To appear in: *Neural Computation*, 10(2): 403-430 (1998).

Probabilistic Interpretation of Population Codes

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

Peter Dayan
Department of Brain & Cognitive Sciences
MIT
Cambridge, MA 02139
dayan@psyche.mit.edu

Alexandre Pouget
Georgetown Institute for Cognitive & Computational Sciences
Georgetown University
Washington, DC 20007-2197
alex@giccs.georgetown.edu

Abstract

We present a general encoding-decoding framework for interpreting the activity of a population of units. A standard population code interpretation method, the Poisson model, starts from a description as to how a single value of an underlying quantity can generate the activities of each unit in the population. In casting it in the encoding-decoding framework, we find that this model is too restrictive to describe fully the activities of units in population codes in higher processing areas, such as the medial temporal area. Under a more powerful model, the population activity can convey information not only about a single value of some quantity but also about its whole distribution, including its variance, and perhaps even the certainty the system has in the actual presence in the world of the entity generating this quantity. We propose a novel method for forming such probabilistic interpretations of population codes and compare it to the existing method.

Acknowledgements: We thank Charlie Anderson, Terry Sanger, and Larry Abbott for helpful discussions, and an anonymous reviewer for useful comments. R.Z. was supported by the McDonnell Foundation, grant JSMF 95-1; P.D. by NIMH R29 MH 55541-01 and the Surdna Foundation; and A.P. by a grant from the DOD, grant DAMD17-93-V-3018.

1 Introduction

Population codes, where information is represented in the activities of whole populations of units, are ubiquitous in the brain. There has been substantial work on how animals should and/or actually do extract information about the underlying encoded quantity (Georgopoulos, Schwartz, & Kettner, 1986; Baldi & Heiligenberg, 1988; Seung & Sompolinsky, 1993; Salinas & Abbott, 1994; 1995; Snippe, 1996; Sanger, 1996). With the exception of Anderson (1994), most of this work has concentrated on the case of extracting a *single value* for this quantity. In this paper, we study ways of characterizing the joint activity of a population as coding a whole *probability distribution* over the underlying quantity.

We will use two main motivating examples throughout this paper. They are both well-studied and classic examples of population codes. The first is place cells in hippocampus of freely moving rats (O'Keefe and Dostrovsky, 1971), which tend to fire when the animal is at a particular part of an environment. The second example is that of motion selective cells in the medial temporal (MT) area of monkeys that are reporting aspects of the motion in a stochastic stimulus made up of dots moving in various directions (Newsome, Britten, & Movshon, 1989). MT 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.

In these cases, treating the activity of such populations of cells as reporting on a single value of the variables they code (e.g., direction of motion) is inadequate. Instead, these are examples of two general situations in which the population must be interpreted as coding probability distributions over these variables:

1. *Insufficient information* exists to define a single value with certainty. The rat may not have enough information from visual or vestibular cues to know exactly where it is. This paper discusses how to make statistical sense of the obvious notion that if the animal is confused whether it is in place $\mathbf{x}_1$ or place $\mathbf{x}_2$, then the place cells that prefer both places should be activated.[1] Hinton (1992) pointed out that one should be able to use patterns of activity over such population codes to report not only a single place in the world, but also variance/uncertainty about that place and other aspects of a probability distribution.

2. *Multiple values* underlie the input. The monkey may have to report the direction of coherent motion embedded in a field of random motion noise. MT cells cannot be characterized as only reporting on the direction of coherent motion, since cells that prefer directions opposite to this are activated by the noise. The population must therefore be characterized as reporting something about the entire distribution of inputs.

In this paper, we first provide a general statistical framework that can be used to understand how the activity of a population of neurons can be considered as *encoding* information about the world, and, concomitantly, the way that this information can be *decoded* from the activity. We illustrate the framework by casting the standard model for population codes in its terms, use it to show why

[1] As far as we are aware, there is as yet no data on how the place cells really fire in such ambiguous cases.

this model is inadequate for representing probability distributions even though it does contain some distributional information, and also describe an existing model for probabilistic interpretation (Anderson & Van Essen, 1994; Anderson, 1994) in terms of the framework. In section 2 we define the encoding and decoding framework, and examine existing models for population codes in its terms. In section 3 we propose a novel method for encoding and decoding probability distributions from population codes. In section 4 we present empirical comparisons between the alternative methods.

2 Population Code Interpretations

2.1 The encoding-decoding framework

The starting point for almost all work on neural population codes is the neurophysiological finding that many neurons respond to a particular variable underlying a stimulus (such as the orientation of a visually presented bar) according to a unimodal tuning function. This function is read out as the mean firing rate of the cell, and is often reasonably well characterized as a Gaussian. The value of the underlying variable at which the peak of the tuning function occurs (the mean of the Gaussian) is called the preferred value for the cell. This form of response not only characterizes cells near the sensory periphery, but also cells that report the results of more complex processing, including receiving information from groups of cells that themselves have these tuning properties (in MT, for instance). A major caveat with almost all work on population codes, including that in this paper, is that the responses of cells actually vary with many quantities in the world other than the particular one that is usually studied. For example, many MT cells are also selective for disparity and can be affected by spatial frequency, making it difficult to ascribe weak firing to non-preferred motion or an incorrect disparity or whatever. In our theory, we will assume that we know the (collection of) variables with respect to which a cell's response is systematically determined; all other variables are treated as noise, so if the response depends on some unknown variable, interpretation will be hampered.

By definition, for each of these populations, the activities of the cells can be interpreted as conveying information about some underlying low-dimensional space. Interpreting population codes thus involves two spaces. First, there is an *explicit* space which consists of the activities $\mathbf{r} = \{r_i\}$ of the cells in the population. Second, these activities are described in terms of an *implicit* space (Zemel & Hinton, 1995), which contains the small number of underlying dimensions (*eg* direction in the MT example above) that the activities of the cells encode, and in which they are described as having tuning functions. The implicit space plays no explicit role, even though any processing on the basis of the activities $\mathbf{r}$ has to be referred to the implicit space. For instance, if the activities across the population are combined in some particular way (as in generating a single value), then the implicit space can be used to work out how much information is lost, and consequently to work out the best method of combination.

This notion of explicit and implicit spaces underlies our framework, which is depicted in Figure 1 in the context of a single experiment. At the top is the measured activities of a population of cells.

The two key questions to be asked about this firing are:

encoding: What is the relationship between the activities $\mathbf{r}$ of the cells and the underlying quantity $\mathbf{x}$ in the world that is represented?

decoding: What information about the quantity $\mathbf{x}$ can be extracted from the activities?

Although it is of active interest (Pouget, Zhang, Deneve, & Latham, 1998), in this paper we do not consider constraints that come from the neural instantiation of the decoding algorithms, and rather pose the decoding question as an abstract problem. Since neurons are generally noisy, it is often convenient to characterize encoding (Figure 1: operations A and B) in a probabilistic way, by specifying:

$$\mathcal{P}[\mathbf{r}|\mathbf{x}] \tag{1}$$

The simplest models make a further assumption of conditional independence of the different units given the underlying quantity:

$$\mathcal{P}[\mathbf{r}|\mathbf{x}] = \prod_i \mathcal{P}[r_i|\mathbf{x}] \tag{2}$$

although others characterize the degree of correlation between the units (e.g., Zohary, Shadlen, & Newsome, 1994).

If the encoding model in equation 1 is true, then a Bayesian decoding model specifies that the information $\mathbf{r}$ carries about $\mathbf{x}$ can be characterized precisely as:

$$\mathcal{P}[\mathbf{x}|\mathbf{r}] \propto \mathcal{P}[\mathbf{r}|\mathbf{x}]\mathcal{P}[\mathbf{x}] \tag{3}$$

where $\mathcal{P}[\mathbf{x}]$ is the prior distribution about $\mathbf{x}$ and the constant of proportionality is set so that

$$\int_{\mathbf{x}} \mathcal{P}[\mathbf{x}|\mathbf{r}]d\mathbf{x} = 1.$$

Note that starting with a deterministic quantity $\mathbf{x}$ in the world, encoding it in the firing rates $\mathbf{r}$, and decoding it (operation C) using equation 3 results in a probability distribution over $\mathbf{x}$. This uncertainty arises from the stochasticity represented by $\mathcal{P}[\mathbf{r}|\mathbf{x}]$. Given a loss function, we could then go on to extract a single value from this distribution (operation D).

For most real cases of population codes, encoding cannot be described so crisply. This paper describes the inadequacy of one particularly pervasive assumption, that the underlying quantity is a single value, for instance, the single position of a rat in an environment, or the single coherent direction of motion of a set of dots in a direction discrimination task. The assumption is pervasive since this is how one typically works out what a population of cells is encoding and how each responds to some particular $\mathbf{x}$. It is inadequate because it cannot capture the subtleties of other experiments, such as those in which rats can be made to be uncertain about their position (Cheng & Gallistel, 1984), or in which one direction of motion predominates yet there are several simultaneous motion directions (Newsome, Britten, & Movshon, 1989). In many cases, the natural characterization is actually a whole probability distribution $\mathcal{P}[\mathbf{x}|\omega]$ over the value of the variable $\mathbf{x}$, where ω represents all the available information. For instance, for the rat, this distribution might be the distribution over its possible locations in the environment.

The rat can clearly expect to be in exactly one position at any one time, and it therefore makes sense to consider the distribution of uncertainty $\mathcal{P}[\mathbf{x}|\omega]$ as to that position. This is not quite true for the monkey; it could be observing many different directions of motion simultaneously. In this paper, we characterize this multiplicity by considering a separation between the direction of motion of a single randomly chosen dot (which gives the equivalent of $\mathcal{P}[\mathbf{x}|\omega]$) and the actual number of dots present at any one time. Following Hinton (1992) we consider the sum total activity over the population as reporting the latter, and the distribution of that activity across the population as reporting the former. This imposes the important constraint that there is some independent standard for how much activity there should be (with which to work out the number of dots) and is clearly not the only possibility.[2]

Note also that there is no requirement that the animal perform decoding as in equation 3, or indeed that it explicitly perform decoding at all. That Wilson and McNaughton (1993) can extract the (x, y) coordinates of a rat in a room on the basis of the activities of 25 of its place cells does not mean that the stages of rodent processing subsequent to the hippocampus actually do this.

