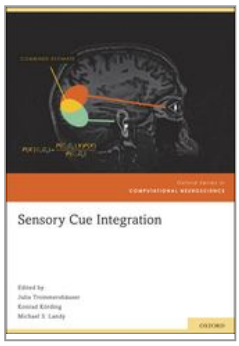


Sensory Cue Integration
Julia Trommershäuser, Konrad Kording, and Michael S. Landy

Print publication date: 2011
Print ISBN-13: 9780195387247
Published to Oxford Scholarship Online: September 2012
DOI: 10.1093/acprof:oso/9780195387247.001.0001

A Neural Implementation of Optimal Cue Integration

Wei Ji Ma
Jeff Beck
Alexandre Pouget

DOI:10.1093/acprof:oso/9780195387247.003.0021

Abstract and Keywords

This chapter lays out a theoretical framework for how optimal cue integration can be implemented by neural populations. The main significance of this framework does not merely lie in understanding multisensory perception in a principled manner, but in the fact that it provides a blueprint for finding neural implementations of other forms of Bayes-optimal computation. Evidence for Bayesian optimality of human behavior has been found in many perceptual tasks, including decision making, visual search, oddity detection, and multiple-trajectory tracking. Probabilistic population coding provides a roadmap for identifying a neural implementation of each of these computations: First the Bayesian model at the behavioral level needs to be worked out, then it needs to be assumed that probability distributions in this model are encoded in neural populations with Poisson-like variability, and finally the neural operations that map onto the desired operations on probability distributions should be identified.

Keywords: cue integration, neural populations, multisensory perception, Bayesian cue combination, probabilistic population coding

INTRODUCTION

This chapter discusses a theoretical framework for how optimal cue integration can be performed by populations of neurons. Cue integration is an interesting behavior from the perspective of neural computation for at least two reasons.

First, it is one of the simplest tasks that demonstrate that humans take uncertainty into account when perceiving the world; in other words, the brain keeps track of “error bars” on its estimates. How neurons encode and manipulate uncertainty (or more generally, probability distributions) has traditionally not been a topic of study in systems neuroscience. Therefore, explaining cue integration from a neural point of view can yield new insights into the format in which neurons represent information and the mechanisms by which they process it. The second, related reason is that cue integration is a textbook example of how a large body of psychophysical data can be used to not only constrain but also construct neural models. Many studies in computational neuroscience focus on constructing neural models that are biophysically realistic and reproduce dynamics observed in physiological experiments. Even when behavioral data are used to constrain such models, there are often many parameter settings that satisfy those constraints. Following an alternative approach, we start from a normative theory of behavior (in this case, optimal cue integration) and use theories of neural coding to link behavioral to neural quantities. This results in a theory stating which neural operations *should* be performed if the brain is to execute certain behaviors optimally. In this approach, biophysical realism is important but only a last step, serving merely to make concrete an implementation that has been found in terms of more abstract neural operations.

This chapter is based on a recent paper by the authors (Ma, Beck, Latham, & Pouget, 2006). Here, however, we will attempt a more didactic approach and highlight the broader context of the work. In previous chapters, we have seen that the problem of cue integration can be formulated in terms of the multiplication of probabilities,

$$p\left(s|x_v, x_A\right) \propto p\left(x_v|s\right) p(x_A|s)$$

(21.1)
(see Eq. 1.7), where s is the stimulus, and x_v and x_A are the noisy observations from two cues (for concreteness, we use the subscripts V for visual and A for auditory). Our first goal is to establish how a neural population can encode a probability distribution over the stimulus. After that, we will examine how the multiplication is implemented.

Before delving into specifics, it is important to comment on the use of the word *optimality*, which has different meanings depending on context and author. In this chapter, we discuss optimality in a very specific sense, namely one in which an observer computes the posterior probability distribution over the task variable of interest (as defined later). This is not **(p.394)** necessarily the same meaning as that of the term “ideal observer, ” which commonly indicates an observer who extracts all information that is present in a physical stimulus. In contrast, a Bayes-optimal observer is optimal in the sense that *during a*

particular computation, such as cue integration, no information is lost. However, it is possible that input information is lost before the computation, that is, that the probability distributions that enter the computation are broader than the ones that could be extracted from the stimulus.

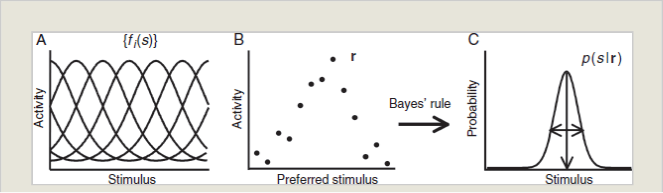
NEURAL VARIABILITY

Our first task is to understand how populations of neurons encode a stimulus. Our starting point is a variable s that is of interest to the organism. This can be the slant of a surface, the width of an object, the spatial location of an event, the speed of a moving object, the identity of a spoken syllable, and so forth. We assume that each presentation of a particular value of this stimulus variable (this particular value is also denoted s) elicits activity in a large population of neurons. Activity is characterized as the total number of evoked action potentials (spikes). An important feature of this response is that when the same value s is presented repeatedly, this spike count typically varies (Tolhurst, Movshon, & Dean, 1982). This variability has been measured in many areas of cortex. In an attempt to model it, people often assume that it obeys a Poisson distribution. This reflects the absence of temporal correlations between the spikes and implies that the probability of a spike count r in response to a stimulus s is

$$p(r|s) = \frac{e^{-\lambda}\lambda^r}{r!} \tag{21.2}$$

In this equation, λ stands for the mean spike count, which in a Poisson process is identical to the variance of the spike count. Observed variability is not exactly Poisson—we will address this issue later. The mean spike count λ depends on two things: the stimulus presented, s, and which neuron in the population is considered, i. Therefore, it can be written as $\lambda = gf_i(s)$, where g is an overall scaling factor (the gain). As a function of s, $f_i(s)$ is called the tuning curve of the i' th neuron; it is typically bell shaped or monotonic. If s is a spatial variable, then $f_i(s)$ is a receptive field. An example of a set of tuning curves (for different i) is shown in Figure 21.1A.

