is proportional to the log of the distribution, which allows the resulting distribution to be concentrated on specific values of the underlying variable.

3.2.2 Distributional Population Coding

The standard Bayesian encoding model assumes that the population responses are caused by a single value $\mathbf{x}^*$ of the relevant underlying variable: $\langle r_i \rangle = f_i(\mathbf{x}^*)$. Under this model, Bayesian decoding takes the observed activities $\mathbf{r}$ and produces probability distributions over $\mathbf{x}$ (Equation 3).

Our extended Bayesian distributional population coding (DPC) method (Zemel, Dayan, & Pouget, 1998) extends the standard encoding model to allow $\mathbf{r}$ to depend on general $\mathcal{P}[\mathbf{x}]$:

$$\langle r_i \rangle = \int_{\mathbf{x}} \mathcal{P}[\mathbf{x}] f_i(\mathbf{x}) d\mathbf{x} \tag{4}$$

Under this model, Bayesian decoding takes the observed activities $\mathbf{r}$ and produces probability distributions over probability distributions over $\mathbf{x}$, $\mathcal{P}[\mathcal{P}(\mathbf{x})|\mathbf{r}]$. For simplicity, we decode using an approximate form of maximum likelihood in distributions over $\mathbf{x}$, finding the $\hat{\mathcal{P}}^{\mathbf{r}}(\mathbf{x})$ that maximizes

$$L\left[\hat{\mathcal{P}}^{\mathbf{r}}(\mathbf{x})\right] \sim \sum_i r_i \log\left[f_i(\mathbf{x}) * \hat{\mathcal{P}}^{\mathbf{r}}(\mathbf{x})\right] - \alpha g\left[\hat{\mathcal{P}}^{\mathbf{r}}(\mathbf{x})\right] \tag{5}$$

where the smoothness term $g[]$ acts as a regularizer.

The distributional *encoding* operation in Equation 4 is quite straightforward – by design, since this represents an assumption about what neural processing prior to the population under study performs. However, the distributional *decoding* operation that we have used (Zemel et al., 1998) involves complicated and non-neural operations. The idea is to understand what information in principle may be conveyed by a population code under this interpretation, and then to judge actual neural operations in the light of this theoretical optimum. Our extended Bayesian approach is a statistical cousin of so-called line-element models, which attempt to account for subjects' performance in cases like transparency using the output of some fixed number of direction-selective mechanisms (Williams, Tweten, & Sekuler, 1991).

Some intuition into the relationship between the two generative models can be gained by inspecting Equations 3 and 5. Both models can be viewed as forming $\hat{\mathbf{r}}$, a predicted value of $\mathbf{r}$ based on some function of $\mathbf{x}$, to match the actual observed rates $\mathbf{r}$: the likelihood $L = \sum_i r_i \log \hat{r}_i$ is maximized when $\{\hat{r}_i\}$ matches $\{r_i\}$. In the standard method, a single value of $\mathbf{x}$ gives rise to the observed rates ($\hat{\mathbf{r}}$ is based on a single value of $\mathbf{x}$), but in the extended method, a full distribution over $\mathbf{x}$ gives rise to the rates, as $f_i(\mathbf{x})$ is convolved with the current estimate $\hat{\mathcal{P}}^{\mathbf{r}}(\mathbf{x})$. This latter approach fits with the objective of finding the best collection over values of $\mathbf{x}$ to match the actual population response.

4 Applying Distributional Population Coding

4.1 Population analysis

We have applied these methods in a number of domains. The most straightforward application of these methods is to decode distributions from the response of a population of neurons. Below we briefly describe two such applications.

4.1.1 Decoding transparent motion

We have applied these methods to simulated MT population responses to multiple motion stimuli, such as the responses shown in Figure 1. We summarize the results here; further details may be found in Zemel & Dayan (1999). The underlying variable of interest in this domain is the motion direction(s), so $\mathbf{x} = \theta$. In the simulations described here, we have used a population of 200 model MT cells, with tuning functions defined by random sampling within physiologically-determined ranges for the parameters: baseline b, amplitude a and width σ. The encoding model comes from the MT data: for a single motion, $\langle r_i|\theta\rangle = f_i(\theta) = b_i + a_i \times \exp[-(\theta - \theta_i)^2/2\sigma_i^2]$, while for two motions, $\langle r_i|\theta_1, \theta_2\rangle = \frac{1}{2}[f_i(\theta_1) + f_i(\theta_2)]$. The noise is taken to be independent and Poisson. For multiple motion stimuli, this encoding model produces the observed neurophysiological response: each unit's expected activity is the average of its responses to the component motions.