We can now cast two existing classes of proposals for population codes in terms of this framework.

2.2 The Poisson Model

Under the Poisson encoding model, the quantity encoded is indeed one particular value, and the activities of the individual units are independent, with the terms in equation 2 specified as:

$$\mathcal{P}[r_i|\mathbf{x}] = e^{-f_i(\mathbf{x})} \frac{(f_i(\mathbf{x}))^{r_i}}{r_i!} \tag{4}$$

The activity r_i could, for example, be the number of spikes the cell emits in a fixed time interval following the stimulus onset. A typical form for the tuning function $f_i(\mathbf{x})$ is Gaussian:

$$f_i(\mathbf{x}) \propto e^{-(\mathbf{x}-\mathbf{x}_i)^2/2\sigma^2} \tag{5}$$

about a preferred value $\mathbf{x}_i$ for cell i. In terms of Figure 1, turning the quantity $\mathbf{x}$ into a set of mean firing rates $f_i(\mathbf{x})$ for the units is operation A; sampling the activities r_i from these means according to a Poisson distribution is operation B. These operations are *descriptive models*; they capture the essence of the results of a collection of experiments rather than being based on a biophysical understanding of how $\mathbf{x}$ actually *causes* the activities $\mathbf{r}$.

[2]We propose that the magnitude of firing can be used to suggest the multiplicity of inputs as well as their properties (i.e., doubling the firing rate could indicate that perhaps there are two stimuli present). Then one could interpret the remaining pattern of activity as implying multiple distributions, one for each possible stimulus. It is likely, however, that nonlinear processes affect the firing rate under multiple stimuli. For example, Snowden (1989) showed that an MT cell's response to motion in its preferred direction can be suppressed by adding motion in an orthogonal direction.

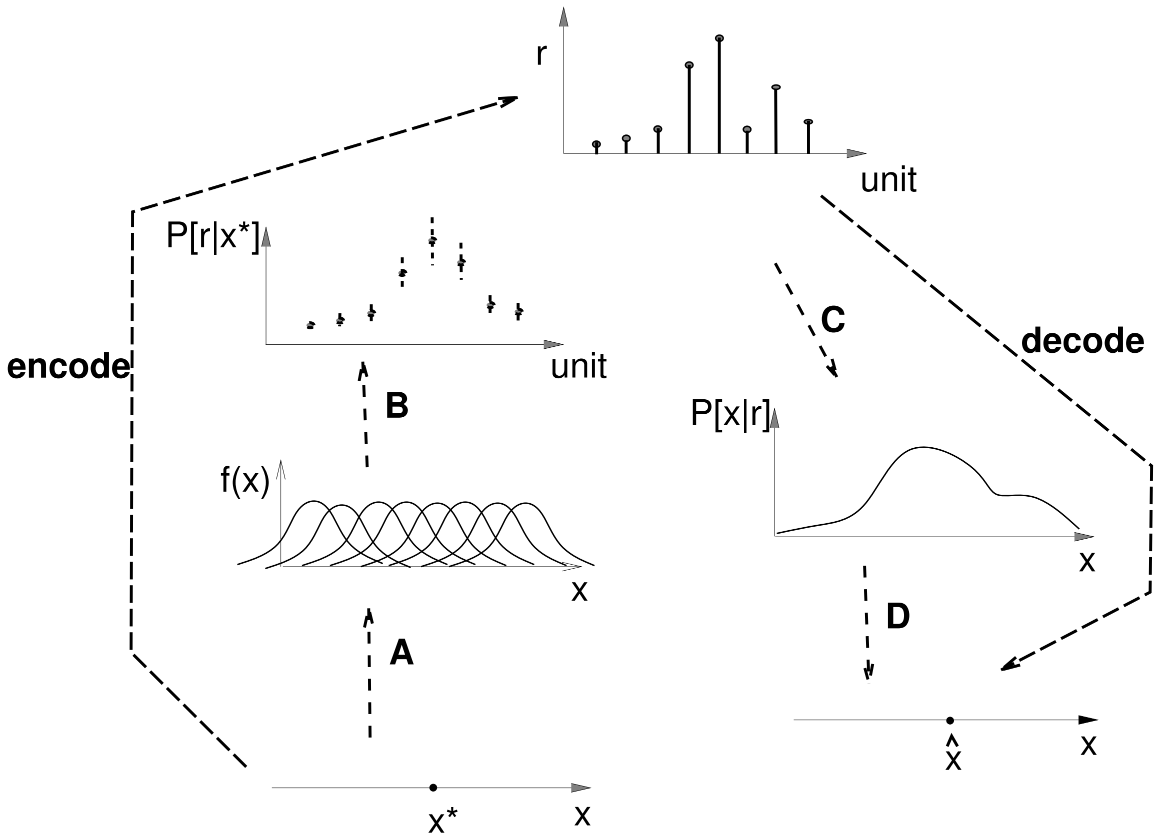

Figure 1: Left: Standard models of population codes, such as the Poisson model, assume an underlying encoding model such as the one illustrated here. The output of the encoding process is shown at the top: the activities of units, corresponding to the explicit, observable representation. These are assumed to have been generated by the independent application of each cell's tuning function and additive noise to the implicit representation shown at the bottom, which in this case is a single value $\mathbf{x}^*$ in the space of underlying variables. Right: Bayesian decoding models describe how to extract $\mathcal{P}[\mathbf{x}|\mathbf{r}]$ from the observed activities, through knowledge of the tuning functions $f(\mathbf{x})$. An estimate of the true value of $\mathbf{x}$ can then be formed according to some loss function. Capital letters label operations.

Several authors have examined maximum likelihood (ML) decoding under the Poisson encoding model (Seung & Sompolinsky, 1993; Salinas & Abbott, 1994; 1995; Snippe, 1996), and have analyzed the performance of other decoding methods relative to ML. These methods all focus on extracting a single value for the underlying parameter. The full probability distribution over $\mathbf{x}$ from this Poisson model is given by (Sanger, 1996; Földiák, 1993):

$$\mathcal{P}[\mathbf{x}|\mathbf{r}] \propto \mathcal{P}[\mathbf{x}] \prod_i e^{-f_i(\mathbf{x})} \frac{(f_i(\mathbf{x}))^{r_i}}{r_i!} \tag{6}$$

Although the Poisson model is simple and straightforward, it suffers from the assumption criticized above, namely that there is just a single value $\mathbf{x}$. If the rat is really uncertain about whether it is at location $\mathbf{x}_1$ or location $\mathbf{x}_2$ in the world, then the standard Poisson model has no formal way of turning that uncertainty into activities. In this case, we argued that the natural characterization of the quantity in the world that the activities of the cells encode is now $\mathcal{P}[\mathbf{x}|\omega]$.

We describe below a method of encoding that takes exactly this approach. However, one might argue that even though there is no formal way of encoding uncertainty in the activities, there is a formal way of decoding activities to produce a probability distribution over $\mathbf{x}$. Perhaps one could form $\mathcal{P}[\mathbf{x}|\mathbf{r}]$ in equation 6 as a decoding of $\mathbf{r}$ to model a diffuse or multimodal $\mathcal{P}[\mathbf{x}|\omega]$. We now show that this approach turns out to be inadequate.

Consider a one dimensional case with tuning functions as in equation 5. Imagine that activities $\mathbf{r}$ are specified directly in some manner on the basis of a whole probability distribution $\mathcal{P}[\mathbf{x}|\omega]$ over $\mathbf{x}$. The goal is to decode the activities $\mathbf{r}$ using equation 6 and actually represent $\mathcal{P}[\mathbf{x}|\omega]$ adequately. From now on, we will use $\hat{\mathcal{P}}^{\mathbf{r}}(\mathbf{x})$ as the decoding distribution over $\mathbf{x}$ that is specified by $\mathbf{r}$. This is no longer $\mathcal{P}[\mathbf{x}|\mathbf{r}]$ unless it is a true Bayesian inverse. From equation 6, and assuming a uniform prior over $\mathbf{x}$, we have:

$$\log \hat{\mathcal{P}}^{\mathbf{r}}(\mathbf{x}) = \mathcal{K} - \sum_i f_i(\mathbf{x}) - \frac{1}{2\sigma^2}\sum_i r_i(\mathbf{x}-\mathbf{x}_i)^2 \quad (7)$$

$$= \mathcal{K}' - \frac{1}{2}\left(\frac{\sum_i r_i}{\sigma^2}\right)\left(\mathbf{x} - \frac{\sum_i r_i \mathbf{x}_i}{\sum_i r_i}\right)^2 \quad (8)$$

by completing the square, if, as in most reasonable cases, there are sufficient units such that $\sum_i f_i(\mathbf{x})$ is constant in $\mathbf{x}$. But the distribution in equation 8 is then Gaussian, with mean μ and variance ξ^2, where:

$$\mu = \frac{\sum_i r_i \mathbf{x}_i}{\sum_i r_i} \qquad \xi^2 = \frac{\sigma^2}{\sum_i r_i}.$$

If we extract a single value from this decoded distribution by simply taking its mean, then this value matches the center-of-gravity estimate (Snippe, 1996). However, if we are interested in the entire distribution, then this standard model cannot capture the range of input distributions $\mathcal{P}[\mathbf{x}|\omega]$ under consideration. First, it will clearly not be possible to match multimodal $\mathcal{P}[\mathbf{x}|\omega]$ with any fidelity, since this decoded Gaussian is unimodal. Second, remembering that $r_i \geq 0$ should reasonably be integers, then unless they are all 0, the variance of $\hat{\mathcal{P}}^{\mathbf{r}}(\mathbf{x})$ is bound to be less than σ^2, and so there is *no* setting for the $\mathbf{r}$ which will match $\mathcal{P}[\mathbf{x}|\omega]$ having widths greater than that of the tuning functions themselves. Thus this Poisson model is incapable of representing distributions that are broader than the tuning functions.