Since there are multiple neurons in the population, their responses to s have to be considered jointly. A population pattern of activity is a set of spike counts $(r_1, r_2, \ldots, r_N)$, where N is the number of neurons in the population. An example is shown in Figure 21.1B.



(p.395) The simplest assumption regarding the neurons' joint activity is that all neurons are uncorrelated. In other words, for a given s, the responses r_i taken across all possible i are independent of each other. That

means that the probability of observing pattern $(r_1, r_2, \ldots, r_N)$ is equal to the product of the probabilities of all individual responses r_i under their respective distributions $p(r_i|s)$: (21.3)

From now on, we will use shorthand notation $\mathbf{r}$ for the vector $p\left(r_1, r_2 \ldots, r_N|s\right) = p\left(r_1\right)$ andΠ for a product. Combining Eqs. 21.2 and 21.3, and $\lambda_i = \{r_i\} = gf_i\left(s\right)$, we find for neural population variability:

Figure 21.1 Schematic illustration of probabilistic population coding. (A) Bell-shaped tuning curves of six neurons in a hypothetical population. (In real data, these do not look nearly as smooth and identical.) Notation: s is the stimulus value, i labels the neuron, and $f_i(s)$ is the average activity of the i' th neuron in response to s. (B) Population pattern of activity elicited by a stimulus (e.g., the orientation of a line segment) on a single trial. Neurons are ordered by their preferred stimuli. (C) Based on this pattern of activity $\mathbf{r}$, Bayes' rule computes the probability distribution $p(s|\mathbf{r})$ over the stimulus (right), providing not only the most likely value of the stimulus (indicated by the arrow) but also its uncertainty (indicated by the double arrow). One can think of the probability distribution in (C) as being *encoded* in the pattern of activity in (B).

$$p\left(\mathrm{r}|s\right) = \prod_{i=1}^{N} p\left(r_i|s\right) = \prod_{i=1}^{N} \frac{e^{-gf_i(s)}\left(gf_i\left(s\right)\right)^{r_i}}{r_i!}.$$

(21.4)

If you know g and $f_i\left(s\right)$, this equation allows you to calculate for each possible pattern of activity $\mathbf{r}$ (and there are a lot of them) the probability that it will occur when the stimulus was s. These probabilities will normally be different for different s. Therefore, the activity pattern $\mathbf{r}$ is informative about s. The observer's brain does not have knowledge of s, and its task is exactly to make a guess about s based on $\mathbf{r}$—in other words, to decode s from $\mathbf{r}$. There are many recipes in the literature to make such a guess, such as winner take all, population vector, and maximum likelihood. Some of these decoding methods are better than others, but they have in common that they all return a single value of s on a single trial.

PROBABILISTIC POPULATION CODES

For optimal cue integration, it is critical that a population of neurons encodes a *probability distribution* over the stimulus on a single trial, not just a best guess about the stimulus. This is where Bayes' rule comes in. In Eq. 21.4, we fixed s and considered the probabilities of different $\mathbf{r}$. Now, we do the reverse: We fix $\mathbf{r}$

(interpreted as the observed pattern of activity on a single trial) and we consider the probabilities of different s. This is the “inverse problem” that the brain has to solve. Bayes' rule explains the latter probability, $p(s|\mathbf{r})$, in terms of the former, $p(s|\mathbf{r})$:

$$p(s|\mathbf{r}) = \frac{p(\mathbf{r}|s)\,p(s)}{p(\mathbf{r})}.$$

(21.5)

This type of coding is called *probabilistic population coding*. The left-hand side is called the posterior distribution, while $p(s|\mathbf{r})$ is called the likelihood function when considered as a function of s. (This is somewhat confusing: Even though $\mathbf{r}$ is the first argument in $p(s|\mathbf{r})$, one can still regard it as a function of s by fixing $\mathbf{r}$ and considering different possible values of s. Some people use a notation like $L_{\mathbf{r}}(s)$ to denote a likelihood over s.) In this equation, the distribution $p(s)$ reflects prior knowledge that the observer has about the stimulus, that is, beliefs held about s before any data ($\mathbf{r}$) are observed. In this chapter, we will choose this prior distribution to be uniform (flat); that is, the observer is completely agnostic (if this distribution is not flat, it is still possible to perform optimal cue integration using neural populations, essentially by regarding the prior information as another cue). Moreover, since the left-hand side is a probability distribution over s, factors on the right-hand side that do not depend on s are irrelevant except for the fact that they serve as a normalization. Therefore, we will from now on use the formulation

$$p(s|\mathbf{r}) \alpha p(\mathbf{r}|s).$$

(21.6)

Combining Eqs. 21.4 and 21.6 allows us to write down the posterior distribution for a population pattern of activity drawn from an independent Poisson distribution:

$$p(s|\mathbf{r}) \alpha \exp \sum_{i=1}^{N} \left(-g f_i(s) + r_i \log f_i(s)\right),$$

(21.7)

where we have absorbed factors independent of s into the proportionality sign.

(p.396) Note that $p(s|\mathbf{r})$ is not a probability distribution in the classical, frequentist sense of “expected frequencies of possible outcomes of a random variable, ” since we are considering only a single trial. Based on a single $\mathbf{r}$, frequentists would only reconstruct a single value of s, the “best guess” we mentioned earlier. Instead, $p(s|\mathbf{r})$ is to be interpreted as the degree of belief in a hypothesized value of the stimulus. Given a single observation $\mathbf{r}$, one asks to what extent one believes that stimulus value s caused $\mathbf{r}$. It should be emphasized

that in this view, Bayesian inference is not characterized by the presence of a nontrivial prior distribution. The framework allows for incorporation of prior beliefs, but the mere fact that we consider $p(s|\mathbf{r})$ a legitimate distribution over s makes our approach Bayesian.

Although Bayes' rule follows directly from the basic properties of probability distributions, its implications in our context are profound.

It means that based on a single pattern $\mathbf{r}$, one cannot only reconstruct the value of s most likely to have caused $\mathbf{r}$ (which is what a maximum-likelihood decoder would do) but also the probability that *any* value of s caused $\mathbf{r}$. This was first proposed in the 1990s (Földiák, 1993; Sanger, 1996). An example is shown in Figure 21.1C. The width of this probability distribution over s is interpreted as the uncertainty about the stimulus (the terms *certainty*, *reliability*, or *fidelity* are sometimes used for the inverse quantity). Note that this width is in general different from the width of the tuning curve. Even though the probability distribution and the population pattern of activity are related and can both be bell shaped, their relationship is indirect, as indicated by Eq. 21.7.

