P. Cisek, T. Drew & J.F. Kalaska (Eds.)
Progress in Brain Research, Vol. 165
ISSN 0079-6123

CHAPTER 32

Probabilistic population codes and the exponential family of distributions

J. Beck[1], W.J. Ma[1], P.E. Latham[2] and A. Pouget[1,*]

[1]*Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY, USA*
[2]*Gatsby Computational Neuroscience Unit, London WC1N 3AR, UK*

Abstract: Many experiments have shown that human behavior is nearly Bayes optimal in a variety of tasks. This implies that neural activity is capable of representing both the value and uncertainty of a stimulus, if not an entire probability distribution, and can also combine such representations in an optimal manner. Moreover, this computation can be performed optimally despite the fact that observed neural activity is highly variable (noisy) on a trial-by-trial basis. Here, we argue that this observed variability is actually expected in a neural system which represents uncertainty. Specifically, we note that Bayes' rule implies that a variable pattern of activity provides a natural representation of a probability distribution, and that the specific form of neural variability can be structured so that optimal inference can be executed using simple operations available to neural circuits.

Keywords: Bayes; neural coding; inference; noise

Introduction

The information available to our senses regarding the external world is ambiguous and often corrupted by noise. Despite such uncertainty, humans not only function successfully in the world, but seem capable of doing so in a manner that is optimal in a Bayesian sense. This has been observed in a variety of cue combination tasks, including visual and haptic cue combination (Ernst and Banks, 2002; Kording and Wolpert, 2004), visual and auditory cue combination (Gepshtein and Banks, 2003), and visual–visual cue combination (Knill and Richards, 1996; Saunders and Knill, 2003; Landy and Kojima, 2001; Hillis et al., 2004). Since cue combination processes lie at the heart of nearly every aspect of human perception, it is important to understand how cue combination tasks can be performed optimally, both in principle and in cortex.

Cue combination can be illustrated with the following example: consider a cat seeking a mouse using both visual and auditory cues. If it is dark and the mouse is partially occluded by its surroundings, there is a high degree of uncertainty in the visual cues available. The mouse may even be hiding in a field of gray mouse-sized rocks and facing in any number of directions, increasing this uncertainty. In such a context, when visual information is highly uncertain, the Bayesian cat would base an estimate of the position of the mouse primarily upon auditory cues. In contrast, when light is abundant, the mouse is easily visible, and auditory input becomes the less reliable of the two cues. In this second case, the cat should rely

*Corresponding author. Tel.: +1 (585) 275 0760;
Fax: +1 (585) 442 9216; E-mail: alex@bcs.rochester.edu

DOI: 10.1016/S0079-6123(06)65032-2

primarily on visual cues to locate the mouse. If the cat's auditory and visual cue-based estimates of the mouse's position (s_a and s_v respectively) are independent, unbiased and Gaussian distributed with standard deviations of σ_a and σ_v, then the optimal estimate of the position of the mouse is given by

$$s_{a+v} = \frac{s_a/\sigma_a^2 + s_v/\sigma_v^2}{1/\sigma_a^2 + 1/\sigma_v^2} \quad (1)$$

Cue combination studies have shown Eq. (1) to be compatible with behavior even when σ_a and σ_v are adjusted on a trial-by-trial basis. Thus one may conclude that the cortex utilizes a neural code that represents the uncertainty, if not an entire probability distribution, for each cue in a way that is amenable to the optimal cue combination as described by Eq. (1).

At first glance, it may seem that cortical neurons are not well-suited to the task of representing probability distributions, as they have been observed to exhibit a highly variable response when presented with identical stimuli. This variability is often thought of as noise, which makes neural decoding difficult in that estimates of various task-relevant parameters become somewhat unreliable. Here, however, we will argue that it is critical to realize that neural variability and the representation of uncertainty go hand-in-hand. For example, suppose that each time our hypothetical cat observes a mouse, a unique pattern of activity reliably occurs in some region of cortex. If this were the case, then observation of that pattern of activity would indicate with certainty that the mouse is in the cat's visual field. Thus, only when the pattern of activity is variable in such a way that it overlaps with patterns of activity for which the mouse is *not* present, can the pattern indicate that the mouse is present only with "some probability." In reality, absolute knowledge is an impractical goal. This is not just because sensory organs are unreliable, but also because many problems faced by biological organisms are both ill-posed (there are an infinite number of three-dimensional configurations that lead to the same two-dimensional image on the retina) and data limited (the signal reaching the brain is too noisy to determine precisely what two-dimensional image produced it). Regardless, the above example indicates that neural variability is not only compatible with the representation of probability distributions in cortex, but is, in fact, expected in this context.

Ultimately, this insight is simply an acknowledgement of Bayes' rule, which states that when the presentation of a given stimulus, s, yields a variable neural response vector $\mathbf{r}$, then for any particular response $\mathbf{r}$, the distribution of the stimulus is given by

$$p(s|\mathbf{r}) = \frac{p(\mathbf{r}|s)p(s)}{p(\mathbf{r})} \quad (2)$$

The construction of a posterior distribution, $p(s|\mathbf{r})$, from a likelihood function, $p(\mathbf{r}|s)$, in this manner corresponds to an ideal observer analysis and is, by definition, optimal. We are not suggesting that a Bayesian decoder is explicitly implemented in cortex, but rather that the existence of Bayes' rule renders such a computation unnecessary, since a variable population pattern of activity already provides a natural representation of the posterior distribution (Foldiak, 1993; Anderson, 1994; Sanger, 1996; Zemel et al., 1998). This view stands in contrast to previous work (Rao, 2004; Deneve, 2005) advocating the construction of a network that represents the posterior distribution by directly identifying neural activity with either the probability of a specific value of the stimulus, the log of that probability, or convolutions thereof.