In fact, even though it is not actually being used to specify the activities of the cells, the underlying assumption in the encoding of the Poisson model (as embodied in equation 4) is that there is just one value of $\mathbf{x}$ which results in the activities of the cells and thus the Poisson model has trouble encoding anything other than delta function $\mathcal{P}[\mathbf{x}|\omega]$. This analysis also applies to the Gaussian encoding model. It is not strictly true if the tuning functions are not Gaussian or the units have some baseline activity (Snippe, 1996; Pouget *et al*, 1998). We see later (Figure 9) a case in which allowing baseline activities permits the Poisson decoding model to produce a multimodal distribution. However, each of the modes is much too narrow. To reinforce this point: if the firing rate were stationary for long enough so that one could collect arbitrary numbers of spikes, then one could estimate the true mean activities $f_i(\mathbf{x})$ for the cells. Decoding using the Poisson model if one knows the means will almost always lead to a delta function estimate.

2.3 The KDE Model

Anderson (1994) and Anderson & Van Essen (1994) defined a new interpretation for population codes in which the notion of representing probability distributions over $\mathbf{x}$ rather than just a single value is essential. This method represents the distribution $\hat{\mathcal{P}}^{\mathbf{r}}(\mathbf{x})$ in terms of kernel density estimates (KDEs), forming 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:

$$\hat{\mathcal{P}}^{\mathbf{r}}(\mathbf{x}) = \sum_i r'_i \psi_i(\mathbf{x}) \tag{9}$$

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

$$r'_i = \frac{r_i}{\sum_j r_j}. \tag{10}$$

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. If the $\psi_i(\mathbf{x})$ are probability distributions, and so are positive, then the range of spatial frequencies in $\mathcal{P}[\mathbf{x}|\omega]$ that they can reproduce in $\hat{\mathcal{P}}^{\mathbf{r}}(\mathbf{x})$ is likely to be severely limited.

In terms of our framework, whereas the Poisson model makes decoding a corollary of (that is, the Bayesian inverse of) the encoding model, the KDE model makes encoding a corollary of the decoding model. Evaluating the KDE model requires us to consider encoding—taking a probability distribution $\mathcal{P}[\mathbf{x}|\omega]$ over $\mathbf{x}$ and producing a set of firing rates $\{r_i\}$ such that $\hat{\mathcal{P}}^{\mathbf{r}}(\mathbf{x})$ in equation 9 approximates $\mathcal{P}[\mathbf{x}|\omega]$ closely. It is the encoding process that standard experiments probe. Presenting single unambiguous stimuli (the experimental procedure that led to the characterization in the Poisson model) amounts to asking the system to encode delta function distributions $\mathcal{P}[\mathbf{x}|\omega] = \delta(\mathbf{x} - \mathbf{x}^*)$ for various $\mathbf{x}^*$. The measured responses are then the mean resulting activities $f_i(\mathbf{x}^*) \sim \langle r_i \rangle$.

One natural way to do encoding is to use the Kullback-Leibler divergence as a measure of the discrepancy between $\mathcal{P}[\mathbf{x}|\omega]$ and $\sum_i r'_i \psi_i(\mathbf{x})$ and use the expectation-maximization (EM) algorithm to fit the $\{r'_i\}$, treating them as mixing proportions in a mixture model (Dempster, Laird, & Rubin, 1987). This relies on $\{\psi_i(\mathbf{x})\}$ being probability distributions themselves.

The *projection method* (Anderson, 1994) is an alternative encoding scheme which does not require the iterations of EM but instead uses the $\mathcal{L}_2$ distance. This allows r_i to be computed as a projection of $\mathcal{P}[\mathbf{x}|\omega]$ onto the tuning functions:

$$r_i = \int_{\mathbf{x}} \mathcal{P}[\mathbf{x}|\omega] f_i(\mathbf{x}) \mathbf{dx} \tag{11}$$

The kernel functions are again assumed to be a fixed implicit property of the cells in this formulation and the optimal $\mathcal{L}_2$ tuning functions $f_i(\mathbf{x})$ are derived as:

$$f_i(\mathbf{x}) = \sum_j A_{ij}^{-1} \psi_j(\mathbf{x}) \quad ; \quad A_{ij} = \int_{\mathbf{x}} \psi_i(\mathbf{x}) \psi_j(\mathbf{x}) d\mathbf{x} \tag{12}$$

These tuning functions are likely to need smoothing or regularizing (Anderson, 1994), particularly if the $\psi_i(\mathbf{x})$ overlap substantially. In this case, with purely linear encoding and decoding operations, the overall scheme is a particular sort of linear filter, and many of its properties can be derived from this perspective.

There is a further aspect of $\mathcal{P}[\mathbf{x}|\omega]$ that we might wish to represent in a population code, namely *certainty*. Consider reducing the contrast of the moving random dots near to threshold. Then, the absolute activity levels of MT cells might represent the certainty that there is actually a stimulus at all. In this case, one might characterize $\mathcal{P}[\mathbf{x}|\omega]$ as a mixture model, with one mixing proportion for the absence of a stimulus, and one mixing proportion for a distribution over directions for the presence of a stimulus. The normalization step in equation 10 prevents the KDE from representing this form of certainty, since $\sum_i r'_i = 1$. However, certainty could easily be captured. For instance, if there is a maximum value $R_{\max}$ for the summed actual activities $\sum_i r_i$, then one could have an indicator variable $z \in \{0, 1\}$ representing the presence (1) or absence (0) of the underlying object and:

$$\begin{aligned}
\mathcal{P}[z = 1|\{r_i\}] &= \textstyle\sum_i r_i / R_{\max} \\
\hat{\mathcal{P}}(\mathbf{x}; \{r_i\}, z) &= \begin{cases} \mathcal{P}[\mathbf{x}] & \text{if } z = 0 \\ \displaystyle\sum_i \frac{r_i}{\sum_j r_j} \psi_i(\mathbf{x}) & \text{if } z = 1 \end{cases}
\end{aligned}$$

where $\mathcal{P}[\mathbf{x}]$ is the prior distribution over $\mathbf{x}$. Note that under this formulation, as the probability of the underlying object's presence approaches zero, the marginalized estimate of the original distribution approaches the prior:

$$\hat{\mathcal{P}}^{\mathbf{r}}(\mathbf{x}) = \hat{\mathcal{P}}(\mathbf{x}; \{r_i\}, z = 1) \mathcal{P}[z = 1|\{r_i\}] + \hat{\mathcal{P}}(\mathbf{x}; \{r_i\}, z = 0)(1 - \mathcal{P}[z = 1|\{r_i\}])$$

Of course, as $\mathcal{P}[z = 1|\{r_i\}]$ tends to 0, the question of the true underlying distribution becomes moot.

3 The Extended Poisson Model

Given its linear decoding method, and a limited number of decoding kernel functions $\psi_i(\mathbf{x})$, we might expect the KDE model to have difficulty capturing in $\hat{\mathcal{P}}^{\mathbf{r}}(\mathbf{x})$ probability distributions $\mathcal{P}[\mathbf{x}|\omega]$

that have high frequencies, such as delta functions. We also saw that the standard Poisson model has the problem of decoding almost any pattern of activities $\mathbf{r}$ into something that rapidly approaches a delta function as the activities increase. Is there any middle ground?

We argued that the problem for the standard Poisson model comes from its encoding model (equation 4), which is based on there being a single underlying value $\mathbf{x}$. We can extend this encoding model to allow the recorded activities $\mathbf{r}$ to depend explicitly on general $\mathcal{P}[\mathbf{x}|\omega]$. The *extended Poisson model* is based on an encoding model in which the activity of cell i is Poisson about a mean, which, in the continuous version is:

$$\langle r_i \rangle = \int_{\mathbf{x}} \mathcal{P}[\mathbf{x}|\omega] f_i(\mathbf{x}) d\mathbf{x}. \tag{13}$$

Note that this equation is identical to the encoding model for the kernel density model (equation 11), except that here the firing rates are stochastic. In the kernel density model, there is no variability in the activities $\{r_i\}$ that encode a particular $\mathcal{P}[\mathbf{x}|\omega]$. Under the extended Poisson model, the activity of each cell is a filtered, sampled version of the underlying probability distribution, and this implies strong constraints on the potential quality of reconstructions.

This model is the most straightforward extension of the conventional Poisson model, and makes roughly similar predictions about observable activities when $\mathcal{P}[\mathbf{x}|\omega]$ is a delta function. However, it offers a much more powerful model for representing $\mathcal{P}[\mathbf{x}|\omega]$ that are not delta functions.

Given $\{r_i\}$ generated using Poisson noise from equation 13, how should one infer $\mathcal{P}[\mathbf{x}|\omega]$? Recall that for the standard Poisson model, encoding a single location in a population ($\mathcal{P}[r_i|\mathbf{x}]$) leads, through decoding, to a probability distribution over possible locations ($\mathcal{P}[\mathbf{x}|\mathbf{r}]$). Here the encoding model takes a whole function ($\mathcal{P}[\mathbf{x}|\omega]$) and stochastically produces a set of numbers ($\{r_i\}$) which provide partial information about that function. The general Bayesian inverse to this is a model that takes the numbers and produces a probability distribution over the functions that could have generated them. In our case, this means that decoding should really produce a probability distribution over probability distributions over the implicit variable $\mathbf{x}$, that is, $\mathcal{P}[\mathcal{P}[\mathbf{x}|\omega]|\mathbf{r}]$. Rather than do this, we choose to summarize this distribution over distributions by an approximation to its most likely member; we perform an approximate form of maximum a posteriori (MAP) decoding, not in the value of $\mathbf{x}$, but in distributions over $\mathbf{x}$. Figure 2 illustrates the entire encoding and decoding process for the case of implicit probability distributions.