As a consequence, if the task is to estimate s from $\mathbf{r}$, the observer has knowledge of the confidence in the decision without the need of a confidence-estimation mechanism separate from the stimulus representation. However, the merits of this type of coding are not limited to representing the confidence about a decision. Since any neural population that encodes a stimulus variable simultaneously also encodes uncertainty about this variable, this code allows for the easy propagation of uncertainty through all brain areas involved in a perceptual computation. The population activity $\mathbf{r}$ that encodes a probability distribution over the stimulus will carry uncertainty information with it whenever it is manipulated. This is the main argument we will lay out in the rest of this chapter.

Other schemes have been proposed to encode probability distributions in neural activity. In some, neural activity (either on a single trial or averaged over many trials) is linearly related to probability (“explicit” coding) or to the log of probability. In those schemes, the width of the tuning curve (when bell shaped) is equal to the width of the probability distribution. Probabilistic population coding stands out by being the only scheme that bases beliefs about the stimulus on the observed neural variability. When tuning curves are not bell shaped but monotonic, the difference between probabilistic population codes and “explicit” coding becomes very clear (see Ma, Beck, & Pouget, 2008, for a review).

CAN ONE “MEASURE” THE POSTERIOR DISTRIBUTION?

Before we consider cue integration using probabilistic population codes, we first address the common question of how the posterior distribution relates to behavior as measured through a psychophysical experiment. It is tempting to

identify the posterior distributions with the distribution of stimulus estimates across many trials, but this is a mistake. The posterior distribution, $p(s|\mathrm{r})$, reflects the observer's beliefs about the stimulus on a single trial. When a decision needs to be made, a single value μ is extracted from the posterior distribution (for example, its mode, mean, or median, depending on the cost function one uses). This value is the model observer's response, $\hat{s} = \mu$. These responses can be collected over many trials, keeping the true stimulus, s_0, the same. This creates a response distribution, $p(\hat{s}|\mathrm{s}_0)$. There is no reason why $p(\hat{s}|\mathrm{s}_0)$ should have the same shape or functional form as $p(s|\mathrm{r})$. The reason that this is nevertheless often believed is because only the Gaussian case is considered. In behavioral modeling, the internal representation of a stimulus is typically taken to be μ, not $\mathbf{r}$, and **(p.397)** from the outset, a Gaussian distribution $p(\mu|s)$ is assumed. Under uniform priors, this choice makes both $p(s|\mu)$ and $p(\hat{s}|s_0)$ Gaussian with the same variance. Operations performed on $p(s|\mu)$ are then directly mirrored in operations on $p(\hat{s}|s_0)$, just like we will see in this chapter. However, in the presence of nonuniform priors (Stocker & Simoncelli, 2006), or when the posterior is non-Gaussian (Körding et al., 2007), this is no longer true and identifying the posterior with the response distribution leads to wrong predictions. It is important to keep this caveat in mind.

Now we can turn to the problem of cue integration. Suppose there are two cues to the same stimulus attribute, which we will call auditory (A) and visual (V) for convenience. Each cue is represented by a neural population of N neurons; we will denote their patterns of activity by r_A and r_V. We assume that these patterns are both drawn from independent Poisson distributions, and even that they have identical tuning curves $f_i(s)$ (there are a lot of assumptions here, and we will relax all of them eventually). They only differ in their gains: The mean activities of the auditory neurons are equal to $g_A f_i(s)$, whereas the mean activities of the visual neurons are $g_V f_i(s)$. A higher gain implies narrower posteriors and less uncertainty. Choosing different gains for the auditory and the visual cues reflects that they come with different degrees of uncertainty. Moreover, gain can change from trial to trial.

We are now interested in the optimal posterior distribution that is encoded by r_A and r_V together, because the psychophysics of cue integration suggests that this distribution is computed in the brain. Using Bayes' rule, the posterior takes the following form:

$$p(s|\mathrm{r}_A, \mathrm{r}_V) \propto p(\mathrm{r}_A, \mathrm{r}_V) \propto p(\mathrm{r}_A, \mathrm{r}_V)\, p(\mathrm{r}_V|s)\,.$$

(21.8)

In going from the second to the third expression, we have assumed that the cues are conditionally independent given the stimulus, just as in the behavioral theory. We substitute Eq. 21.4 and absorb all factors independent of s into the proportionality sign. This gives

$$p\left(s|\mathrm{r}_A,\mathrm{r}_V\right) \qquad \alpha\left(\prod_{i=1}^{N} e^{-gAfi(s)} f_i(s)^{rAi}\right) \times\left(\prod_{i=1}^{N} e^{-gvf_i(s)} f_i(s)^{rVi}\right).$$

(21.9)

This can be rewritten as:

$$p\left(s|\mathrm{r}_A,\mathrm{r}_V\right)\alpha\exp\sum_{i=1}^{N}\left(-\left(g_A+\mathrm{g}_V\right)f_i\left(s\right)+\left(\mathrm{r}_{Ai}+\mathrm{r}_{Vi}\right)\log f_i\left(s\right)\right).$$

(21.10)

This is the posterior distribution encoded by r_A and r_V together. However, r_A and r_V are separate populations—what we want instead is a single multisensory population. To implement optimal cue integration neurally means to ask what operation we can perform on r_A and r_V such that the resulting multisensory population encodes the optimal posterior distribution, Eq. 21.10. In that way, we would not lose any information about s. The right-hand side of Eq. 21.10 suggests the answer: addition. We construct a new population pattern of activity, r_{AV}, by summing the activities of corresponding pairs of neurons in the auditory and visual populations:

$$\mathrm{r}_{AV}=\mathrm{r}_A+\mathrm{r}_V.$$

(21.11)

The output population pattern, r_{AV}, will still obey independent Poisson variability across many trials, since the sum of two Poisson processes is again Poisson. The mean activity of the i' th neuron in the output population in response to s is $\left(g_A+g_V\right)f_i\left(s\right)$. Therefore, it encodes a posterior distribution that is given by:

$$p\left(s|\mathrm{r}_{AV}\right)\alpha\exp\sum_{i=1}^{N}\left(-\left(g_A+\mathrm{g}_V\right)f_i\left(s\right)+\left(\mathrm{r}_{Ai}+\mathrm{r}_{Vi}\right)\log f_i\left(s\right)\right).$$

(21.12)

This distribution is identical to the one in Eq. 21.10. We conclude that adding **(p.398)** independent Poisson population patterns of activity implements a multiplication of the probability distributions over the stimulus that are encoded

in those patterns. We will now attempt to generalize this framework to other forms of neural variability.