It is not immediately clear whether or not optimal cue combination, or other desirable operations, can be performed. We will address this issue through the construction of a Probabilistic Population Code (PPC) that is capable of performing optimal cue combination via linear operations. We will then show that when distributions over neuronal activity, i.e., the stimulus-conditioned neuronal responses, $p(\mathbf{r}|s)$, belong to the exponential family of distributions with linear sufficient statistics, then optimal cue combination (and other type of Bayesian inference, such as integration over time) can be performed through simple linear combinations. Members of this family of likelihood functions, $p(\mathbf{r}|s)$, will then be shown to be compatible with populations of neurons that have arbitrarily shaped tuning curves, arbitrary

covariance matrices, and can represent arbitrary posterior distributions.

Probabilistic Population Codes

We define a PPC as any code that uses Bayes' rule to optimally and *accessibly* encode a probability distribution. Here, we say that a code is accessible to a given neural circuit when that circuit is capable of performing the operations necessary to perform Bayesian inference and computation. For instance, in this work, we will be assuming that neural circuits are, at the very least, capable of performing linear operations and will seek the population code for which cue combination can be performed with some linear operation. To understand PPCs in a simplified setting, consider a Poisson distributed populations of neurons for which the tuning curve of neuron indexed by i is $f_i(s)$. In this case

$$p(\mathbf{r}|s,g) = \prod_i \frac{e^{-gf_i(s)}(gf_i(s))^{r_i}}{r_i!} \tag{3}$$

where r_i is the response or spike count of neuron i and g the amplitude, or gain, of population. When the prior is flat, i.e., $p(s)$ does not depend on s, application of Bayes' rule yields a posterior distribution that takes the form

$$\begin{aligned} p(s|\mathbf{r},g) &= \frac{p(\mathbf{r}|s,g)p(s|g)}{p(\mathbf{r}|g)} \\ &= \frac{1}{Lp(\mathbf{r}|g)} \prod_i \frac{e^{-gf_i(s)}(gf_i(s))^{r_i}}{r_i!} \\ &= \frac{1}{Lp(\mathbf{r}|g)} \left(\prod_i \frac{1}{r_i!} \right) \exp\left(\sum_i r_i \log g - gf_i(s) \right) \\ &\quad \exp\left(\sum_i r_i \log f_i(s) \right) \\ &= \frac{1}{Lp(\mathbf{r}|g)} \left(\prod_i \frac{1}{r_i!} \right) \exp\left(\sum_i r_i \log g - gc \right) \\ &\quad \exp\left(\sum_i r_i \log f_i(s) \right) \\ &\quad \propto \exp\left(\sum_i r_i \log(f_i(s)) \right) \end{aligned} \tag{4}$$

where $1/L = p(s)$, and we have assumed that tuning curves are sufficiently dense so that $\sum_i f_i(s) = c$. Because this last line of the equation represents an unnormalized probability distribution over s, we may conclude that the constant of proportionality depends only on $\mathbf{r}$ and is thus also independent of the gain g.

From Eq. (4), we can conclude that, if we knew the shape of the tuning curves $f_i(s)$, then for any given pattern of activity $\mathbf{r}$ (and any gain, g), we could simply plot this equation as a function of s to obtain the posterior distribution (see Fig. 1). In the language of a PPC we say that knowledge of the likelihood function, $p(\mathbf{r}|s)$, automatically implies knowledge of the posterior distribution $p(s|\mathbf{r})$. In this simple case of independent Poisson neurons, knowledge of the likelihood function means knowing the shape of the tuning curves. As we will now demonstrate, this knowledge is also sufficient for the identification of an optimal cue combination computation.

To this end, suppose that we have two such populations, $\mathbf{r}_1$ and $\mathbf{r}_2$, each of which encodes some piece of independent information about the same stimulus. In the context of our introductory example, $\mathbf{r}_1$ might encode the position of a mouse given visual information while $\mathbf{r}_2$ might encode the position of the mouse given auditory information. When the two populations are conditionally independent given the stimulus and the prior is flat, the posterior distribution of the stimulus given both population patterns of activity is simply given by the product of the posterior distributions given each population independently

$$\begin{aligned} p(s|\mathbf{r}_1,\mathbf{r}_2) &\propto p(\mathbf{r}_1,\mathbf{r}_2|s) \\ &\propto p(\mathbf{r}_1|s)p(\mathbf{r}_2|s) \\ &\propto p(s|\mathbf{r}_1)p(s|\mathbf{r}_2) \end{aligned} \tag{5}$$

Thus, in this case optimal cue combination corresponds to the multiplication of posteriors and subsequent normalization.

As illustrated in Fig. (2), a two layer network which performs the optimal cue combination operation would combine the two population patterns of activity, $\mathbf{r}_1$ and $\mathbf{r}_2$, into a third population, $\mathbf{r}_3$, so that

$$p(s|\mathbf{r}_3) = p(s|\mathbf{r}_1,\mathbf{r}_2) \tag{6}$$

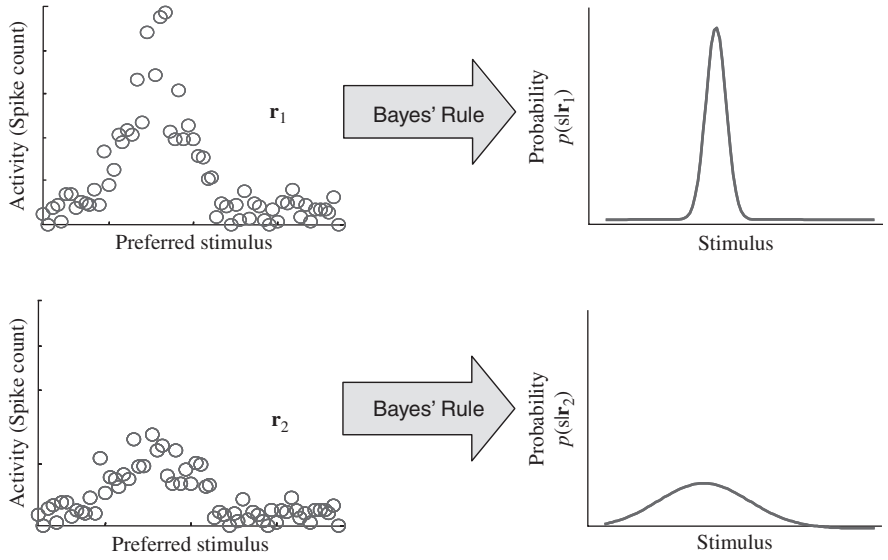

Fig. 1. Two population patterns of activity, $\mathbf{r}_1$ and $\mathbf{r}_2$, which were drawn from the distribution given by Eq. (3) with Gaussian-shaped tuning curves. Since the tuning curve shape is known, we can compute posterior $p(s|\mathbf{r})$ for any given $\mathbf{r}$, either by simply plotting the likelihood, $p(\mathbf{r}|s)$, as a function of the stimulus s or, equivalently, by using Eq. (4).