We approximate $\mathcal{P}[\mathbf{x}|\omega]$ as a piece-wise constant histogram which takes the value $\hat{\phi}_j$ in $(\mathbf{x}_j, \mathbf{x}_{j+1}]$, and $f_i(\mathbf{x})$ by a piece-wise constant histogram that take the values f_{ij} in $(\mathbf{x}_j, \mathbf{x}_{j+1}]$. Then we model activities $\{r_i\}$ as being independent Poisson random variables whose means are (*cf* equation 13):

$$\langle r_i \rangle = \sum_j \hat{\phi}_j f_{ij}$$

Then, the true inverse distribution is:

$$\mathcal{P}\left[\{\hat{\phi}_j\}|\{r_i\}\right] \propto \mathcal{P}\left[\{\hat{\phi}_j\}\right] e^{-\sum_{ij} \hat{\phi}_j f_{ij}} \prod_i \left[\sum_j \hat{\phi}_j f_{ij}\right]^{r_i} \tag{14}$$

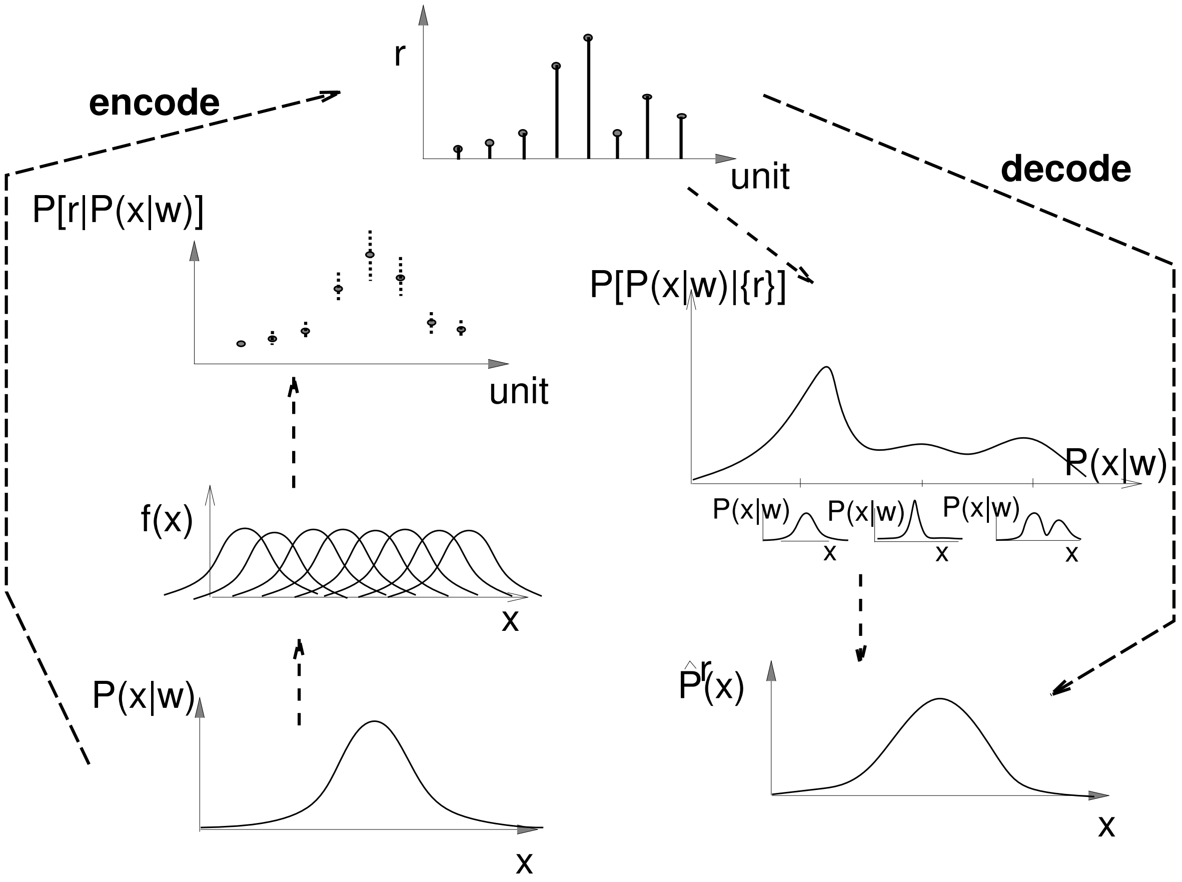

Figure 2: A set of firing rates may also be interpreted as encoding a probability distribution in implicit space. Decoding the rates now involves forming a probability distribution over possible implicit distributions, $\mathcal{P}\left[\mathcal{P}[\mathbf{x}|\omega]|\mathbf{r}\right]$. The decoding distribution $\hat{\mathcal{P}^{\mathbf{r}}}(\mathbf{x})$ may be formed from this distribution over distributions by integrating or through a maximum *a posteriori* computation. The extended Poisson model forms $\hat{\mathcal{P}^{\mathbf{r}}}(\mathbf{x})$ using an approximate form of maximum likelihood in distributions over $\mathbf{x}$.

---

where $\mathcal{P}\left[\{\hat{\phi}_j\}\right]$ is the prior over the $\{\hat{\phi}_j\}$.

If the system is translation invariant, then $\sum_i f_{ij} = f$ is constant for all j. Also $\sum_j \hat{\phi}_j = 1$, because it represents a probability distribution, therefore $e^{-\sum_{ij}\hat{\phi}_j f_{ij}} = e^{-f}$, which is constant. Taking logs,

$$\log \mathcal{P}\left[\{\hat{\phi}_j\}|\{r_i\}\right] = \mathcal{K} + \log \mathcal{P}\left[\{\hat{\phi}_j\}\right] + \sum_i r_i \log\left[\sum_j \hat{\phi}_j f_{ij}\right] \tag{15}$$

where $\mathcal{K}$ is a constant. Maximum a posteriori decoding in this context requires finding the set of $\{\hat{\phi}_j\}$ that sum to 1 and maximize this expression. If $\log \mathcal{P}\left[\{\hat{\phi}_j\}\right]$ is dominated by a smoothness prior such as

$$\sum_j \left(\hat{\phi}_j - \hat{\phi}_{j+1}\right)^2 \tag{16}$$

then we are left with the following expression:

$$AP(\{\hat{\phi}_j\}) = \sum_i r_i \log\left[\sum_j \hat{\phi}_j f_{ij}\right] - \epsilon \sum_j \left(\hat{\phi}_j - \hat{\phi}_{j+1}\right)^2 \tag{17}$$

where ϵ is a weighting coefficient on the smoothness prior.

Thus the extended Poisson model creates a decoding distribution $\hat{\mathcal{P}^{\mathbf{r}}}(\mathbf{x})$ that approximates MAP inference from the distribution over distributions $\mathcal{P}\left[\mathcal{P}[\mathbf{x}|\omega]|\{r_i\}\right]$.

The values of this decoding distribution can be found in a number of ways. One simple method involves adjusting $\{\hat{\phi}_j\}$ via simple gradient ascent in $AP(\{\hat{\phi}_j\})$. Alternatively, one can use a version of EM to find the MAP values. For this, one interprets $\sum_j \hat{\phi}_j f_{ij}$ as a mixture model for case i, where $\{\hat{\phi}_j\}$ are the mixing proportions and $\{f_{ij}\}$ are the fixed values of the underlying distributions. r_i is then the weighting for case i, and the resulting expression in equation 15 therefore acts just like a likelihood. In practice a smoothness prior such as equation 16 is required for regularization. For the experiments described below, we implemented a crude approximation by averaging adjacent $\hat{\phi}_j$ after each EM iteration.

With this method of decoding in mind, we now see how the extended Poisson model competes with the KDE model as a way of representing probability distributions. By comparison with the linear equation 9, equation 17 offers a *non-linear* way of combining a set of activities $\{r_i\}$ to give a probability distribution $\hat{\mathcal{P}^{\mathbf{r}}}(\mathbf{x})$ over the underlying variable $\mathbf{x}$. The computational complexities of equation 17 are irrelevant, since decoding is only an implicit operation that the system need never actually perform.

Another way of looking at the difference between the models is that the extended Poisson model is a Bayesian decoding method, and thus involves a multiplication of tuning functions (assuming the units are independent), while the KDE model is a basis function method, in which decoding entails a summation of kernels. We will see the consequences of this difference in the simulations below.

Finally, just as for the KDE model, this form of decoding does not capture well the certainty in the presence of a stimulus. A natural extension to the model in equation 17 is to estimate $\hat{\phi}_j$ just as before, but set the approximation to $\hat{\mathcal{P}^{\mathbf{r}}}(\mathbf{x})$ to be $\hat{\phi} \times \hat{\phi}_j$ in the interval $(\mathbf{x}_j, \mathbf{x}_{j+1}]$, where

$$\hat{\phi} = \min\left[1, \frac{\sum_i r_i}{\sum_{i,j} f_{ij}}\right] \tag{18}$$

4 Comparing the Models

We now compare the ability of these interpretation methods to extract the probabilistic information contained in a population code. The primary question of interest is whether we can find a set of deterministic activities $\{r_i\}$ that make $\hat{\mathcal{P}}^{\mathbf{r}}(\mathbf{x})$ a close approximation to an interesting implicit distribution $\mathcal{P}[\mathbf{x}|\omega]$. We take three instances of implicit distributions intended to model the primary cases in which interpreting a population code as a probability distribution is essential:

1. Some uncertainty exists about the location of the target. The aim is to approximate correctly the mean and the uncertainty about this mean.
2. The target could be in one of two locations. A bimodal distribution could arise due to insufficient information, as in the case of the rat, or due to the actual existence of multiple values.
3. Uncertainty exists about the presence of the target. Here the magnitude of the integral under the implicit distribution is intended to capture the degree of certainty in the presence of the target.