POISSON-LIKE VARIABILITY

We assumed earlier that neurons fire with Poisson statistics and that their noise is independent. Both assumptions are often not completely satisfied. In cortical neurons, spike count variance is often proportional to, but not equal to, spike count mean. The ratio variance/mean is called the *Fano factor*, and its measured values range from 0.3 to 1.8 (Gur & Snodderly, 2006; Tolhurst et al., 1982). Moreover, neurons are not independent (when conditioned on the stimulus) but exhibit correlations (Averbeck, Latham, & Pouget, 2006). Therefore, a more general treatment is needed. Fortunately, there is a family of distributions that is more general than independent Poisson variability but leaves the mechanism for implementing optimal cue integration intact. This family is the exponential family with linear sufficient statistics, also called Poisson-like variability. It takes the following form:

$$p\left(r|s\right) = \frac{\Phi\left(\mathrm{r}\right)}{\eta\left(s\right)}e^{h(s)\text{-r}}$$

(21.13)

where $\Phi\left(\mathrm{r}\right)$ is an arbitrary function of $\mathbf{r}$, $\mathbf{h}(s)$is a vector-valued function of s that we will specify later, and $\eta(s)$ serves as a normalization (since this is a probability distribution over $\mathbf{r}$). The exponent contains the inner product of $\mathbf{h}(s)$ with the population pattern of activity. To gain some intuition for Eq. 21.13, it helps to see what Φ, h, and η are for independent Poisson variability:

$$\begin{aligned}\Phi\left(\mathrm{r}\right) &= \frac{1}{\prod_i r_i!}, \eta\left(s\right) = e^{g\sum_i fi(s)} \\ h_i\left(s\right) &= \log f_i\left(s\right).\end{aligned}$$

(21.14)

Thus, $\mathbf{h}$ and η both depend on the tuning curve, whereas Φ contains all factors that only depend on $\mathbf{r}$. In general, when variability is Poisson-like but not necessarily independent Poisson, there is a relationship between $\mathbf{h}(s)$ and the tuning curve. By using the definition, Eq. 21.13, it is possible to show that the derivative of $\mathbf{h}$ satisfies:

$$\mathrm{h'}\left(s\right) = \sum^{-1}\left(s\right)\mathrm{f'}\left(s\right),$$

(21.15)

where $\sum^{-1}\left(s\right)$ is the inverse of the covariance matrix of the population and $\mathbf{f}(s)$ is the mean activity. For independent neurons, the covariance matrix is a diagonal matrix, with the variances of the neural activities on the diagonal. In

the case of independent Poisson variability, it is

$$\Sigma(s) = \begin{pmatrix} f_1(s) & 0 & \cdots & 0 \\ 0 & f_2(s) & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & f_N(s) \end{pmatrix}. \tag{21.16}$$

Therefore, Eq. 21.15 implies that $h_i'(s) = \frac{f_i'(s)}{f_i(s)}$, which is, up to a constant, equivalent to the second part of Eq. 21.14. This confirms that independent Poisson variability is indeed a special case of Poisson-like variability.

DEALING WITH "IRRELEVANT" VARIABLES

When a task involves estimating a stimulus, such as a spatial location of an event, the size of a visual object, or the direction of motion of a moving object, neural activity is often affected by parameters that are not of interest to the observer. For example, the visual object might come with greater or lesser contrast, or the motion might be more or less coherent. Such variables, which affect the difficulty of the task but are irrelevant when estimating the stimulus, are called *nuisance parameters*. Nuisance parameters are extremely common in any real-world perceptual task, for example, in object recognition (Kersten, Mamassian, & Yuille, 2004). The brain is faced with this problem of estimating the **(p.399)** stimulus while not knowing the value of the nuisance parameters. In this section, we describe under what conditions Poisson-like variability solves this problem. This section is more technical than the rest of this chapter and can be skipped without affecting the understanding of the overall line of reasoning. The main results of this section are Eqs. 21.21 and 21.22.

We will examine the specific situation in which nuisance parameters affect the gain of the population. For example, if $\mathbf{r}$ is a visual population, then the gain g could be determined by the contrast of the stimulus. In the formulation of the Poisson-like family, Eq. 21.13, we did not include the gain. Yet each of the factors in Eq. 21.13 could in principle depend on it. For example, in Eq. 21.14, the expression for $\eta(s)$ contains the gain. This does not pose a problem if the gain is known, but in general this is not the case. There are two types of solutions to this problem. The first is to use an external mechanism to estimate the gain and substitute its estimated value. However, this requires extra computational resources and it is not an optimal solution. The second solution is the Bayes-optimal one: When a parameter whose value is unknown influences the observations, it is averaged out (also called "integrated out" or "marginalized out"). This is done as follows:

$$p(s|\mathrm{r}) \propto p(\mathrm{r}|s) = \int p(\mathrm{r}|s,g)\, p(g)\, dg, \tag{21.17}$$

where $p(g)$ is some prior distribution over g.In writing this integral, we have assumed that g does not depend on s. To avoid the complications of marginalizing out gain, we will restrict ourselves to the situation where the s and g dependencies of $p(\mathbf{r}|s,g)$ can be separated, that is, where $p(\mathbf{r}|s,g)$ can be written as the product of a factor that only depends on s and $\mathbf{r}$ and one that only depends on g and $\mathbf{r}$. Looking back at Eq. 21.13, this means that it would take the following form:

$$p(\mathrm{r}|s,g) = \frac{\Phi(\mathrm{r},g)}{\eta(s)} e^{h(s)\cdot \mathrm{r}}$$

(21.18)

How does this help? If we substitute Eq. 21.18 into the integral of Eq. 21.17, then we find:

$$\begin{aligned} & p(s|\mathrm{r}) \alpha \int \frac{\Phi(\mathrm{r},g)}{\eta(s)} e^{h(s)\cdot} p(g)\, dg \\ & = \frac{e^{h(s)\cdot \mathrm{r}}}{\eta(s)} \int \Phi(\mathrm{r},g)\, p(g) \\ & \alpha \frac{e^{h(s)\cdot \mathrm{r}}}{\eta(s)}, \end{aligned}$$

(21.19)

where we can go from the second to the third line because the integral does not depend on s, no matter what $p(g)$ is.