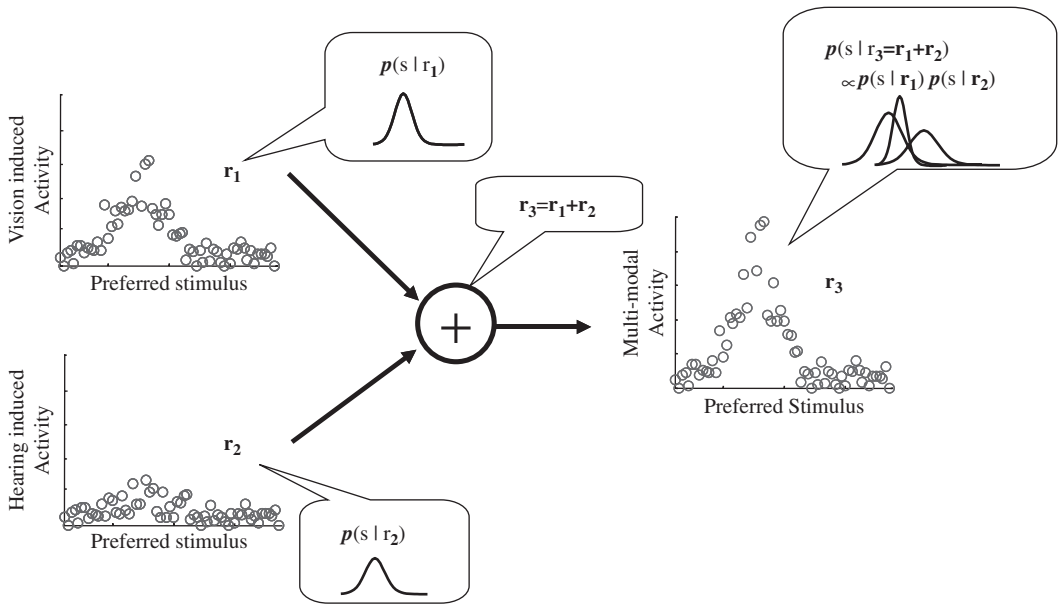

Fig. 2. On the left, the activities of two independent populations, $\mathbf{r}_1$ and $\mathbf{r}_2$, are added to yield a third population pattern $\mathbf{r}_3$. In the insets, we plot the posterior distributions associated with each of the three activity patterns are shown. In this case, optimal cue combination corresponds to a multiplication of the posteriors associated with the two independent populations.

For identically tuned populations, $\mathbf{r}_3$ is simply the sum, $\mathbf{r}_1+\mathbf{r}_2$. Since $\mathbf{r}_3$ is the sum of two Poisson random variables with identically shaped tuning curves, it too is a Poisson random variable with the same shaped tuning curve (but with a larger amplitude). As such, the Bayesian decoder applied to $\mathbf{r}_3$ takes the same form as the Bayesian decoder for $\mathbf{r}_1$ and $\mathbf{r}_2$. This implies

$$\begin{aligned} p(s|\mathbf{r}_3) &\propto \exp\left(\sum_i r_{3i}\log(f_i(s))\right) \\ &\propto \exp\left(\sum_i (r_{1i}+r_{2i})\log(f_i(s))\right) \\ &\propto \exp\left(\sum_i r_{1i}\log(f_i(s))\right)\exp\left(\sum_i r_{2i}\log(f_i(s))\right) \\ &\propto p(s|\mathbf{r}_1)p(s|\mathbf{r}_2) \\ &\propto p(s|\mathbf{r}_1,\mathbf{r}_2) \end{aligned} \tag{7}$$

and we conclude that multiplication of the two associated posterior distributions corresponds to the addition of two population codes. When the tuning curves are Gaussian, this result can also be obtained through a consideration of the variance of the maximum likelihood estimate obtained from the posterior distribution associated with the population $\mathbf{r}_3$. This results from the fact that, for Gaussian tuning curves, the log of $f_i(s)$ is quadratic in s and thus the posterior distribution is also Gaussian with maximum likelihood estimate, $\hat{s}(\mathbf{r})$, and estimate variance, $\sigma^2(\mathbf{r})$. These quantities are related to the population pattern activity $\mathbf{r}$, via the expressions

$$\begin{aligned} \hat{s}(\mathbf{r}) &= \frac{\sum_i s_i r_i}{\sum_i r_i} \\ \frac{1}{\sigma^2(\mathbf{r})} &= \frac{1}{\sigma_{tc}^2}\sum_i r_i \end{aligned} \tag{8}$$

where s_i is the preferred stimulus of the ith and σ_{tc} gives the width of the tuning curve. The estimate, $\hat{s}(\mathbf{r})$, is the well-known population vector decoder, which is known to be optimal in this case (Snippe, 1996). Now, we use the fact that the expression for the mean and variance of the posterior associated with $\mathbf{r}_3$ is the same as the expression associated with $\mathbf{r}_1$ and $\mathbf{r}_2$. This implies that

$$\begin{aligned} \frac{1}{\sigma_3^2(\mathbf{r}_3)} &= \frac{1}{\sigma_{tc}^2}\sum_i r_{3i} = \frac{1}{\sigma_{tc}^2}\sum_i r_{1i}+r_{2i} \\ &= \frac{1}{\sigma_1^2(\mathbf{r}_1)}+\frac{1}{\sigma_2^2(\mathbf{r}_2)} \end{aligned} \tag{9}$$

and

$$\begin{aligned} \hat{s}_3(\mathbf{r}_3) &= \frac{\sum_i r_{3i}s_i}{\sum_i r_{3i}} = \frac{\sum_i r_{1i}s_i+\sum_i r_{2i}s_i}{\sum_i r_{1i}+\sum_i r_{2i}} \\ &= \frac{\hat{s}_1(\mathbf{r}_1)/\sigma_1^2(\mathbf{r}_1)+\hat{s}_2(\mathbf{r}_2)/\sigma_2^2(\mathbf{r}_2)}{1/\sigma_1^2(\mathbf{r}_1)+1/\sigma_2^2(\mathbf{r}_2)} \end{aligned} \tag{10}$$