Finally, we also examine the noise robustness of the methods.

In each of these cases, we consider a simplified situation in which the location of the object only varies along one dimension. This eases visualization of network activity and decoded distributions, but each model readily could be extended to include other dimensions. For each model, the population code consisted of 50 units, with preferred values $\mathbf{x}_i$ spaced uniformly in the range of $\mathbf{x} = [-10., 10.]$. Associated with each unit was a Gaussian distribution: $\mathcal{N}(\mathbf{x}_i, \sigma = 0.3)$. In the KDE model, these 50 Gaussians were the kernels $\psi_i(\mathbf{x})$, while in the Poisson and extended Poisson models they represented the tuning functions $f_i(\mathbf{x})$.

For the KDE model, we used two methods to find the set of activities that provided the best fit between the true implicit distribution $\mathcal{P}[\mathbf{x}|\omega]$ and the estimate $\hat{\mathcal{P}}^{\mathbf{r}}(\mathbf{x})$ (equation 9). The projection method computes the rates $\{r_i\}$ as a projection of $\mathcal{P}[\mathbf{x}|\omega]$ onto the tuning functions, where the optimal tuning functions are derived from the fixed kernels (see equations 11 and 12). The EM method adapts $\{r_i\}$ to minimize the Kullback-Leibler divergence between the estimated and true distributions.

For the Poisson and extended Poisson models, we again computed the rates $\{r_i\}$ as a projection of $\mathcal{P}[\mathbf{x}|\omega]$ onto the tuning functions, where the tuning functions are now the fixed $\{f_i(\mathbf{x})\}$. In the Poisson model we decode directly into the maximum likelihood values of equation 8, while in the extended Poisson model we decode using a version of EM to find the MAP values of equation 17. In all of these experiments, we limited the activities $\{r_i\}$ to assume integer values. Note that in the case of the KDE model, these activities are then normalized as part of the decoding process (equation 10). Finally, in all simulations presented here, $R_{\max} = 50$, and the number of histogram bins in the Extended Poisson model was 500. Table 1 summarizes these methods.

Operation	Extended Poisson	KDE (Projection)	KDE (EM)
Encode $\langle r_i \rangle$	$\langle r_i \rangle = h\left[\int_{\mathbf{x}} \mathcal{P}[\mathbf{x}\vert\omega] f_i(\mathbf{x}) d\mathbf{x}\right]$ $f_i(\mathbf{x}) = R_{\max}\mathcal{N}(\mathbf{x}_i, \sigma)$	$\langle r_i \rangle = h\left[R_{\max}\int_{\mathbf{x}} \mathcal{P}[\mathbf{x}\vert\omega] f_i(\mathbf{x}) d\mathbf{x}\right]$ $f_i(\mathbf{x}) = \sum_j A_{ij}^{-1}\psi_j(\mathbf{x})$ $A_{ij} = \int_{\mathbf{x}} \psi_i(\mathbf{x})\psi_j(\mathbf{x}) d\mathbf{x}$	$\langle r_i \rangle = h\left[R_{\max} r'_i\right]$ r'_i to max. L
Decode $\hat{\mathcal{P}}^{\mathbf{r}}(\mathbf{x})$	$\hat{\mathcal{P}}^{\mathbf{r}}(\mathbf{x})$ to max. L $\hat{r}_i = \int_x \hat{\mathcal{P}}^{\mathbf{r}}(\mathbf{x}) f_i(\mathbf{x}) d\mathbf{x} \approx \sum_j \phi_j f_{ij}$	$\hat{\mathcal{P}}^{\mathbf{r}}(\mathbf{x}) = \sum_i r'_i \psi_i(\mathbf{x})$ $r'_i = r_i / \sum_j r_j$	$\hat{\mathcal{P}}^{\mathbf{r}}(\mathbf{x}) = \sum_i r'_i \psi_i(\mathbf{x})$
Likelihood	$L = \log \mathcal{P}\left[\{\hat{\phi}_j\} \vert \{r_i\}\right] \approx \sum_i r_i \log \hat{r}_i$		$L = \int_{\mathbf{x}} \mathcal{P}[\mathbf{x}\vert\omega] \log \hat{\mathcal{P}}^{\mathbf{r}}(\mathbf{x}) d\mathbf{x}$
Error	$G = \sum_i r_i \log(r_i / \hat{r}_i)$	$E = \int_{\mathbf{x}} \left[\hat{\mathcal{P}}^{\mathbf{r}}(\mathbf{x}) - \mathcal{P}[\mathbf{x}\vert\omega]\right]^2 d\mathbf{x}$	$G = \int_{\mathbf{x}} \mathcal{P}[\mathbf{x}\vert\omega] \log \frac{\mathcal{P}[\mathbf{x}\vert\omega]}{\hat{\mathcal{P}}^{\mathbf{r}}(\mathbf{x})} d\mathbf{x}$

Table 1: A summary of the key operations with respect to the framework of the three interpretation methods compared here. $h[]$ is a rounding operator to ensure integer firing rates, and $\psi_i(\mathbf{x}) = \mathcal{N}(\mathbf{x_i}, \sigma)$ are the kernel functions for the KDE method. The certainty calculation is not included in this table.

We also require some quantitative way of comparing the reconstructions of the various models. Although the Kullback-Leiber distance (the implicit metric for both extended Poisson and KDE-EM methods) is the natural measure of the difference between two probability distributions, it cannot be used here, since the reconstructions from the KDE-Projection method are not strict distributions (they are not non-negative). We therefore used the somewhat less informative squared error measure:

$$E = \sum_j \left[\hat{\mathcal{P}}^{\mathbf{r}}(\mathbf{x}_j) - \mathcal{P}[\mathbf{x}_j \vert \omega]\right]^2 \tag{19}$$

4.1 Uncertainty in target location

For these simulations, the implicit distribution is a Gaussian:

$$\mathcal{P}[\mathbf{x}|\omega] = \mathcal{N}(0, \tau)$$

All three methods form good estimates for a range of values of τ, the width of the true implicit distribution (for example, see Figure 3). However, as predicted, both KDE methods are unable to accurately represent narrow distributions, e.g., $\tau = 0.2$ (Figure 4). In general, the critical variable in the fidelity of the KDE estimate is the ratio of the widths σ of the decoding kernels and the

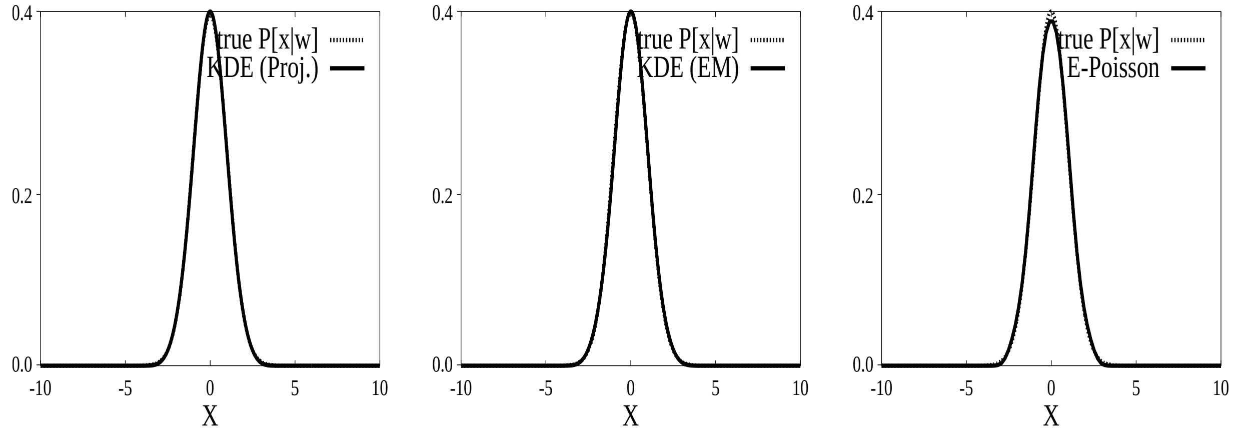

Figure 3: All three methods can reconstruct the original implicit Gaussian distribution over a range of values of τ, the standard deviation of this Gaussian. Here $\tau = 1.0$.

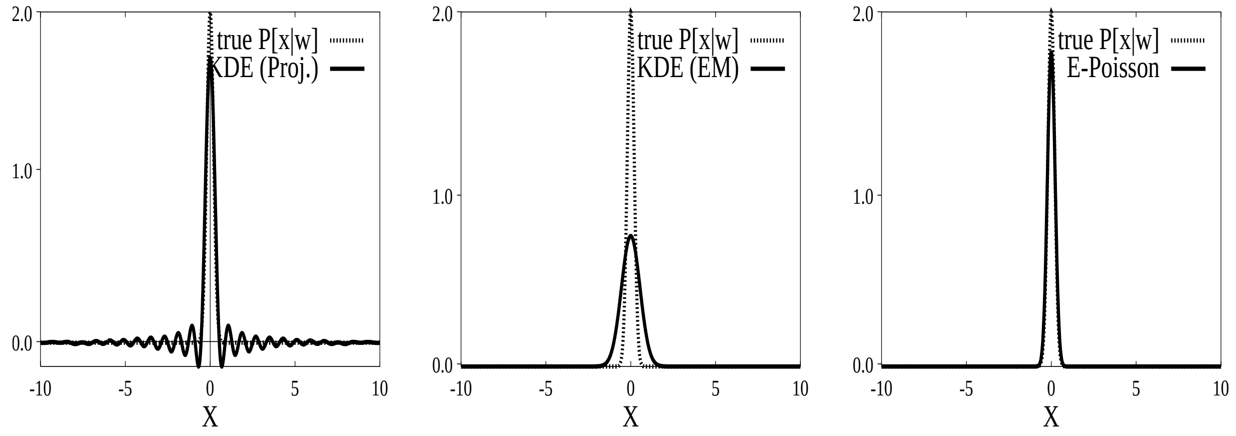

Figure 4: The KDE method has difficulty capturing the high frequency information in the implicit Gaussian distribution when its standard deviation $\tau = 0.2$ is smaller than the kernel widths ($\sigma = 0.3$).

widths τ in the true implicit distribution. The extended Poisson method is able to capture both narrow and wide implicit distributions, so it can form accurate interpretations of distributions with a relatively high variance as well as delta functions.