For the distributional population coding methods to decode successfully when there are two motions in the input (θ_1 and θ_2), the extracted distribution must at least have two modes. Standard Bayesian decoding (Equation 3) fails to satisfy this requirement. First, if the response profile $\mathbf{r}$ is unimodal (cf. the 30° plot in Figure 1), convolution with unimodal kernels $\{\log f_i(\theta)\}$ produces a unimodal $\log \mathcal{P}[\theta|\mathbf{r}]$, peaked about the average of the two directions. The basis function method (Equation 1) suffers from the same problem, and also fails to be adequately sharp for single value inputs.

Surprisingly, the standard Bayesian decoding method also fails on bimodal response profiles. If the baseline response $b_i = 0$, then $\mathcal{P}[\theta|\mathbf{r}]$ is Gaussian, with mean $\sum_i r_i\theta_i / \sum_{i'} r_{i'}$ and variance $1/\sum_i r_i/\sigma_i^2$ (Snippe, 1996; Zemel et al., 1998). If $b_i > 0$, then, for the extracted distribution to have two modes in the appropriate positions, $\log[\mathcal{P}[\theta_1|\mathbf{r}]/\mathcal{P}[\theta_2|\mathbf{r}]]$ must be small. However, the variance of this quantity is $\sum_i \langle r_i\rangle \left(\log[f_i(\theta_1)/f_i(\theta_2)]\right)^2$, which is much greater than 0 unless the tuning curves are so flat as to be able to convey only little information about the stimuli. Intuitively, the noise in the rates causes $\sum r_i \log f_i(\theta)$ to be greater around one of the two values, and exponentiating to form $\mathcal{P}[\theta|\mathbf{r}]$ selects out this one value. Thus the standard method can only extract one of the two motion components from the population responses to transparent motion.

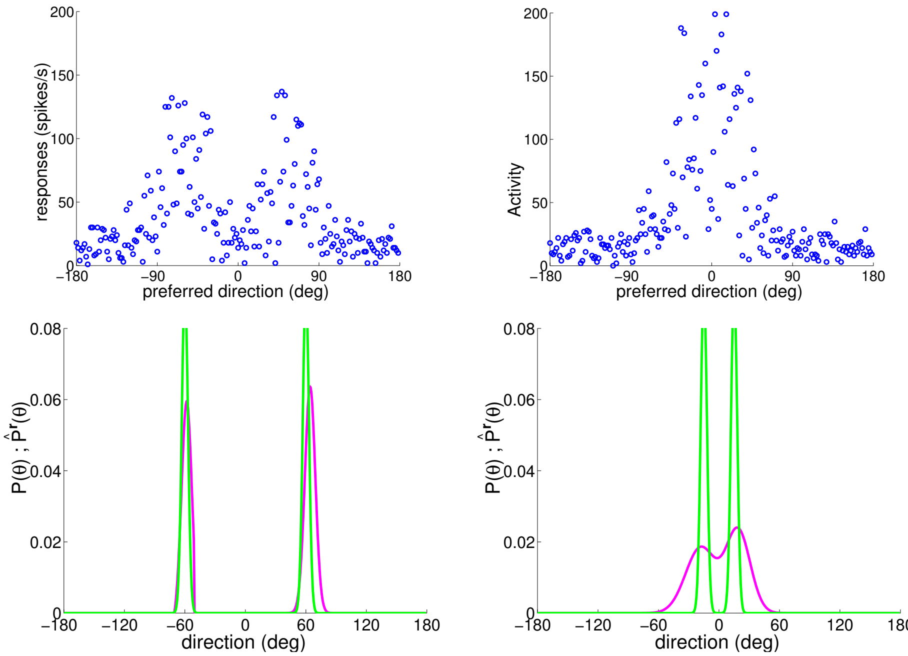

Figure 2: (A) On a single simulated trial, the population response forms a bimodal activity profile when $\Delta\theta = 120°$. (B) The reconstructed (darker) distribution closely matches the true input distribution for this trial. (C) As $\Delta\theta \to 10°$, the population response is no longer bimodal, instead has a noisy unimodal profile, and (D) the reconstructed distribution no longer has two clear modes.