Comparison with Eq. (1) demonstrates optimality. Moreover, optimality is achieved on a trial-by-trial basis, since the estimate of each population is weighted by a variance which is computed from the actual population pattern of activity.

It is also important to note that, in the computation above, we did not explicitly compute the posterior distributions, $p(s|\mathbf{r}_i)$, and then multiply them together. Rather we operated on the population patterns of activity (by adding them together) and then simply remarked (Fig. 2) that we could have applied Bayes' rule to optimally decode these population patterns and noted the optimality of the computation. This is the essence of a PPC. Specifically, a PPC consists of three things: (1) a set of operations on neural responses $\mathbf{r}$; (2) a desired set of operations in posterior space, $p(s|\mathbf{r})$; and (3) the family of likelihood functions, $p(\mathbf{r}|s)$, for which this operation pair is optimal in a Bayesian sense. In the cue combination example above, the operation of addition of population patterns of activity (list item 1) was shown to correspond to the operation of operation of multiplication (and renormalization) of posterior distributions (list item 2), when the population patterns of activity were drawn from likelihood functions which corresponded to an independent Poisson spiking populations with identically shaped tuning curves (list item 3). Moreover, simple addition was shown to be optimal regardless of the variability of each cue, i.e., unlike other proposed cue combination

schemes (Rao, 2004; Navalpakkam and Itti, 2005), this scheme does not require that the weights of the linear combination be adjusted on a trial-by-trial basis.

Generalization to the exponential family with linear sufficient statistics

So far we have relied on the assumption that populations consist of independent and Poisson spiking neurons with identically shaped tuning curves. However, this is not a limitation of this approach. As it turns out, constant coefficient linear operations can be found that correspond to optimal cue combination for a broad class of Poisson-like likelihood functions described by the so-called exponential family with linear sufficient statistics. This family includes members that can have any tuning curve shape, any correlation structure, and can represent any shape of the posterior, i.e., not just Gaussian posteriors. Below, we show that, in the case of contrast-invariant turning curves, the requirement that optimal cue combination occur via linear combination of population patterns of activity with fixed coefficients only limits us to likelihood functions for which the variance is proportional to the mean.

Optimal cue combination via linear combination of population codes

Consider two population codes, $\mathbf{r}_1$ and $\mathbf{r}_2$, encoding visual and auditory location cues, which are jointly distributed according to $p(\mathbf{r}_1, \mathbf{r}_2|s)$. The goal of an optimal cue combination computation is to combine these two populations into a third population pattern of activity $\mathbf{r}_3 = \mathbf{F}(\mathbf{r}_1, \mathbf{r}_2)$, so that an application of the Bayes rule yields

$$p(s|\mathbf{r}_3) = p(s|\mathbf{r}_1, \mathbf{r}_2) \tag{11}$$

Note that optimal cue combination can be trivially achieved by selecting any invertible function $\mathbf{F}(\mathbf{r}_1, \mathbf{r}_2)$. To avoid this degenerate case, we assume the lengths of the vectors $\mathbf{r}_1$, $\mathbf{r}_2$, and $\mathbf{r}_3$ are the same. Thus the function $\mathbf{F}$ cannot be invertible. The likelihood of $\mathbf{r}_3$ is related to the likelihood of $\mathbf{r}_1$ and $\mathbf{r}_2$ via the equation

$$p(\mathbf{r}_3|s) = \int p(\mathbf{r}_1, \mathbf{r}_2|s)\delta(\mathbf{r}_3 - F(\mathbf{r}_1, \mathbf{r}_2))d\mathbf{r}_1 d\mathbf{r}_2 \tag{12}$$

Application of Bayes' rule and the condition of optimality [Eq. (11)], indicates that an optimal cue combination operation, $\mathbf{F}(\mathbf{r}_1, \mathbf{r}_2)$, depends on the likelihood, $p(\mathbf{r}_1, \mathbf{r}_2|s)$.

Interestingly, if the likelihood is in the exponential family with linear sufficient statistics, a linear function $\mathbf{F}(\mathbf{r}_1, \mathbf{r}_2)$ can be found such that Eq. (11) is satisfied. Members of this family take the functional form

$$p(\mathbf{r}_1, \mathbf{r}_2|s) = \frac{\phi_{12}(\mathbf{r}_1, \mathbf{r}_2)}{\eta_{12}(s)} \exp(\mathbf{h}_1^{\mathrm{T}}(s)\mathbf{r}_1 + \mathbf{h}_2^{\mathrm{T}}(s)\mathbf{r}_2) \tag{13}$$

where the superscript "T" denotes transpose, $\phi_{12}(\mathbf{r}_1, \mathbf{r}_2)$ is the so-called measure function, and $\eta_{12}(s)$ the normalization factor, often called the partition function. Here $\mathbf{h}_1(s)$ and $\mathbf{h}_2(s)$ are vector functions of s, which are called the stimulus-dependent kernels associated with each population. We make the additional assumption that $\mathbf{h}_1(s)$ and $\mathbf{h}_2(s)$ share a common basis $\mathbf{b}(s)$ which can also be represented as vector of functions of s. This implies that we may write $\mathbf{h}_i(s) = \mathbf{A}_i\mathbf{b}(s)$ for some stimulus independent matrix $\mathbf{A}_i$ ($i = 1, 2$). We will now show that, when this is the case, optimal combination is performed by the linear function

$$\mathbf{r}_3 = \mathbf{F}(\mathbf{r}_1, \mathbf{r}_2) = \mathbf{A}_1^{\mathrm{T}}\mathbf{r}_1 + \mathbf{A}_2^{\mathrm{T}}\mathbf{r}_2 \tag{14}$$

Moreover, this we will show that the likelihood function $p(\mathbf{r}_3|s)$ is also in the same family of distributions as $p(\mathbf{r}_1, \mathbf{r}_2|s)$. This is important, as it demonstrates that this approach — taking linear combinations of firing rates to perform optimal Bayesian inference — can be either repeated iteratively over time or cascaded from one population to the next.