4.2 Multiple locations

A multimodal implicit distribution can be described as a mixture of two Gaussians:

$$\mathcal{P}[\mathbf{x}|\omega] = 0.5 * \mathcal{N}(\mathbf{x}_1, \tau) + 0.5 * \mathcal{N}(\mathbf{x}_2, \tau)$$

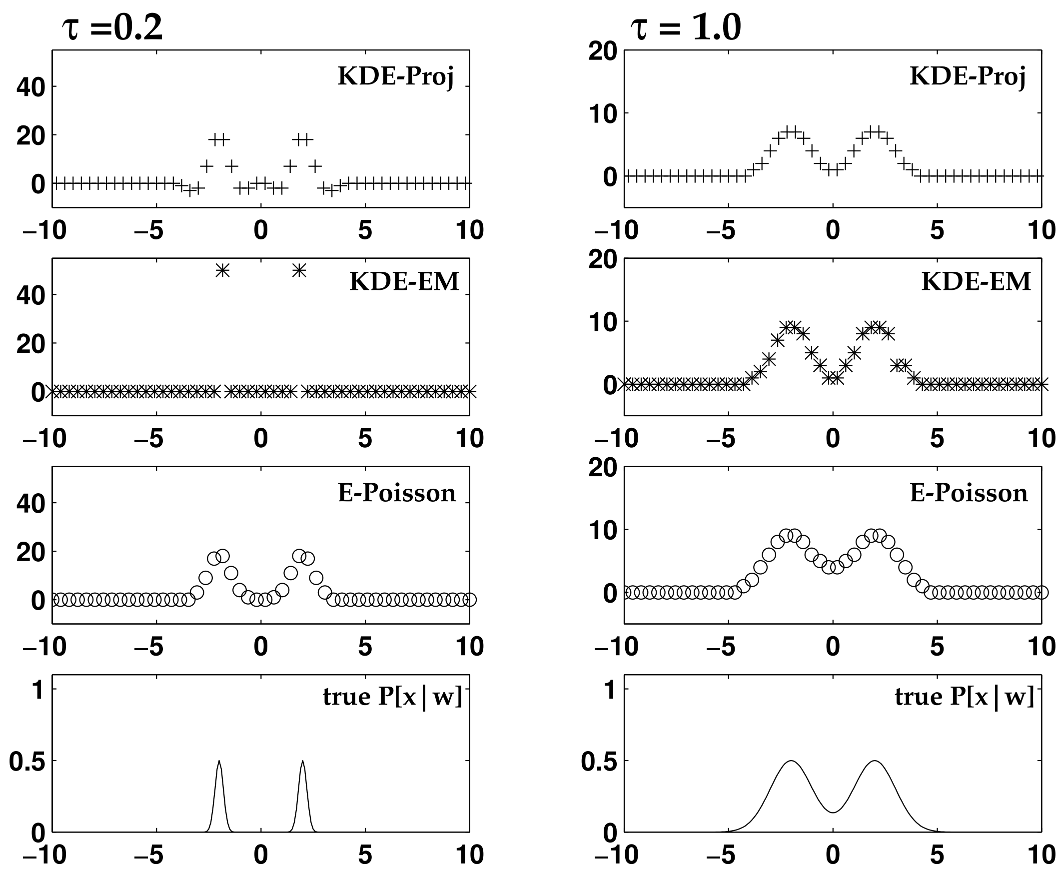

Figure 5: Each method's encoding model specifies how the expected firing rates of units in the population are based on the implicit distribution. The implicit distribution here is a bimodal Gaussian with standard deviation τ. The firing rates on the left are for $\tau = 0.2$, while $\tau = 1.0$ on the right. In both cases, $R_{\max} = 50$.

Here we can model the situation in which the rat is uncertain whether the target is in location $\mathbf{x}_1$ or $\mathbf{x}_2$. The variable τ describes an additional degree of uncertainty about the exact location of each potential target. For these simulations, we let $\mathbf{x}_1 = 2$ and $\mathbf{x}_2 = -2$.

To get a sense of the different encoding models, the expected activities of the 50 units in the population code are plotted in Figure 5. For $\tau = 1.0$, The KDE-EM method has converged to a local minimum, which is why one of the firing rates looks odd. However, because the distribution being modeled is so broad, this makes very little difference to the net quality of reconstruction.

Applying the respective decoding models to these expected firing rates, we find results similar to those of the previous section. Both KDE methods can capture a bimodal implicit distribution where the width exceeds that of the kernels (Figure 6), yet they cannot accurately reconstruct

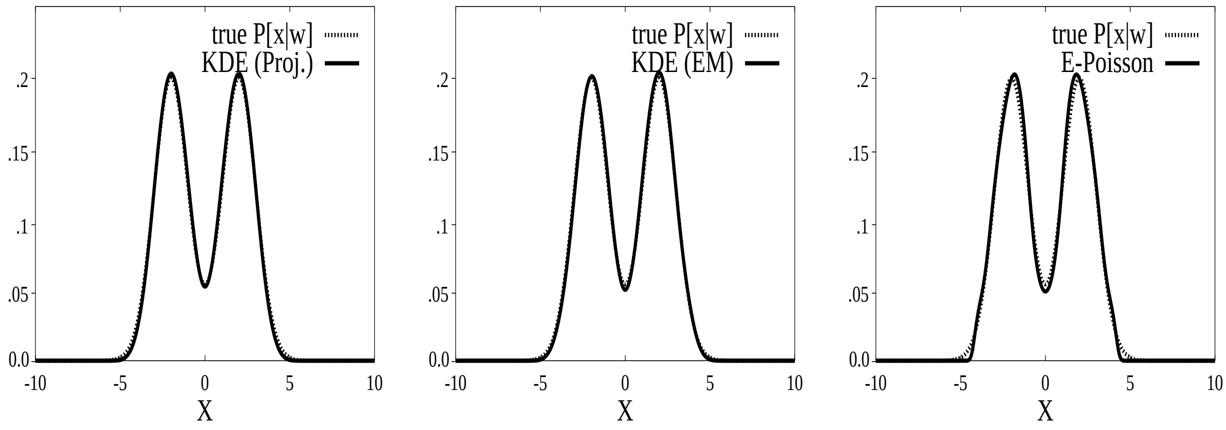

Figure 6: All three methods provide a good fit to the bimodal Gaussian distribution when its variance is sufficiently large ($\tau = 1.0$).

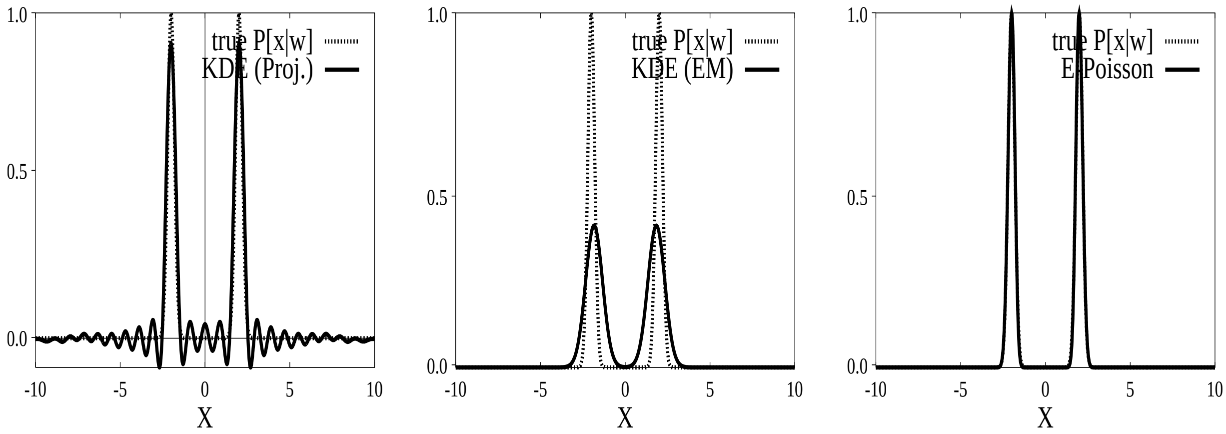

Figure 7: The KDE method again has difficulty capturing the high frequency information in the implicit bimodal Gaussian distribution when its variance $\tau = 0.2$ is smaller than the kernel widths ($\sigma = 0.3$).

narrow distributions (Figure 7). The extended Poisson model matches both types of implicit distributions with high fidelity.

For the sake of comparison, Figure 8 shows the decoded distribution formed by the simple Poisson model (equation 8). Here we use the same encoding model as in the extended Poisson method, in order to allow the unit activities to convey information about the entire implicit distribution. Nonetheless, the simple decoding model always produces a delta function estimate. An estimated delta function fits the case of the narrow unimodal implicit distribution shown in Figure 4, but cannot match the bimodal implicit distribution here.