It turns out that another constraint must be satisfied. To derive this constraint, we first compute the derivative of $\eta(s)$ with respect to s, keeping in mind that $\eta(s)$ is a normalization factor:

$$\begin{aligned} \frac{d}{ds}\log\eta(s) &= \frac{1}{\eta(s)}\frac{d}{ds}\int \Phi(\mathrm{r},g)\, e^{h(s)\cdot \mathrm{r}} d\mathrm{r} \\ &= \mathrm{h'}(s)\cdot \int \frac{r\Phi(\mathrm{r},g)\, e^{h(s)\cdot \mathrm{r}}}{\eta(s)} \\ &= \mathrm{h'}(s)\cdot(\mathrm{r}) = \mathrm{h'}(s)\cdot g\mathrm{f}(s). \end{aligned}$$

(21.20)

Now, differentiating both sides with respect to g gives $0 = \mathrm{h'}(s)\cdot \mathrm{f}(s)$. Substituting this back into Eq. 21.20 implies that $d\eta/ds = 0$. Surprisingly, this condition is not very hard to meet. For example, in the independent Poisson case, Eq. 21.14, it seems as though η depends both on s and on g. However, if tuning curves are translation invariant and many of them span the stimulus space, as in Figure 21.1A, then the sum $\sum_i f_i(s)$ will be nearly independent of s.

This means that η only depends on g and can therefore be absorbed into $\Phi(\mathrm{r},\mathrm{g})$.

To allow Eq. 21.19 to be true, it is also important that $\mathbf{h}(s)$ does not depend on g. Both the mean activity and the covariance matrix can depend on the gain g, and in general they will, but the combination $\sum^{1}(s,g)\mathrm{f}'(s,g)$ (from Eq. 21.15) cannot. Since $\mathrm{f}(s,g)=g\mathrm{f}(s)$, this means that the covariance matrix must be of the form $\sum(s,g)=g\sum(s)$. On the diagonal, this **(p.400)** means that the variance scales with the gain, in other words, that the Fano factor is constant but not necessarily equal to 1 (as it is in a Poisson process). Off diagonal, it means that the entries $\langle r_i r_i\rangle - g^2 f_i(s) f_i(s)$ should be proportional to as well. In conclusion, we find that neural variability must be of the form

$$p(\mathrm{r}|s)=\Phi(\mathrm{r},g)\,e^{h(s)\cdot\mathrm{r}},$$

(21.21)
and the posterior distribution over s is simply

$$p(\mathrm{r}|s)=\alpha e^{h(s)\cdot\mathrm{r}}.$$

(21.22)
In the more general case in which nuisance parameters $\mathbf{c}$ exist that affect the tuning curves and the covariance matrix, but not through a simple gain modulation, then Poisson-like variability requires that $\sum^{-1}(s,\mathrm{c})\mathrm{f}'(s,c)$ be independent of $\mathbf{c}$.

Whether cortical variability is approximately Poisson-like is an open question that can in principle be addressed by analyzing population recordings. We just stated that Poisson-like variability requires that the elements of the covariance matrix, including the variance, scale with the gain of the population. This condition seems to be roughly satisfied in cortical neurons (Gur & Snodderly, 2006; Tolhurst et al., 1982), but further study is needed. Furthermore, if variability is Poisson-like, the locally optimal linear decoder (Seriès, Latham, & Pouget, 2004) should extract all available information. This property can be falsified by trying other decoders, such as support vector machines (Bishop, 2006). Moreover, the optimal decoder of Poisson-like activity is completely determined by $\mathbf{h}(s)$ and should therefore be independent of nuisance parameters.

OPTIMAL CUE INTEGRATION WITH POISSON-LIKE POPULATIONS

We claim that the same addition operation that implements optimal cue integration in the independent Poisson case, $\mathrm{r}_{AV}=\mathrm{r}_A+\mathrm{r}_V$, does the same in the more general Poisson-like case. We verify this by first calculating the distribution of the sum random variable, r_{AV}, for a given stimulus s and input gains g_V and g_A. That is done in the same way one would calculate the distribution of the total number rolled with two dice, namely by summing (or integrating) over all possible values of one of the terms in the sum:

$$
\begin{aligned}
&p\left(\mathbf{r}_{AV}|s,g_V,g_A\right)\\
&=\int p\left(\mathbf{r}_V|s,g_V\right)\\
&\times p\left(\mathbf{r}_A=\mathbf{r}_{AV}-\mathbf{r}_V|s,g_A\right)d\mathbf{r}_V\\
&=\int \Phi_V\left(\mathbf{r}_V,g_V\right)e^{\mathbf{h}(s)\cdot\mathbf{r}V}\\
&\times\Phi_A\left(\mathbf{r}_{AV}-\mathbf{r}_V,g_A\right)e^{h(s)\cdot(rAV-\mathbf{r}V)}d\mathbf{r}_V\\
&=\int \Phi_A\left(\mathbf{r}_{AV}-\mathbf{r}_V,g_A\right)e^{h(s)\cdot\mathbf{r}_{AV}dr_V}\\
&\times\Phi_A\left(\mathbf{r}_{AV}-\mathbf{r}_V,g_A\right)e^{h(s)\cdot AV}dr_V\\
&=e^{h(s)\cdot AV}\int \Phi_V\left(\mathbf{r}_V,g_V\right)\\
&\times\Phi_A\left(\mathbf{r}_{AV}-\mathbf{r}_V,g_A\right)d\mathbf{r}.
\end{aligned}
$$