Optimality of this operation is most easily demonstrated by computing the likelihood, $p(\mathbf{r}_3|s)$, applying Bayes' rule to obtain $p(s|\mathbf{r}_3)$ and then showing that $p(s|\mathbf{r}_3) = p(s|\mathbf{r}_1, \mathbf{r}_2)$. Combining

Eqs. (12–14) above indicates that

$$\begin{aligned} p(\mathbf{r}_3|s) &= \int \frac{\phi_{12}(\mathbf{r}_1,\mathbf{r}_2)}{\eta_{12}(s)} \exp(\mathbf{h}_1^{\mathrm{T}}(s)\mathbf{r}_1 \\ &\quad + \mathbf{h}_2^{\mathrm{T}}(s)\mathbf{r}_2)\delta(\mathbf{r}_3 - \mathbf{A}_1^{\mathrm{T}}\mathbf{r}_1 - \mathbf{A}_2^{\mathrm{T}}\mathbf{r}_2)d\mathbf{r}_1 d\mathbf{r}_2 \\ &= \int \frac{\phi_{12}(\mathbf{r}_1,\mathbf{r}_2)}{\eta_{12}(s)} \exp(\mathbf{b}^{\mathrm{T}}(s)\mathbf{A}_1^{\mathrm{T}}\mathbf{r}_1 \\ &\quad - \mathbf{b}^{\mathrm{T}}(s)\mathbf{A}_2^{\mathrm{T}}\mathbf{r}_2)\delta\left(\mathbf{r}_3 - \mathbf{A}_1^{\mathrm{T}}\mathbf{r}_1 - \mathbf{A}_2^{\mathrm{T}}\mathbf{r}_2\right)d\mathbf{r}_1 d\mathbf{r}_2 \\ &= \frac{\exp(\mathbf{b}^{\mathrm{T}}(s)\mathbf{r}_3)}{\eta_{12}(s)} \int \phi_{12}(\mathbf{r}_1,\mathbf{r}_2) \\ &\quad \delta(\mathbf{r}_3 - \mathbf{A}_1^{\mathrm{T}}\mathbf{r}_1 - \mathbf{A}_2^{\mathrm{T}}\mathbf{r}_2)d\mathbf{r}_1 d\mathbf{r}_2 \\ &= \frac{\phi_3(\mathbf{r}_3)}{\eta_{12}(s)} \exp(\mathbf{b}^{\mathrm{T}}(s)\mathbf{r}_3) \end{aligned} \tag{15}$$

where $\phi_3(\mathbf{r}_3)$ is a new measure function that is independent of s. This demonstrates that $\mathbf{r}_3$ is also a member of this family of distributions with a stimulus-dependent kernel drawn from the common basis associated with $\mathbf{h}_1(s)$ and $\mathbf{h}_2(s)$. As in the independent Poisson case, the Bayesian decoder applied to a member of this family of distributions takes the form

$$p(s|\mathbf{r}_1,\mathbf{r}_2) \propto \frac{\exp(\mathbf{h}_1^{\mathrm{T}}(s)\mathbf{r}_1 + \mathbf{h}_2^{\mathrm{T}}(s)\mathbf{r}_2)}{\eta_{12}(s)} \tag{16}$$

and we may conclude that optimal cue combination has been performed by this linear operation, since

$$\begin{aligned} p(s|\mathbf{r}_3) &\propto \exp\frac{(\mathbf{b}^{\mathrm{T}}(s)\mathbf{r}_3)}{\eta_{12}(s)} \\ &\propto \exp\frac{(\mathbf{b}^{\mathrm{T}}(s)\mathbf{A}_1^{\mathrm{T}}\mathbf{r}_1 + \mathbf{b}^{\mathrm{T}}(s)\mathbf{A}_2^{\mathrm{T}}\mathbf{r}_2)}{\eta_{12}(s)} \\ &\propto \exp\frac{(\mathbf{h}_1^{\mathrm{T}}(s)\mathbf{r}_1 + \mathbf{h}_2^{\mathrm{T}}(s)\mathbf{r}_2)}{\eta_{12}(s)} \\ &\propto p(s|\mathbf{r}_1,\mathbf{r}_2) \end{aligned} \tag{17}$$

Note that the measure function $\phi_{12}(\mathbf{r}_1,\mathbf{r}_2)$, which was defined in Eq. (13), is completely arbitrary so long as it does not depend on the stimulus.

Nuisance parameters and gain

In the calculation above, we assumed that the likelihood function, $p(\mathbf{r}|s)$, is a function *only* of the stimulus. In fact, the likelihood often depends on what are commonly called *nuisance parameters*. These are quantities that affect the response distributions of the individual neural populations, but that the brain would like to ignore when performing inference. For example, it is well-known that contrast and attention strongly affect the gain and information content of a population, as was the case in the independent Poisson example above. For members of the exponential family of distributions, this direct relationship between gain and information content is, in fact, expected. Recalling that the posterior distribution takes the form

$$p(s|\mathbf{r}) \propto \exp(\mathbf{h}^{\mathrm{T}}(s)\mathbf{r}) \tag{18}$$

it is easy to see that a large amplitude population pattern of activity is associated with a significantly sharper distribution than a low amplitude pattern of activity (see Fig. 1). From this we can conclude that amplitude or gain is an example of a nuisance parameter of particular interest as it is directly related to the variance of the posterior distribution.