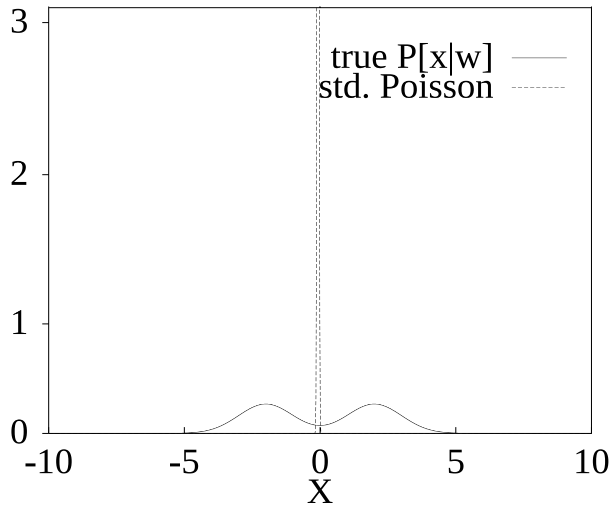

Figure 8: The simple Poisson decoding model leads to a delta function for the estimated implicit distribution. Here the implicit distribution is the bimodal Gaussian, $\tau = 1.0$. The unit activities $\mathbf{r}$ are the same as for the extended Poisson model, and the decoding is done using equation 8.

This result is predicted based on the the analysis presented in section 2.2. This analysis, however, applied to the case of Gaussian tuning functions in the absence of any baseline activity. Using simulations, we examine what happens when the units have some baseline activity. Here it is necessary to change the encoding model, so that the unit's expected firing rate is the sum of a Gaussian tuning function and a baseline firing rate, R_b:

$$f_i(\mathbf{x}) \propto e^{-(\mathbf{x}-\mathbf{x}_i)^2/2\sigma^2} + R_b \tag{20}$$

Again we form the decoding distribution as the Bayesian inverse of this new encoding model

$$\mathcal{P}[\mathbf{x}|\mathbf{r}] \propto \mathcal{P}[\mathbf{x}] \prod_i e^{-f_i(\mathbf{x})} \frac{(f_i(\mathbf{x}))^{r_i}}{r_i!}$$

For the simulations, we again use the encoding model of the extended Poisson model, but now use this modified simple Poisson model for decoding. An example of the results is shown in Figure 9. The decoded distribution is able to take on multiple modes, yet it is always highly peaked, due to the likelihood being a product of Poisson probabilities. Thus this decoding method can recover multiple implicit values, but cannot capture any uncertainty in these values.

4.3 Uncertainty in object presence

The next set of simulations addresses the issue of object presence. We use a dampened Gaussian to model the situation in which uncertainty exists in the actual presence of the target:

$$c\mathcal{N}(0, \tau), 0 \leq c \leq 1$$

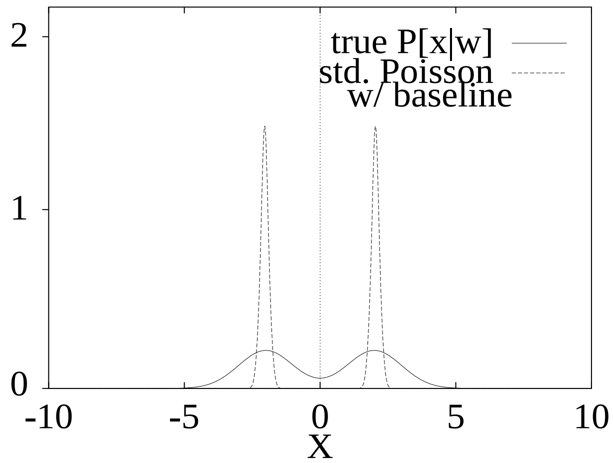

Figure 9: When we extend the simple Poisson decoding model to include baseline firing rates ($R_b = 5.0$ here), the estimated implicit distribution can be multimodal. Yet this decoded distribution does not contain any variance about these values. The implicit distribution is again the bimodal Gaussian, $\tau = 1.0$, and the unit activities $\mathbf{r}$ are the same as for the extended Poisson model.

Ideally, c would not be restricted to be less than 1. Instead it could take on any positive value, and thereby describe the actual number of instances of the implicit variable. Here we consider the simpler situation in which only one instance of the object is represented.

In these simulations, we compare the KDE model using the projection method to the extended Poisson model. We set τ to be 1.0, because both methods are able to match this distribution with high fidelity when $c = 1.0$. The primary empirical result is that while both methods have difficulty as the presence c gets close to 0, both are able to recover a variety of non-normalized Gaussian implicit distributions, as shown in Figure 10. The main reason for the poor performance as c decreases is that the firing rates are forced to be integers. Clearly, increasing $R_{\max}$ would extend the range of good reconstructions.

4.4 Noise robustness

In the previous simulations, the activities of the units in the population code were equal to their expected responses according to the respective encoding models. We now examine the robustness of the interpretation methods by treating the unit activities as Poisson random variables.

Each simulation involves a stochastic sampling of the unit responses. The robustness of a method is estimated by computing the expected error in the decoding distribution with respect to the true implicit distribution, that is, averaging the squared-error metric (equation 19) over 50 stochastic

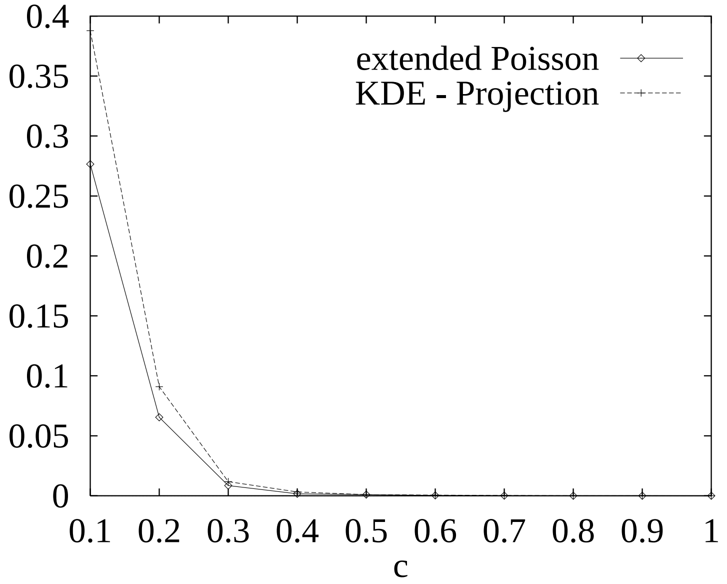

Figure 10: This plot shows the squared error (equation 19) for different values of $\mathbf{c}$, the total integral under the implicit distribution, for both the extended Poisson and KDE projection methods. In these simulations, $R_{\max} = 50$.

---

trials.

The results of this set of simulations again match the predictions based on the contrast between the methods, as shown in Figure 11. Because the decoding in the Extended Poisson model in nonlinear, we predict that this model will be more robust to noise than the KDE model, in which decoding is linear.

This figure makes it appear that both methods produce perfect reconstructions for larger values of τ. This is in fact not true, and is largely due to the magnitude of the expected error for small values of τ. For the Extended Poisson model, the regularization removes one component of the error: inaccuracies in the shape of the reconstructed distribution. This process is more effective as τ gets larger. However, the smoothing does not remove the second component: the *bias* that is present if the centers of the two Gaussians in the reconstructed distribution are incorrect. This bias component of the error is slightly higher for low values of τ, but relatively constant and non-zero for $\tau > 0.4$.

5 Discussion

We have presented a theoretical framework for understanding population codes which generalizes naturally to the important case in which the population provides information about a whole probability distribution over an underlying quantity rather than just a single value. We used the

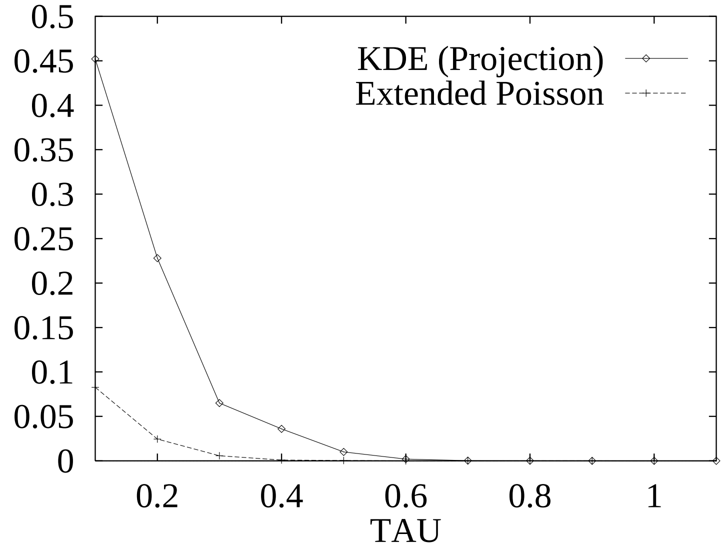

Figure 11: The expected error $< E >$ in the decoded distribution, averaged over 50 samples from the firing rate distributions, is plotted against different values of τ, the width of the bimodal Gaussian implicit distribution. Note that the results for only one of the KDE methods is shown because both KDE methods involve the same decoding operation.

---

framework to analyze two existing models, and to suggest and evaluate a third model for encoding such probability distributions.

More informally, we have tried to examine the consequences of the seemingly obvious step of saying that if a rat, for instance, is uncertain about whether it is at one of two places, then place cells representing both places could be activated. The complications come because the structure of the interpretation changes; for instance, one can no longer think of maximum likelihood methods to extract a single value from the code directly.

We are not suggesting that the uncertainty is generated at any particular step in the processing system. Rather, it is a type of information that is potentially contained in the population activity, about which inferences can be made from one level to the next. So the rat need not be aware that he is uncertain; we are not positing anything about 'conscious' uncertainty. Instead different firing patterns corresponding to different distributions over the underlying implicit variables are all that is needed to infer the uncertainty at the level above. This leads to the prediction that place cells corresponding to multiple locations will be active when the rat is uncertain as to its location. This uncertainty may be present even at the sensor level: a population of orientation selective cells should have a different pattern of firing to a bar at one orientation and a blurred image of the same bar in which the orientation is "uncertain". This fuzzier bar (as well as the sharper bar) can be represented in terms of a probability distribution over orientation.