Our extended Bayesian method, on the other hand, matches the generating distribution for bimodal response patterns (Figure 2). For unimodal response patterns, such as those generated by double motion stimuli with $\Delta\theta = 30°$, DPC also consistently recovers the generating distribution. The bimodality of the reconstructed distribution begins to break down around $\Delta\theta = 10°$, which is also the point at which subjects are unable distinguish two motions from a single broader band of motion directions (Mather & Moulden, 1983).

It has been reported (Treue, personal communication) that for angles $\Delta\theta < 10°$, subjects can tell that all points are not moving in parallel, but are uncertain whether they are moving in two discrete directions or within a directional band. Our model qualitatively captures this uncertainty, reconstructing a broad distribution with two small peaks for directional differences between 7° and 10°. Our model also matches psychophysical performance on other aspects of these stimuli, and generates a number of testable predictions.

4.1.2 Decision noise

A second application of distribution population coding concerns a well-studied set of psychophysical and neurophysiological results for a dot-motion task in which monkeys must discriminate the single coherent direction of motion amidst a random motion background. These data pose a notorious puzzle for population coding: it seems that there are *single* cells in monkey MT that are just as accurate as the whole monkey, i. e., an ideal observer judging the direction of stimulus motion based on the response of optimally stimulated MT neurons could match the monkey's

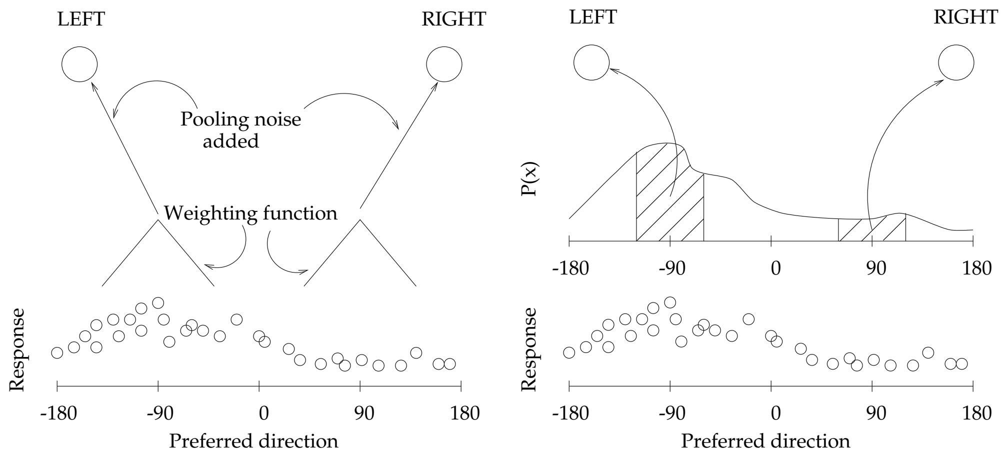

Figure 3: (A) The value assigned to each pool in the Shadlen et al. (1996) model is a random variable whose mean is a weighted combination of responses of neurons whose preferred directions are close to the target direction. The larger pooled signal dictates the response. (B) The output of the extended Bayesian DPC model is the integral under the inferred distribution within a window around the target direction.

psychophysical performance (Britten et al., 1992). The puzzle is that monkeys are not much better than their single cells, since one would expect them to be able to average away substantial noise. That is, neural coding seems unreasonably inefficient. Shadlen, Britten, Newsome, and Movshon (1996) analyzed this puzzling inefficiency and proposed a model, which includes a number of factors that contribute to the poor population performance. Even after taking account of factors such as correlations between the activities of cells with similar orientation preferences, their model requires the addition of extra noise in the decision-making stage to reduce the accuracy appropriately. Distributional population coding offers an alternative hypothesis to account for the apparent inefficiency: that the activity of the population should be seen as coding the whole distribution of inputs rather than just a single direction of motion in the face of noise. Here we outline the application of this approach to the decision-noise paradigm.

The experimental paradigm involves measuring neuronal responses while monkeys perform a two-alternative forced discrimination of motion direction. The monkeys choose from opposite directions in displays containing a unidirectional motion signal embedded in a field of random motion noise. The axis of the direction discrimination is tailored to the preferred direction of each individual neuron (but we will call it 'up' or 'down'). The *coherence*, or strength of the signal is varied by changing the percentage of dots moving in the single direction.