(21.23)

Note that in the transition from the second to the third line, it is essential that $\mathbf{h}(s)$ is the same in both populations. As a consequence of Eq. 21.23, the posterior encoded in a multisensory population pattern of activity can be computed in analogy to Eq. 21.22. We find

$$p\left(s|\mathbf{r}_{AV}\right)\alpha e^{h(s)\cdot\mathbf{r}AV}.$$

(21.24)

When we substitute $\mathbf{r}_{AV}=r_A+\mathbf{r}_V$ (i.e., giving the random variable $\mathbf{r}_{AV}$ the value $\mathbf{r}_A+\mathbf{r}_V$, it follows that

$$p\left(s|\mathbf{r}_{AV}=\mathbf{r}_A+\mathbf{r}_V\right)\alpha p\left(\mathbf{r}_A|s\right)p\left(\mathbf{r}_V|s\right),$$

(21.25)

where we again assumed that $\mathbf{h}(s)$ is the same function for auditory and visual inputs. Thus, just like for independent Poisson variability, optimal cue integration is achieved in Poisson-like populations by adding input population activities.

Our assumption that $\mathbf{h}(s)$ is the same in both input populations is violated if tuning curves or **(p.401)** covariance matrices differ between auditory and visual areas. However, different $\mathbf{h}(s)$ can be dealt with as long as they can be linearly mapped onto a common basis of functions, that is, $\mathbf{h}_A\left(s\right)=\mathbf{W}_AH\left(s\right)$ and $\mathbf{h}_V\left(s\right)=\mathbf{W}_VH\left(s\right)$, where W_A and W_V are stimulus-independent matrices and $\mathbf{H}(s)$ is the common basis. Then, it can be shown that the linear combination

$$\mathbf{r}_{AV}=\mathbf{W}_A^T\mathbf{r}_A+\mathbf{W}_V^T\mathbf{r}\mathrm{V}$$

(21.26)

(where the superscript “T” denotes a transpose) implements optimal cue integration (for details, see the Supplement of Ma et al., 2006). This more

general operation is depicted in Figure 21.2. Notably, the weights do not depend on neural gain or on uncertainty. Having to adjust the weights every time gain or uncertainty changes would make a neural implementation much more difficult. By using Poisson-like variability, the brain can avoid this problem, since the weights are learned once and for all.

A PHYSIOLOGICAL PREDICTION

Our most general prediction for physiology is that the activity in a multisensory area involved in optimal cue integration is equal to a linear combination of the activities in the input populations (Eq. 21.26). If one would present only one of both inputs, then only one of the terms in $\mathbf{r}_{AV}$ (either $\mathbf{W}_A^T\mathbf{r}_A$ or $\mathbf{W}_V^T\mathbf{r}_\mathrm{V}$) would be nonzero. It follows that the activity evoked in a multisensory area by both cues presented simultaneously is predicted to be approximately equal to the sum of the activities evoked by each individual cue separately. This property is called *additivity*. Recent physiological studies have begun to test this prediction (see Chapter 16). Earlier work had claimed that multisensory interactions are characterized by *superadditivity* of multisensory responses (Stein & Meredith, 1993), that is, that the multisensory activity evoked **(p.402)** by both cues is more than the sum of the activities evoked by the individual cues. However, this notion has become largely discredited, based on new physiological data (see Ma & Pouget, 2008, for a review).

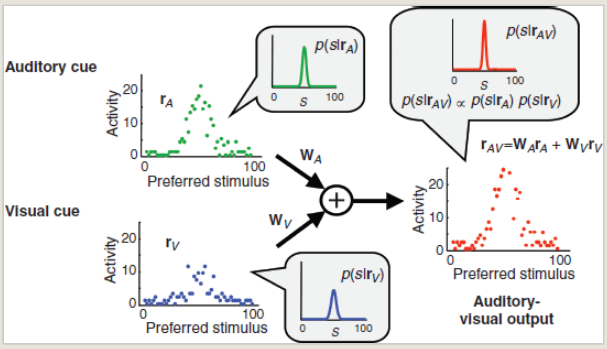


Figure 21.2 Optimal cue integration with probabilistic population codes. The cues elicit activity in input populations $\mathbf{r}_A$ and $\mathbf{r}_\mathrm{V}$, indicated by green and blue dots. The dialogue boxes show the probability distributions over the stimulus encoded in each population on a single trial. A simple linear combination of the population patterns of activity, $\mathbf{r}_{AV} = \mathbf{W}_A\mathbf{r}_A + \mathbf{W}_V\mathbf{r}_\mathrm{V}$, guarantees optimal cue integration, if neural variability is Poisson-like. Optimal cue integration means that the probability

RELATING BACK TO BEHAVIOR

We will now examine how the neural operation $\mathbf{r}_{AV} = \mathbf{r}_A + \mathbf{r}_V$ relates to the behavioral equations for multisensory mean and variance that we encountered in Chapter 1. In behavioral modeling of cue integration discussed in this book, it is assumed that the posterior distribution $p(s|\mathbf{r})$ is Gaussian. Therefore, it is exponential with a quadratic function of s in the exponent:

distribution over the stimulus encoded in the multisensory population is a product of the distributions encoded in the unisensory populations, i.e. $p(s|\mathbf{r}_{AV}) \alpha p(s|\mathbf{r}_A)\, p(s|\mathbf{r}_V)$. The synaptic weight matrices $\mathbf{W}_A$ and $\mathbf{W}_V$ depend on the tuning curves and covariance matrices of the input populations, but they do not have to be adjusted over trials.

$$p(s|\mathbf{r}) \alpha e^{-\frac{1}{2}s^2 a(\mathbf{r}) + s b(\mathbf{r})},$$

(21.27)

where $a(\mathbf{r})$ and $b(\mathbf{r})$ are functions of $\mathbf{r}$. Comparing with Eq. 21.22, we see that these functions must be of the form $a(\mathbf{r}) = a \cdot \mathbf{r}$ and $b(\mathbf{r}) = b \cdot \mathbf{r}$, where now $\mathbf{a}$ and $\mathbf{b}$ are constant vectors. From Eq. 21.27, we can find the mean μ and variance σ^2 of the Gaussian, since the exponent of a Gaussian is of the form $-\frac{(s-\mu)^2}{2\sigma^2} = -\frac{s^2}{2\sigma^2} + \frac{s\mu}{\sigma^2}$+constant. They are given by

$$\frac{1}{\sigma^2} = a(\mathbf{r}) = \mathbf{a} \cdot \mathbf{r}$$

(21.28)

and

$$\frac{\mu}{\sigma^2} = b(\mathbf{r}) = \mathbf{b} \cdot \mathbf{r}.$$

(21.29)