One main result of our framework is a method for encoding and decoding probability distributions that is the natural extension of the standard Poisson model for encoding and decoding single values. We also showed that this standard model is inadequate for coding probability distributions. Under the new encoding model, the activity of a cell has Poisson statistics about a mean which is dependent on the integral of the whole encoded probability distribution, weighted by the tuning function of the cell. The behavior of this encoding model is appropriate in circumstances such as the positional uncertainty of the rat. We suggested a particular decoding model, based on an approximation to maximum likelihood decoding to a discretized version of the whole probability distribution. The resulting calculations require a form of the expectation-maximization algorithm (EM), regularized by a smoothing operation. We showed that this non-linear decoding method works well in a variety of circumstances, reconstructing broad, narrow and multimodal distributions more accurately than either the standard Poisson model or the kernel density model. Stochasticity is built into our method, since the units are supposed to have Poisson statistics, and it is therefore also quite robust to noise.

Various aspects of this scheme merit discussion. First, we readily acknowledge that the decoding model is quite non-biological, involving an implausible iterative computation. The point of our particular decoding model was to show explicitly a lower bound to the veracity with which a set of activities can code a distribution. One might expect the subsequent stages of processing in the brain to do one of two things with the information in the population:

1. integrate it with information represented in other population codes to form a combined population code (*eg* combining uncertain information about the relative position of two landmarks to generate the activity of a population code formed of place cells);
2. extract a single value from it to control behavior (*eg* pull a lever to report the best supported direction for the motion of the dots or choose where to explore for food in a maze).

In both cases, the extraction is presumably performed through standard neural operations such as taking non-linear weighted sums and possibly products of the activities. We are interested in how much information is preserved by such operations, as measured against the non-biological standard of our decoding method.

The former issue—how to integrate two or more population codes to generate the output in the form of another population code—was stressed by Hinton (1992), who noted that it directly relates to Ballard's (1981) notion of generalized Hough transforms. This question is particularly important because of the apparent ubiquity of population coding in the brain. It is not at all obvious how simple and local combination methods could be capable of preserving and combining probabilistic information in the population, and we are presently studying this question, using the EM-based decoder to generate targets and using local learning rules. One interesting theoretical concern is that the population code output of such a combination might not have exactly the same form as the population code inputs. For instance, it might not be completely accurate to characterize the cells as having Poisson statistics based on a Gaussian tuning function. In this case, one could formally calculate the true statistical interpretation of the combined code. However, in the brain, there does not appear to be a great difference between the population codes near to the input and the population codes in deeper areas, such as MT. This actually places a strong constraint on the method of combination.

One special concern for combination is how to understand noise. For instance, the visual system can be behaviorally extraordinarily sensitive, detecting just a handful of photons. However, the outputs of real cells at various stages in the system are quite noisy, with apparent Poisson statistics. If noise is added at every stage of processing and combination, then the final population code will not be very faithful to the input. There is much current and confusing research on the issue of the creation and elimination of noise in cortical synapses and neurons (Shadlen & Newsome, 1994; Mainen & Sejnowski, 1995; Bair & Koch, 1996). Correlated noise presents extra and different concerns.

A further concern for combination is the basis function strategy apparently adopted in parietal cortex, in which the position of an object in space is reported by neurons that have a retinotopic visual receptive field and multiplicative modulation from the position of the eyes in their orbits (Andersen *et al*/, 1985; Zipser & Andersen, 1988; Pouget & Sejnowski, 1995; Salinas & Abbott, 1996). Multiplicative modulation based on the locus of attention has also been found in V4 (Connor *et al*, 1996), and it has been suggested as being a general computational strategy (Poggio, 1990; Pouget & Sejnowski, 1997). The statistical effects of the multiplicative modulation remain to be investigated.

The second issue—extracting a single value from the population—is also important, particularly at the interface into the motor system. Some empirical data about how this is accomplished comes from the work on biasing the choices of the monkeys as to the directions of motion of the random dot stimuli through electrical micro-stimulation in MT (Salzman & Newsome, 1994). If the dots were moving in one direction but the electrical stimulation favored a different direction, then the monkeys would typically choose one or other of the two directions, rather than something like the mean direction. In our framework, we would regard the simultaneous activity of the MT cells which prefer the two directions as encoding a distribution, and postulate an output extraction process (thought to be in LIP, Shadlen & Newsome, 1996) that chooses a single value on the basis of this distribution. Short of knowing exactly the effects of the electrical stimulation on the activity of the MT cells, it is hard to use the experiment to confirm or reject this hypothesis.

A last issue that we have addressed is that of certainty or magnitude. Hinton's (1992) idea of using the sum total activity of a population to code the certainty in the existence of the quantity they represent is attractive, provided that there is some independent way of knowing what the scale is for this total. We used this scaling idea for both the kernel density and the extended Poisson models. In fact, we can go one stage further, and interpret greater activity still as representing information about the existence of multiple objects or multiple motions. However, this treatment seems less appropriate for the place cell system; the rat is presumably always certain that it is somewhere. There, it has been suggested that the absolute level of activity could be coding something different, such as the familiarity of a location.

An entire collection of cells is a terrible thing to waste on representing just a single value of some quantity. Representing a whole probability distribution, at least with some fidelity, is not more difficult, provided that the interpretation of the encoding and decoding are clear. Here we have suggested some steps in this direction.

References

Andersen, R. A., Essick, G. K., and Siegel, R. M. (1985). Encoding of spatial location in posterior parietal neurons. *Science*, 230:456–458.

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

Anderson, C. H. and Van Essen, D. C. (1994). Neurobiological computational systems. In *Computational Intelligence Imitating Life*, pages 213–222. IEEE Press, New York.

Bair, W. and Koch, C. (1996). Temporal precision of spike trains in extrastriate cortex of the behaving macaque monkey. *Neural Computation*, 8(6):1185–1202.

Baldi, P. and Heiligenberg, W. (1988). How sensory maps could enhance resolution through ordered arrangements of broadly tuned receivers. *Biological Cybernetics*, 59:313–318.

Ballard, D. H. (1981). Generalizing the Hough transfrom to detect arbitrary shapes. *Pattern Recognition*, 13(2):111–122.

Cheng, K. and Gallistel, C. R. (1984). Testing the geometric power of an animal's spatial representation. In *Animal Cognition*, pages 409–423. Erlbaum, Hillsdale, NJ.

Connor, C. E., Gallant, J. L., Preddie, D. C., and Van Essen, D. C. (1996). Responses in area V4 depend on the spatial relationship between stimulus and attention. *Journal of Neurophysiology*, 75:1306–1308.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B*, 39:1–38.

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*, pages 55–60. Kluwer Academic Publishers, Norwell, MA.

Georgopoulos, A. P., Schwartz, A. B., and Kettner, R. E. (1986). Neuronal population coding of movement direction. *Science*, 243:1416–1419.

Hinton, G. E. (1992). How neural networks learn from experience. *Scientific American*, 267(3):145–151.

Mainen, Z. F. and Sejnowski, T. J. (1995). Reliability of spike timing in neocortical neurons. *Science*, 268(5216):1503–1506.

Newsome, W. T., Britten, K. H., and Movshon, J. A. (1989). Neuronal correlates of a perceptual decision. *Nature*, 341:53–54.

O'Keefe, J. and Dostrovsky, J. (1971). The hippocampus as a spatial map: Preliminary evidence from unit activity in the freely-moving rat. *Brain Research*, 34:171–175.

Poggio, T. (1990). A theory of how the brain might work. *Cold Spring Harbor Symposium on Quantitative Biology*, 55:899–910.

Pouget, A. and Sejnowski, T. J. (1995). Spatial representations in the parietal cortex may use basis functions. In Tesauro, G., Touretzky, D. S., and Leen, T. K., editors, *Advances in Neural Information Processing Systems 7*, pages 157–164. MIT Press, Cambridge, MA.

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

Pouget, A., Zhang, K., Deneve, S., and Latham, P. (1998). Statistically efficient estimation using population codes. *Neural Computation*, 10(2):373–401.

Salinas, E. and Abbott, L. F. (1994). Vector reconstruction from firing rates. *Journal of Computational Neuroscience*, 1:89–107.

Salinas, E. and Abbott, L. F. (1995). Transfer of coded information from sensory to motor networks. *Journal of Neuroscience*, 15(10):6461–6474.

Salinas, E. and Abbott, L. F. (1996). A model of multiplicative neural responses in parietal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 93:11956–11961.

Salzman, C. D. and Newsome, W. T. (1994). Neural mechanisms for forming a perceptual decision. *Science*, 264:231–237.

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

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

Shadlen, M. N. and Newsome, W. T. (1994). Noise, neural codes and cortical organization. *Current Opinion in Neurobiology*, 4:569–579.

Shadlen, M. N. and Newsome, W. T. (1996). Motion perception: Seeing and deciding. *Proceedings of the National Academy of Sciences, USA*, 93:628–633.

Snippe, H. P. (1996). Parameter extraction from population codes: a critical assessment. *Neural Computation*, 8(3):511–530.

Snowden, R. (1989). Motions in orthogonal directions are mutually suppressive. *Journal of the Optical Society of America*, 6(7):1096–1101.

Wilson, M. A. and McNaughton, B. L. (1993). Dynamics of the hippocampal ensemble code for space. *Science*, 261:1055–1058.

Zemel, R. S. and Hinton, G. E. (1995). Developing population codes by minimizing description length. *Neural Computation*, 7(3):549–564.

Zipser, D. and Andersen, R. A. (1988). A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons. *Nature*, 331:679–684.

Zohary, E., Shadlen, M. N., and Newsome, W. T. (1994). Correlated neuronal discharge rate and its implications for psychophysical performance. *Nature*, 370(6485):140–3.