In the Shadlen et al. (1996) model, the final decision about direction is based on pooling the activities across the population (see Figure 3A). The details of the model include correlated noise in the activity of MT neurons; scaled inclusion in the pool of neurons whose preferred direction did not match the discrimination direction; and, crucially, noise in the output of each pool. The final decision was based on which pool had the larger value (see Figure 3A).

We attribute the apparent inefficiency to a different cause, suggesting that it is unreasonable to think of the output of all the neurons as being combined directly to make a decision about one of two directions of motion. Rather, the population can be taken as encoding *extra* information, namely a whole distribution over the directions of motion in the input, and inferences about

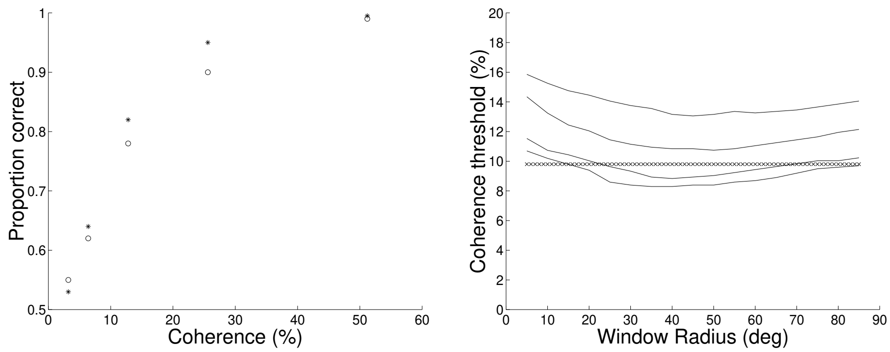

Figure 4: (A) The open symbols depict the monkey's performance on one direction discrimination experiment as a function of stimulus strength (Britten et al., 1992). The closed symbols show the proportion correct of the DPC model; each symbol is the average of 200 trials in which the unit responses were determined by the encoding model (Equation 4). The integration window radius is 20°, and the population contains 100 units. Britten et al. summarize these results by a fitting a function of the form $p = 1 - 0.5\exp[-(c/\alpha)^{\beta}]$ to the points, where p is the proportion correct, c is the coherence level, α is the coherence level supporting 'threshold' performance (82% correct), and β is the function slope. The key empirical parameter that Shadlen et al. (1996) attempt to model is α. (B) The output of the DPC model depends on the radius of the integration window around each target direction and the number of units in the population ($N = 50, 100, 150, 200$ from top to bottom). The model's performance is summarized by the threshold coherence level (α). The horizontal line is the empirical mean $\alpha = 9.8\%$ reported in Shadlen et al. (1996).

whether there is more 'up' than 'down' in the input have to be based on the distribution that can be decoded from this (see Figure 3B). To put it another way, the population only looks inefficient if it is treated as coding the proportions of motion in two directions, not if it is treated as coding the proportions of motion in all directions.

We have applied the extended Bayesian distributional population coding (DPC) model to this domain. Figure 4A shows that the model's performance degrades as the stimulus coherence decreases, in a manner consistent with the monkey's psychophysical performance. Figure 4B shows how the model's performance depends on the radius of the window around the target direction, as well as the number of units in the population. This plot suggests a bias/variance trade-off in the size of the window – a larger window involves more units in the target discrimination (through the decoding process), but also increases the effect of the motion noise.

These results demonstrate that our model is generally consistent with the monkey performance over a range of parameter values; yet, like the penultimate models in Shadlen et al. (1996), for the best parameter values it exceeds the accuracy of the monkeys. However, we have not yet included the effects of correlations in the activities of cells.

4.2 Lateral interactions

Distributional population coding can provide a method of analyzing population responses in domains other than motion processing. We briefly summarize one application here, concerning another well-studied area in population coding: orientation tuning in primary visual cortex. One class of network models posits that the sharp tuning of striate cells arises from broadly tuned feedforward LGN inputs which are substantially sharpened by recurrent connections within the striate population. These *recurrent* models can account for a wide range of the empirical data (reviewed in Sompolinsky and Shapley, 1997).