We can model this gain dependence by writing the likelihood function for populations 1 and 2 as $p(\mathbf{r}_1,\mathbf{r}_2|s,g_1,g_2)$ where g_k is the gain parameter for population k ($k = 1,2$). Although we could apply this Bayesian formalism and treat g_1 and g_2 as part of the stimulus, if we did that the likelihood for $\mathbf{r}_3$ would contain the term $\exp(\mathbf{b}^{\mathrm{T}}(s,g_1,g_2)\ \mathbf{r}_3)$ [see Eq. (16)]. This is clearly inconvenient, as it means we would have to either know g_1 and g_2, or integrate these quantities out of the posterior distribution to extract the a posterior distribution for the stimulus alone, i.e., we would have to find a neural operation which effectively performs the integral

$$p(s|\mathbf{r}) = \int p(s,g|\mathbf{r})dg \tag{19}$$

Fortunately, it is easy to show that this problem can be avoided if the nuisance parameter does not appear in the stimulus-dependent kernel, i.e., when the likelihood for a given population takes the form

$$p(\mathbf{r}|s,g) = \phi(\mathbf{r},g)\exp(\mathbf{h}^{\mathrm{T}}(s)\mathbf{r}) \tag{20}$$

When this is the case, the posterior distribution is given by

$$p(s|\mathbf{r}, g) \propto \exp(\mathbf{h}^{\mathrm{T}}(s)\mathbf{r}) \tag{21}$$

and the value of g does not affect posterior distribution over s and thus does not affect the optimal combination operation. If $\mathbf{h}(s)$ were a function of g, this would not necessarily be true. Note that the normalization factor, $\eta(s, g)$, from Eq. (16) is not present in Eq. (20). This is because the posterior is only unaffected by g when $\eta(s, g)$ factorizes into a term that is dependent only on s and a term that is dependent only on g and this occurs only when $\eta(s, g)$ is independent of s. Fortunately, this seemingly strict condition is satisfied in many biologically relevant scenarios, and seems to be intricately related to the very notion of a tuning curve. Specifically, when silence is uninformative ($\mathbf{r} = 0$ gives no information about the stimulus), it can be shown that the normalization factor, $\eta(s, g)$, is dependent only on g

$$\begin{aligned} p(s) &= p(s|\mathbf{r} = \mathbf{0}, g) \\ &= \frac{\phi(\mathbf{0}, g)p(s)}{\eta(s, g)} \left(\int \frac{\phi(\mathbf{0}, g)p(s')}{\eta(s', g)} ds' \right)^{-1} \\ &= \frac{p(s)}{\eta(s, g)} \left(\int \frac{p(s')}{\eta(s', g)} ds' \right)^{-1} \end{aligned} \tag{22}$$

Since the second term in the product on the right hand side is a function only of g, equality holds only when $\eta(s, g)$ is independent of s.

Relationship between the tuning curves, the covariance matrix and the stimulus-dependent kernel h(s)

At this point we have demonstrated that the family of likelihood function described above is compatible with the identification of linear operations on neural activity with the optimal cue combination of associated posterior distribution. What remains unclear is whether or not this family of likelihood functions is capable of describing the statistics of neural populations. In this section, we show that this family of distribution is applicable to a very wide range of tuning curves and covariance matrices, i.e., members of this family of distributions can model the s-dependence of the first and second order statistics of any population. We will then show that when the shape of the tuning curve is gain invariant, we expect to observe that the covariance matrix should also be proportional to gain. This is an important result, as it is a widely observed property of the statistics of cortical neurons.

We begin by showing that the tuning curve and covariance matrix of a population pattern of interest are related to the stimulus-dependent kernel by a simple relationship obtained via consideration of the derivative of the mean of the population pattern of activity, $\mathbf{f}(s, g)$, with respect to the stimulus as follows:

$$\begin{aligned} \mathbf{f}'(s, g) &= \frac{d}{ds} \int \mathbf{r}\phi(\mathbf{r}, g) \exp(\mathbf{h}^{\mathrm{T}}(s)\mathbf{r}) d\mathbf{r} \\ &= \int \mathbf{r}\mathbf{r}^{\mathrm{T}}\mathbf{h}'(s)\phi(\mathbf{r}, g) \exp(\mathbf{h}^{\mathrm{T}}(s)\mathbf{r}) d\mathbf{r} \\ &= \int \mathbf{r}\mathbf{r}^{\mathrm{T}}\mathbf{h}'(s)p(\mathbf{r}|s, g) d\mathbf{r} \\ &= \langle \mathbf{r}\mathbf{r}^{\mathrm{T}} \rangle_{s,g} \mathbf{h}'(s) - \mathbf{f}(s, g)\mathbf{f}^{\mathrm{T}}(s, g)\mathbf{h}'(s) \\ &= \mathbf{\Gamma}(s, g)\mathbf{h}'(s) \end{aligned} \tag{23}$$

Here $\langle \rangle_{s,g}$ is an expected value conditioned on s and g, $\mathbf{\Gamma}(s,g)$ the covariance matrix and we have used the fact that $\mathbf{f}^{\mathrm{T}}(s, g)\mathbf{h}'(s) = 0$ for all distributions given by Eq. (20), which follows from the assumption that silence is uninformative. Next, we rewrite Eq. (23) as

$$\mathbf{h}'(s) = \mathbf{\Gamma}^{-1}(s, g)\mathbf{f}'(s, g) \tag{24}$$

and observe that, in the absence of nuisance parameters, a stimulus-dependent kernel can be found for any tuning curve and any covariance matrix, regardless of their stimulus-dependence. Thus this family of distributions is as general as the Gaussian family in terms of its ability to model the first and second order statistics of any experimentally observed population pattern of activity. However, when nuisance parameters, such as gain, are present the requirement that the stimulus-dependent kernel, $\mathbf{h}(s)$, be independent of these parameters restricts the set of tuning curves and covariance matrices that are compatible with this family of distributions. For example, when the tuning curve shape is gain invariant

(i.e., $\mathbf{f}'(s,g) = g\bar{\mathbf{f}}'(s)$ where $\bar{\mathbf{f}}(s)$ is independent of gain), $\mathbf{h}$(s) is independent of the gain if the covariance matrix is proportional to the gain. Since variance and mean are both proportional to the gain, their ratio, known as the Fano factor, is also constant. Thus we conclude that constant Fano factors are associated with neurons that implement a linear PPC using tuning curves which have gain invariant shape. Hereafter, likelihood functions with these properties will be referred to as "Poisson-like" likelihoods.