Since $\mathbf{r}_{AV} = \mathbf{r}_A + \mathbf{r}_V$ for optimal cue integration, applying the inner product with $\mathbf{a}$ gives (from Eq. 21.28):

$$\frac{1}{\sigma_{AV}^2} = \frac{1}{\sigma_A^2} + \frac{1}{\sigma_V^2},$$

(21.30)

which is the *single-trial* version of our well-known equation for optimal combination of variances. For the mean, Eq. 21.29 gives

$$\frac{\mu_{AV}}{\sigma_{AV}^2} = \frac{\mu_A}{\sigma_A^2} + \frac{\mu_V}{\sigma_V^2}$$

(21.31)

(the extra assumption here is that the trial-to-trial fluctuations in the inverse variance are small). This is the *single-trial* version of the optimal combination of means. Now, we can look at the effect across many trials. This requires a way to turn $p\left(s|\mathrm{r}\right)$ into a single estimate of the stimulus, $\hat{s}\left(\mathrm{r}\right)$, on each trial. The optimal estimator is the maximum-likelihood estimator, which chooses the s that makes $p\left(\mathrm{r}|\mathrm{s}\right)$ maximal. Since we have assumed that the prior is uniform, this is also the estimator that maximizes $p\left(\mathrm{s}|\mathrm{r}\right)$. For a Gaussian $p\left(\mathrm{s}|\mathrm{r}\right)$, this is simply the mean of the Gaussian. We can now calculate the variance of this estimate using the so-called Cramèr-Rao bound, which states that the inverse variance (sampled across many trials) of an optimal estimator is given by the *Fisher information*:

$$\frac{1}{\sigma_{\text{estimator}}^2} = I\left(s\right) \equiv -\left\langle \frac{\partial^2}{\partial s^2} \log p\left(\mathrm{r}|s\right) \right\rangle,$$

(21.32)

where the average $\langle . \rangle$ is over $\mathbf{r}$ drawn from $p\left(\mathrm{r}|s\right)$. Fisher information and the Cramèr-Rao can be applied to any distribution. For Poisson-like variability, there are several ways to express Fisher information:

$$\begin{aligned} I\left(s\right) &= -\mathrm{h}''\left(s\right) \cdot g\mathrm{f}\left(s\right) = g\mathrm{h}'\left(s\right) \cdot \sum\left(s\right) \cdot \mathrm{h}'\left(s\right) \\ &= g\mathrm{f}'\left(s\right) \cdot \sum{}^{-1}\left(s\right) \mathrm{f}'\left(s\right). \end{aligned}$$

(21.33)

This offers a particularly easy way to check optimality of $\mathrm{r}_{AV} = \mathrm{r}_A + \mathrm{r}_V$. Taking the average on both sides, we find $\mathrm{g}_{AV} = \mathrm{g}_A + \mathrm{g}_V$, and since $I\left(s\right)$ is proportional to g according to Eq. 21.33, it follows from Eq. 21.32 the estimate's inverse variances sum:

$$\frac{1}{\sigma_{\text{estimator, }\ AV}^2} = \frac{1}{\sigma_{\text{estimator, }\ AV}^2} + \frac{1}{\sigma_{\text{estimator, }\ V}^2}.$$

(21.34)

This is the relationship found in behavior. Similarly, for the mean estimate, we have from Eq. 21.29:

$$\left\langle \hat{s} \right\rangle = \sigma^2 \mathrm{b} \cdot \left\langle \mathrm{r} \right\rangle = \sigma_{\text{estimator}}^2 \mathrm{b} \cdot g\mathrm{f}\left(s\right),$$

(21.35)

and therefore, $g_{AV} = g_A + g_V$ implies

(p.403)

$$\frac{\left\langle \hat{s} \right\rangle}{\sigma_{\text{estimator, }\ AV}^2} = \frac{\left\langle \hat{s} \right\rangle_A}{\sigma_{\text{estimator, }\ AV}^2} + \frac{1}{\sigma_{\text{estimator, }\ AV}^2},$$

(21.36)

which is the behavioral result for the mean multisensory estimate.

OPTIMAL CUE INTEGRATION WITH BIOPHYSICAL POPULATIONS

So far, we have used abstract neurons completely characterized by their firing rates and therefore without any dynamics. As a proof of principle, it is important to show that the same scheme can be implemented with a population of biologically more realistic neurons. We did this by tuning the parameters of a network of conductance-based integrate-and-fire neurons such that the network would mimic the linear combination operation (Eq. 21.26). The output of this network satisfied the same equations that describe human observers (Eqs. 21.34 and 21.36). It is not known how to capture, in general, the network behavior of conductance-based integrate-andfire neurons in equations only involving firing rates, and therefore a direct mapping between this network and the firing-rate neurons cannot be made. However, our approach provides an example of top-down driven neural modeling: The computational model at the behavioral level is used to construct the neural theory at an abstract level, and this in turn is used to guide a more physiologically realistic implementation. The more realistic implementation thus serves as a feasibility check, not as the centerpiece of the computational approach.

WHAT IS OPTIMAL ABOUT CUE INTEGRATION?

One might wonder whether an optimal perceptual strategy would not lead to separate, veridical percepts of the auditory and the visual stimuli. This is not necessarily the case. Although in cue-integration experiments, small conflicts between the *presented* stimuli are introduced by the researcher, small conflicts between the best *estimates* of the auditory and visual stimuli exist even in the absence of artificial conflict. This is due to the variability in the neural response, which leads to variability in the perceived auditory and visual stimuli (μ_A and μ_V in Eq. 21.31). Thus, even when the true auditory and visual stimuli are physically completely in agreement, the brain still has to solve the cue-integration problem. Of course, it is important that in experimental settings, the artificial conflicts are kept sufficiently small and infrequent, so that they can be mistaken for naturally occurring ones. If conflicts are too large or too frequent, subjects may notice that the stimuli have different sources and develop a tendency to separate their percepts. Such perception can still be modeled using Bayesian models (Körding et al., 2007; Sato, Toyoizumi, & Aihara, 2007; also Chapters 1 and 2 in this volume).