Carandini and Ringach (1997) studied a simplified recurrent model that replicates many observed characteristics of a hypercolumn of striate cells. The core of their model is a single equation implementing a feedback filter with a center-surround weighting function. When the input to the cell population contains two orientations, their model exhibits peculiar responses: if the two orientations differ by less than 45°, the model responds as if the input contained a single orientation at the mean of the two values; if the angular difference between the two orientations exceeds 45°, the model responds as if they were nearly orthogonal. The model also cannot signal the presence of three orientations, and creates spurious orthogonal orientations to noisy single orientation inputs. These authors analyzed their model and showed that these effects of attraction and repulsion between orientations are unavoidable in a broad class of recurrent models of orientation selectivity.

This model thus makes strong predictions about the visual response of cortical neurons to stimuli containing multiple orientations. Relevant empirical data are scarce. DeAngelis et al. (1992) found that the response of a striate cell to its optimal orientation is reduced when the stimulus contains a second orientation. Cell responses to pairs of non-optimally oriented stimuli are not known. Nonetheless, the predictions of the model appear dubious: it predicts that the neurons will be unable to veridically encode multiple orientations, even when the orientations differ by nearly 45°.

We have proposed an alternative model (Zemel & Pillow, 1999) that is able to maintain a broader range of activity patterns in the population for multiple orientation stimuli. The dynamics in this model are very similar to the Carandini-Ringach model, with three modifications. Our model utilizes optimized recurrent weights in the population; shunting inhibition of the excitatory potential; and a nonlinear activation function.

The model uses the extended Bayesian DPC method to determine what information about orientation is present in the model cell responses. This approach is essential in this case, since the central question is whether information about multiple orientations is preserved in the resulting patterns. As was the case in the motion domain, a key feature of this method is that unimodal response profiles do not necessarily indicate that only one value is present in the input; instead, a broad unimodal profile is consistent with a stimulus containing two nearby orientations. With respect to DPC, a novel aspect of this application is that the encoding model (Equation 4) is used to optimize the recurrent weights within the population: target values are obtained by convolving the tuning functions (preferred orientation and tuning width) of the model striate cells with the set of orientations present in the stimulus.

The response of the model to stimuli containing two orientations, of varying angular difference, is shown in Figure 5. The Carandini-Ringach model is unable to veridically encode these stimuli. Our model forms different stable activity patterns to all three pairs, and the decoding method forms a bimodal distribution in each case, suggesting that two orientations are present in the stimulus.