Constraint on the posterior distribution over s

In addition to being compatible with a wide variety of tuning curves and covariance matrices, Poisson-like likelihoods can be used to represent many types of posterior distributions, including non-Gaussian ones. Specifically, as in the independent Poisson case, when the prior $p(s)$ is flat, application of Bayes rule yields

$$p(s|\mathbf{r}) \propto \exp(\mathbf{h}^{\mathrm{T}}(s)\mathbf{r}) \tag{25}$$

Thus, the log of the posterior is a linear combination of the functions that make up the vector $\mathbf{h}(s)$, and we may conclude any posterior distribution may be well approximated when this set of functions is "sufficiently rich." Of course, it is also possible to restrict the set of posterior distributions by an appropriate choice for $\mathbf{h}(s)$. For instance, if a Gaussian posterior is required, we can simply restrict the basis of $\mathbf{h}(s)$ to the set quadratic functions of s.

An example: combining population codes

To illustrate this point, in Fig. 3 we show a simulation in which there are three input layers in which the tuning curves are Gaussian sigmoidal with a positive slope, and sigmoidal with a negative slope (Fig. 3a). The parameters of the individual tuning curves, such as the widths, slopes, amplitude, and baseline activity, are randomly selected. Associated with each population is a stimulus-dependent kernel, $\mathbf{h}_k(s\mathrm{k} = 1, 2, 3$. Since the set of Gaussian functions of s form a basis, $\mathbf{b}(s)$, each of these stimulus-dependent kernels can be represented as a linear combination of these functions, i.e., $\mathbf{h}_k(s) = \mathbf{A}_k\mathbf{b}(s)$. Thus the linear combination of the input activities, $\mathbf{A}_1^{\mathrm{T}}\mathbf{r}_1 + \mathbf{A}_2^{\mathrm{T}}\mathbf{r}_2 + \mathbf{A}_3^{\mathrm{T}}\mathbf{r}_3$, corresponds to the product of the three posterior distributions. To ensure that all responses are positive, a baseline is removed, and the output population pattern of activity is given by

$$\begin{aligned}\mathbf{r}_4 = {} & \mathbf{A}_1^{\mathrm{T}}\mathbf{r}_1 + \mathbf{A}_2^{\mathrm{T}}\mathbf{r}_2 + \mathbf{A}_3^{\mathrm{T}}\mathbf{r}_3 \\ & - \min(\mathbf{A}_1^{\mathrm{T}}\mathbf{r}_1 + \mathbf{A}_2^{\mathrm{T}}\mathbf{r}_2 + \mathbf{A}_3^{\mathrm{T}}\mathbf{r}_3)\end{aligned} \tag{26}$$

Note that the choice of a Gaussian basis, $\mathbf{b}(s)$, yields more or less Gaussian-shaped tuning curves (Fig. 3c) in the output population and that the resulting population pattern of activity is highly correlated. Additionally, for this basis, the removal of the baseline can be shown to have no affect of the resulting posterior, regardless of how that baseline is chosen.

Figure 3b–d shows the network activity patterns and corresponding probability distributions on a given trial. As can be seen in Fig. 3d, the probability distribution encoded by the output layer is identical to the distribution obtained from multiplying the input distributions.

Discussion

We have shown that when the neuronal variability is Poisson-like, i.e., it belongs to the exponential family with linear sufficient statistics, Bayesian inference such as the one required for optimal cue combination reduces to simple linear combinations of neural activities.

In the case in which probability distributions are all Gaussian, reducing Bayesian inference to linear combination may not appear to be so remarkable, since Bayesian inferences are linear in this case. Indeed, as can be seen in Eq. (1), the visual–auditory estimate is obtained through a linear combination of the visual estimate and auditory estimate. However, in this equation, the weights of the linear combination are related to the variance of each estimate, in such a way that the combination favors the cue with the smallest variance, i.e., the most reliable cue. This is problematic, because it implies that the weights must be adjusted every time the reliability of the cues