THE BIG PICTURE

In this chapter, we have laid out a theoretical framework for how optimal cue integration can be implemented by neural populations. The main significance of this framework does not merely lie in understanding multisensory perception in a principled manner, but in the fact that it provides a blueprint for finding neural implementations of other forms of Bayes-optimal computation. Evidence for Bayesian optimality of human behavior has been found in many perceptual tasks,

including decision making (Beck et al., 2008), visual search (Ma, Navalpakkam, Beck, & Pouget, 2008; Vincent, Baddeley, Troscianko, & Gilchrist, 2009), causal inference (Körding et al., 2007), oddity detection (Hospedales & Vijayakumar, 2009), and multiple-trajectory tracking (Ma & Huang, 2009). Probabilistic population coding provides a roadmap for identifying a neural implementation of each of these computations: First work out the Bayesian model at the behavioral level, then assume that probability distributions in this model are encoded in neural populations with Poisson-like variability, and finally identify the neural operations that **(p.404)** map onto the desired operations on probability distributions. This general scheme for neural computation is illustrated in Figure 21.3. We believe that this approach to neural computation provides unprecedented power for relating mental functions to their underlying neural mechanisms.

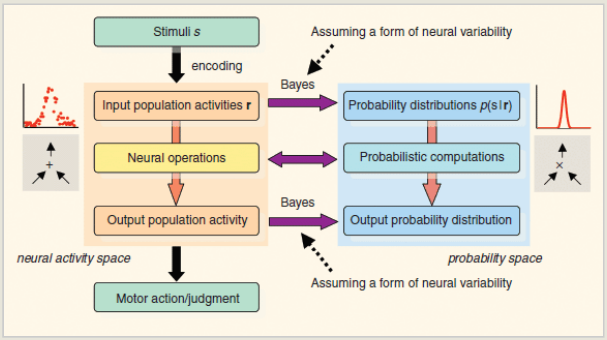


Figure 21.3 Schematic of perceptual computation using probabilistic population codes. One or multiple stimuli elicit population patterns of activity. Each pattern encodes a probability distribution over the stimulus through Bayes' rule. In perceptual tasks, these probability distributions have to be manipulated in specific ways to achieve optimality (e.g., multiplication in cue integration). The key problem is to establish a “dictionary” between such probabilistic computations (e.g., multiplication) and neural operations on population patterns of activity (e.g., addition), assuming a form of neural variability (e.g., Poisson-like). Using those neural operations, the brain will retain full probabilistic information about the variable(s) of interest at all stages of computation. Eventually, a motor action is generated or a high-level

judgment is made. (From Ma, Beck, & Pouget 2008).

REFERENCES

Bibliography references:

Averbeck, B. B., Latham, P. E.,
& Pouget, A. (2006). Neural correlations, population coding, and computation. *Nature Reviews Neuroscience*, *7*, 358–366.

Beck, J. M., Ma, W. J., Kiani, R., Hanks, T. D., Churchland, A. K., Roitman, J. D., … Pouget, A. (2008). Bayesian decision-making with probabilistic population codes. *Neuron*, *60*, 1142–1145.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Cambridge, England: Springer.

Földiák, P. (1993). The “ideal homunculus”: Statistical inference from neural population responses. In F. Eeckman & J. Bower (Eds.), *Computation and neural systems* (pp. 55–60). Norwell, MA: Kluwer Academic Publishers.

Gur, M., & Snodderly, D. M. (2006). High response reliability of neurons in primary visual cortex (V1) of alert, trained monkeys. *Cerebral Cortex*, *16*, 888–895.

Hospedales, T., & Vijayakumar, S. (2009). Multi-sensory oddity detection as Bayesian inference. *PLoS ONE*, *4*(1), e4205.

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annual Review of Psychology*, *55*, 271–304.

Körding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., & Shams, L. (2007). Causal inference in multisensory perception. *PLoS ONE*, *2*(9), e943.

(p.405) Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, *9*, 1432–1438.

Ma, W. J., Beck, J. M., & Pouget, A. (2008). Spiking networks for Bayesian inference and choice. *Current Opinion in Neurobiology*, *18*, 217–222.

Ma, W. J., & Huang, W. (2009). No capacity limit in attentional tracking: Evidence for probabilistic inference under a resource constraint. *Journal of Vision*, *9*(11):3, 1–30.

Ma, W. J., Navalpakkam, V., Beck, J. M., & Pouget, A. (2008). *Bayesian theory of visual search*. Program No. 616.1. 2008 Neuroscience Meeting Planner. Washington, DC: Society for Neuroscience. Online.

Ma, W. J., & Pouget, A. (2008). Linking neurons to behavior in multisensory perception: A computational review. *Brain Research 1242*, 4–12.

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

Sato, Y., Toyoizumi, T., & Aihara, K. (2007). Bayesian inference explains perception of unity and ventriloquism aftereffect: Identification of common sources of audiovisual stimuli. *Neural Computation*, *19*, 3335–3355.

Seriès, P., Latham, P., & Pouget, A. (2004). Tuning curve sharpening for orientation selectivity: Coding efficiency and the impact of correlations. *Nature Neuroscience*, *10*, 1129–1135.

Stein, B. E., & Meredith, M. A. (1993). *The merging of the senses*. Cambridge, MA: MIT Press.

Stocker, A. A., & Simoncelli, E. P. (2006). Noise characteristics and prior expectations in human visual speed perception. *Nature Neuroscience*, *9*, 578–585.

Tolhurst, D., Movshon, J., & Dean, A. (1982). The statistical reliability of signals in single neurons in cat and monkey visual cortex. *Vision Research*, *23*, 775–785.

Vincent, B. T., Baddeley, R. J., Troscianko, T., & Gilchrist, I. D. (2009). Optimal feature integration in visual search. *Journal of Vision*, *9*(5):15, 1–11.