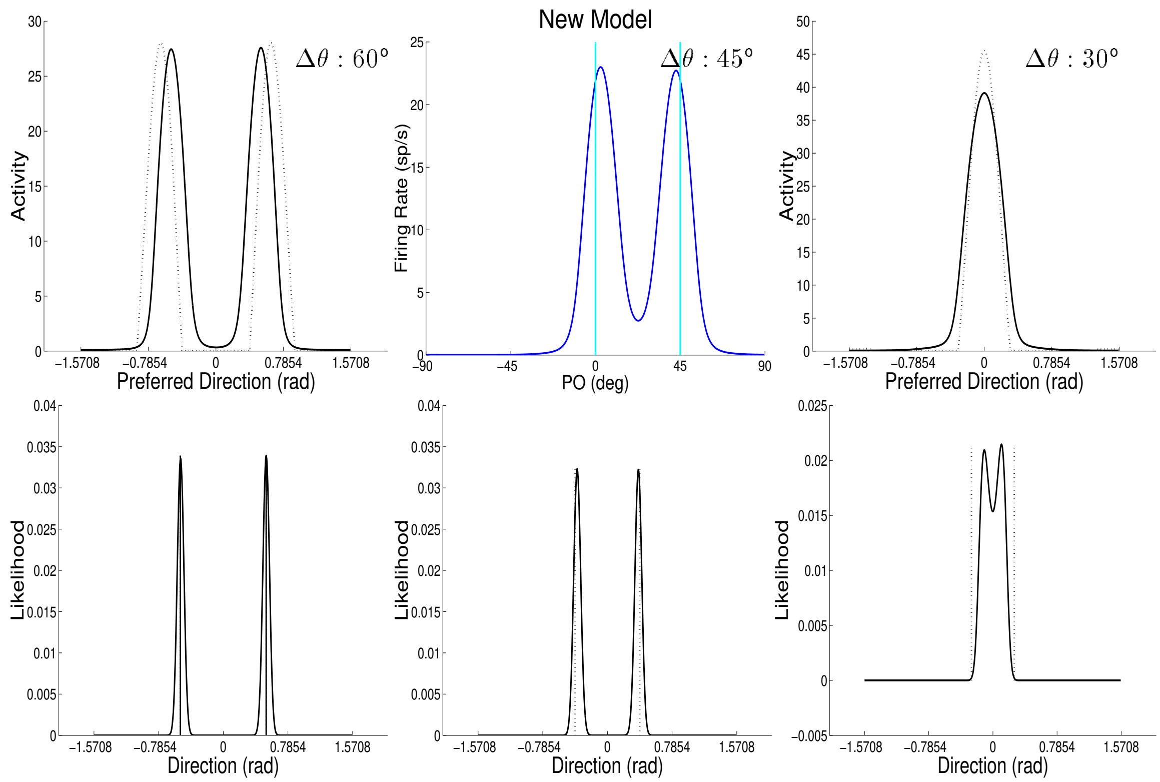

Figure 5: The top row of plots depicts stable response profiles of model cells to stimuli containing two orientations. The three columns correspond to differing angular spread between the pair of orientations. In the top row of plots, the thicker dots show the cell responses in our model, the thinner dots show the responses of the Carandini-Ringach model to the same inputs. The bottom row of plots depicts the information about orientation decoded from those response profiles. Here the thin dotted lines are the true orientations in the model inputs, and the thick lines are the orientations encoded in the model responses. The contrast between the models' responses in the middle column shows that ours can stably maintain bimodal activity patterns even when the modes are not orthogonal. And information about both orientations is retained even as the angular difference shrinks below 45°.

4.3 Cortical belief network

DPC provides a method of interpreting general probability distribution from a population of units representing a particular variable. The larger goal is to understand how probabilities, or distributional information in one population can affect the responses of another population. A network composed of many populations representing different, related variables must properly combine distributions.

This problem of combining distributions can be cast as a formulation of a type of *belief network*, in which associated with each population code is a probability distribution over a particular variable or set of variables, and the links between the populations express conditional dependencies between these variables.

Consider a simple situation in which an animal is stationary while an object is moving and the animal wants to estimate time-to-collision. Various forms of information represented in different neural populations are useful in this task, including one population of cells representing the direction of motion of an object, and a second population encoding the distance from the object to the animal. These two variables, along with others such as the object speed, can be used to infer the time-to-collision of the object, also encoded in a population of neurons.

We cast this problem in the probabilistic framework of a belief network, which sets up a description of the computation to be implemented in populations of neurons. The cortical belief network framework for this example is cartooned in Figure 6. Links between a population encoding direction and a second population encoding time-to-collision (T) the typical relations between these variables. For example, a very short T value is consistent only with object motions directly towards the animal, while longer times are consistent with a range of directions. Based on these conditional relations, and assuming that direction and distance are conditionally independent given time-to-collision, then a distribution over T can be inferred from distributions over the other two variables.

Given true probability distributions $\mathcal{P}[\theta|\omega]$ and $\mathcal{P}[Z|\omega]$ over the direction (θ) and distance (Z) of the object (here ω represents the underlying information in the environment, including visual and extra-retinal cues), inferring time-to-collision (T) requires calculating:

$$\begin{aligned} \mathcal{P}[T|\omega] &= \int_{\theta,Z} \mathcal{P}[\theta|\omega]\mathcal{P}[Z|\omega]\mathcal{P}[T|\theta,Z]d\theta dZ \\ &\propto \mathcal{P}[T]\int_{\theta,Z} \mathcal{P}[\theta|\omega]\mathcal{P}[Z|\omega]\mathcal{P}[\theta,Z|T]d\theta dZ \end{aligned} \tag{6}$$

where $\mathcal{P}[T]$ is the prior distribution over T. This equation governs computation in the belief network (shown in dashed lines in Figure 6), which forms a target for the computation carried out in the neurons.

We have therefore reduced the computational problem to one of mapping activities $\mathbf{r}^\theta$ and $\mathbf{r}^Z$ into activities $\mathbf{r}^T$ for which $\mathcal{P}[T|\mathbf{r}^T]$ is a good approximation to the integration in Equation 6, where $\mathcal{P}[\theta|\omega]$ is what $\mathbf{r}^\theta$ represents and $\mathcal{P}[Z|\omega]$ is what $\mathbf{r}^Z$ represents.

An important point is that the belief network—the probability distributions and the conditional links—are not actually implemented. Instead it provides an interpretation for the network of neurons. Thus this framework provides a dual representation in which the probabilities are implicitly encoded, while the population responses are explicit.