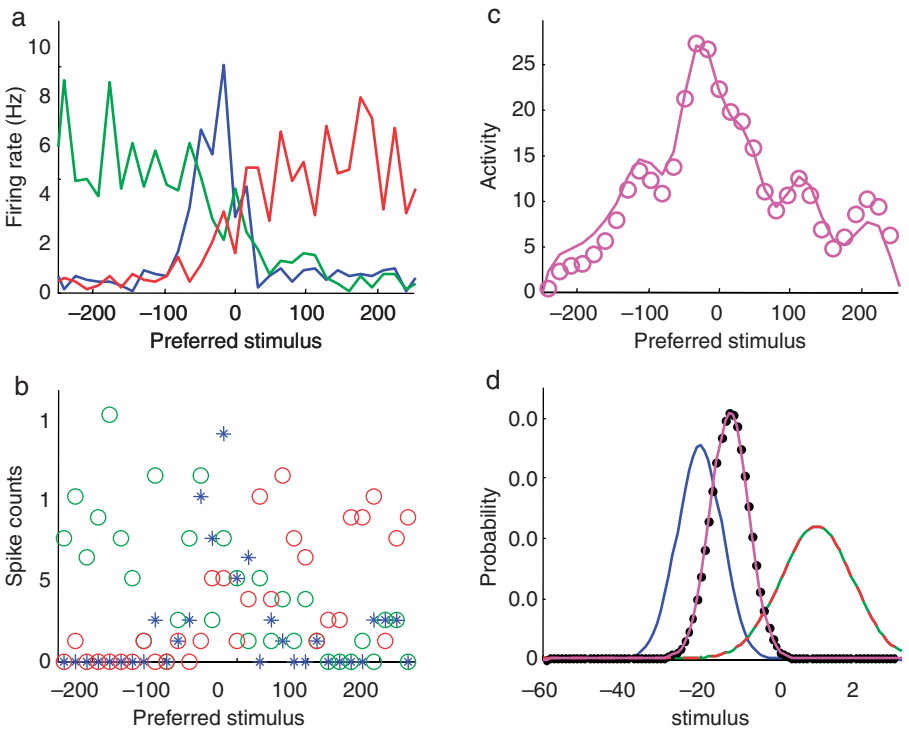

Fig. 3. Inference with non-translation invariant Gaussian and sigmoidal tuning curves. (a) Mean activity in the three input layers when $s = 0$. Blue curves: input layer with Gaussian tuning curves. Red curves: input layers with sigmoidal tuning curves with positive slopes. Green curves: input layers with sigmoidal tuning curves with negative slopes. The noise in the curves is due to variability in the baseline, widths, slopes, and amplitudes of the tuning curves, and to the fact that the tuning curves are not equally spaced along the stimulus axis. (b) Activity in the three input layers on a given trial. These activities were sampled from Poisson distributions with means as in a. Color legend as in a. (c) Solid lines: mean activity in the output layer. Circles: output activity on a given trial, obtained by a linear combination of the input activities shown in b. (d) Blue curves: probability distribution encoded by the blue stars in b (input layer with Gaussian tuning curves). Red-green curve: probability distribution encoded by the red and green circles in b (the two input layers with sigmoidal tuning curves). Magenta curve: probability distribution encoded by the activity shown in c (magenta circles). Black dots: probability distribution obtained with Bayes rule (i.e., the product of the blue and red-green curves appropriately normalized). The fact that the black dots are perfectly lined up with the magenta curve demonstrates that the output activity shown in **c** encodes the probability distribution expected from Bayes rule.

changes. By contrast, with the approach described in this chapter, there is no need for such weight adjustments. If the noise is Poisson-like, a linear combination with fixed coefficients works across any range of cue reliability. This is the main advantage of our approach. Moreover, it explains how humans remain optimal even when the reliability of the cue changes from trial to trial, without having to invoke any trial-by-trial adjustment of synaptic weights.

Another appealing feature of our theory is that it suggests an explanation for why all cortical neurons exhibit Poisson-like noise. In early stages of sensory processing, the statistics of spike trains are not necessarily Poisson-like, and differ across sensory systems. In the subcortical stages of the auditory system, spike timing is very precise and spike trains are oscillatory, reflecting the oscillatory nature of sound waves. By contrast, in the LGN (the thalamic relay of the visual system), the spike trains in response to static stimuli are neither very precise nor oscillatory. Performing optimal multisensory integration with such differences in spike statistics is a difficult problem. If our theory is correct, the cortex solves the problem by first reformatting all spike trains into the Poisson-like

family, so as to reduce optimal integration to simple sums of spikes. This transformation is particularly apparent in the auditory system. In the auditory cortex of awake animals, the response of most neurons show Poisson-like statistics, in sharp contrast with the oscillatory spike train seen in early subcortical stages.

The idea that the cortex uses a format that reduced Bayesian inference to linear combination is certainly appealing, but one should not forget that many aspects of neural computation are highly nonlinear. In its present form, our theory does not require those nonlinearities. However, we have only considered one type of Bayesian inference, namely, products of distributions. There are other types of Bayesian inference, such as marginalization, that are just as important. In fact, marginalization is needed to perform the optimal nonlinear computations which are needed for most sensorimotor transformations. We suspect that nonlinearities will be needed to implement marginalization optimally in neural hardware when the noise is Poisson-like, and may also be necessary to implement optimal cue combination when noise is not Poisson-like. We intend to investigate these and related issues in future studies. It is also important to note that, in its most general formulation, a PPC does not necessarily assign equality of the operations of addition of neural responses to optimal cue combination of related posteriors. Rather this is just a particular example of a PPC which seems to be compatible with neural statistics.

References

Anderson, C. (1994) Computational Intelligence Imitating Life. IEEE Press, New York, pp. 213–222.

Deneve, S. (2005) Neural Information Processing System. MIT Press, Cambridge, MA.

Ernst, M.O. and Banks, M.S. (2002) Humans integrate visual and haptic information in a statistically optimal fashion. Nature, 415: 429–433.

Foldiak, P. (1993) In: Eeckman F. and Bower J. (Eds.), Computation and Neural Systems. Kluwer Academic Publishers, Norwell, MA, pp. 55–60.

Hillis, J.M., Watt, S.J., Landy, M.S. and Banks, M.S. (2004) Slant from texture and disparity cues: optimal cue combination. J. Vis., 4(12): 967–992.

Knill, D.C. and Richards, W. (1996) Perception as Bayesian Inference. Cambridge University Press, New York.

Kording, K.P. and Wolpert, D.M. (2004) Bayesian integration in sensorimotor learning. Nature, 427: 244–247.

Landy, M.S. and Kojima, H. (2001) Ideal cue combination for localizing texture-defined edges. J. Opt. Soc. Am. A., 18(9): 2307–2320.

Navalpakkam, V. and Itti, L. (2005) Optimal cue selection strategy. In: Neural Information Processing System. MIT Press, Cambridge, MA.

Rao, R.P. (2004) Bayesian computation in recurrent neural circuits. Neural Comput., 16: 1–38.

Sanger, T. (1996) Probability density estimation for the interpretation of neural population codes. J. Neurophysiol., 76: 2790–2793.

Saunders, J.A. and Knill, D. (2003) Perception of 3D surface orientation from skew symmetry. Vision Res., 41(24): 3163–3183.

Snippe, H.P. (1996) Parameter extraction from population codes: a critical assessment. Neural Comput., 8: 511–529.

Zemel, R., Dayan, P. and Pouget, A. (1998) Probabilistic interpretation of population code. Neural Comput., 10: 403–430.