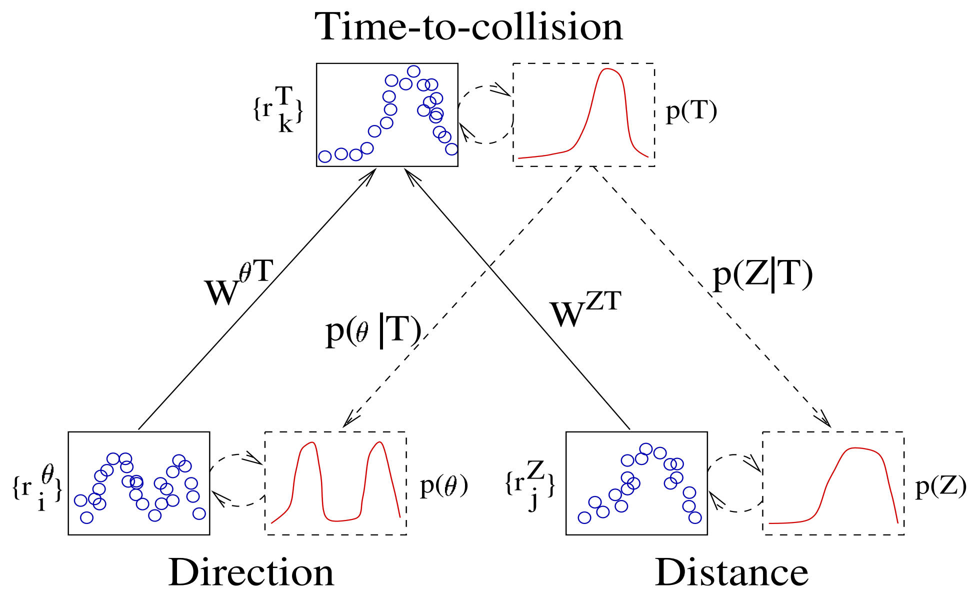

Figure 6: An example of a cortical belief network formulated to combine population code representations of probability distributions. The circular elements in the solid boxes depict explicit network components (neurons), while the curves in the dashed boxes depict implicit interpretations of those neural responses.

This model makes explicit predictions as to how information from different sources are combined as the values and certainty of the sources are manipulated. We are currently testing these predictions in psychophysical experiments on curvature judgments based on a combination of shading and texture cues. Preliminary results suggest that the model provides a natural explanation for a diverse set of phenomena, including cue enhancement and veto.

5 Discussion

The main hypothesis advanced in this chapter is that cortical populations encode distributions or likelihoods over variables. Standard approaches consider that, at any point in time, a population of neurons represents a single "best-guess" about a relevant variable or set of variables. An alternative novel interpretation of cortical activities is that they jointly specify likelihood across a range of values.

Different forms of such representations have been proposed, each with a corresponding method of extracting probability distributions from neural activities. A basis function approach (Anderson, 1995), and a standard Bayesian model are both capable of decoding distributions, but neither approach explicitly contains a description of how the population encodes uncertainty or multiple values. We have proposed a model, distributional population coding (DPC), that does contain an explicit decoding model, and formulates the decoding problem as the Bayesian inverse of this encoding model.

We then examined the performance of these various representational schemes to an example problem. Transparent motion provides an ideal paradigm for population coding of probability distributions: multiple values are encoded in each trial, making standard single-value interpretation inapplicable, and the combination of physiological data and behavioral data respectively specify neural activities and information presumably inferred from those activities. The basis function model and the standard Bayesian model both perform poorly in this paradigm. DPC, which treats neuronal responses and the animal's judgments as being sensitive to the entire distribution of an encoded variable, has been shown to be consistent with both single-cell responses and behavioral decisions, even matching subjects' threshold behavior. DPC also provides a novel perspective on another motion paradigm, suggesting that the judgement as to the single coherent motion direction amidst random motion is not hampered by decision noise but instead sensitive to the full array of motion directions present during the experimental trial.

We are currently applying this same model to several other motion experiments, including one in which subjects had to determine whether a motion stimulus consisted of a number of discrete directions or a uniform distribution (Williams et al., 1991). We are investigating whether our model can explain the nonmonotonic relationship between the number of directions and the judgments. Finally, experiments showing the effect of target uncertainty on population responses (Basso & Wurtz, 1998; Bastian et al,. 1998) are also handled naturally by the DPC approach.

DPC also provides a new approach to the issue of orientation tuning in primary visual cortex. It suggests that the representational capability of a V1 population should be expanded to include multiple values and uncertainty. A consequence of this view are specific predictions as to how a V1 population will respond to images containing ambiguous (low-contrast, noisy, and/or curved) edges, or containing multiple edges. We have developed a recurrent model of orientation tuning within the framework of the DPC approach, which is capable of encoding a range of stable activity profiles. This new model makes a number of testable predictions concerning V1 responses that differ from another recent model.

Finally, the current model is intended to describe the information available in a population at one stage in the processing stream. Extending the model to address how this information gets into the population, and the neural decoding and decision mechanisms subsequent to it, requires constructing a network of populations, each encoding distributions over different variables. A cortical belief network, introduced in the final section, is a method for carrying out probabilistic inference in populations of neurons.

References

[1] Anderson, C. H. (1994). Basic elements of biological computational systems. *International Journal of Modern Physics C*, *5*(2), 135–137.

[2] Anderson, C. H. (1995). Unifying perspectives on neuronal codes and processing. In *XIX International workshop on condensed matter theories*. Caracas, Venezuela.

[3] Basso, M. A. & Wurtz, R. H. (1998). Modulation of neuronal activity in superior colliculus by changes in target probability. *Journal of Neuroscience*, *18*(18), 7519-34.

[4] Bastian, A., Riehle, A., Erlhagen, W., & Schoner, G. (1998). Prior information preshapes the population representation of movement direction in motor cortex. *Neuroreport*, *9*(2), 315-319.

[5] Britten, K. H., Shadlen, M. N., Newsome, W. T., & Movshon, J. A. (1992). The analysis of visual

motion: A comparison of neuronal and psychophysical performance. *Journal of Neuroscience*, *12*(12), 4745–4765.

[6] Grunewald, A. (1996). A model of transparent motion and non-transparent motion aftereffects. In D. S. Touretzky, M. C. Mozer, & M. E. Hasselmo (Eds.), *Advances in Neural Information Processing Systems 8* (pp. 837–843). Cambridge, MA: MIT Press.

[7] Mather, G. & Moulden, B. (1983). Thresholds for movement direction: two directions are less detectable than one. *Quarterly Journal of Experimental Psychology, 35*, 513-518.

[8] Pouget, A. and Sejnowski, T. J. (1997). Spatial transformations in the parietal cortex using basis functions. *Journal of Cognitive Neuroscience*, 9(2):222–237.

[9] 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.

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

[11] Shadlen, M. N., Britten, K. H., Newsome, W. T., & Movshon, J. A. (1996). A computational analysis of the relationship between neuronal and behavioral responses to visual motion. *Journal of Neuroscience*, *16*(4), 1486–510.

[12] Simoncelli, E., Adelson, E., and Heeger, D. (1991). Probability distributions of optical flow. In *Proceedings 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 310–315.

[13] Simoncelli, E. and Heeger, D. (1994). A velocity representation model for MT cells. *Investigative Opthamology and Visual Science Supplement*, 35:1827.

[14] Snippe, H. P. (1996). Theoretical considerations for the analysis of population coding in motor cortex. *Neural Computation*, *8*(3):29–37.

[15] 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.

[16] Weiss, Y., & Adelson, T. (1998). Slow and Smooth: Combination of Local Motion Signals. *Vision Research*, *31*(2), 275–286.

[17] Williams, D., Tweten, S., & Sekuler, R. (1991). Using metamers to explore motion perception. *Vision Research*, *31*(2), 275–286.

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

[19] Zemel, R. S. & Dayan, P. (1999). Distributional population codes and multiple motion models. *Advances in Neural Information Processing Systems 11*, M. S. Kearns, S. A. Solla, & D. A. Cohn, eds., MIT Press, Cambridge, MA, 1999.

[20] Zemel, R. S. & Pillow, J. (1999). Encoding multiple orientations in a recurrent network. *Proceedings of the Computational Neuroscience Society, 8*, Kluwer Press, 1999.

[21] Zhang, K., Ginzburg, I., McNaughton, B. L., and Sejnowski, T. J. (1998). Interpreting neuronal population activity by reconstruction: A unified framework with application to hippocampal place cells. *Journal of Neurophysiology*, 79(2):1017–1044.