A neural implementation of optimal cue integration

Wei Ji Ma, Jeff Beck, and Alexandre Pouget

This chapter discusses a 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, that 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 physiology. Even when behavioral data are used to constrain these 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. The result of this is 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 to confirm what has been found using more abstract neurons.

This chapter is based on a recent paper by the same authors (Ma, Beck, Latham, & Pouget, 2006). Here, however, we will attempt a more didactic approach, emphasize the broader context of the work, and answer frequently asked questions. In previous chapters, we have seen that the problem of cue integration can be formulated in terms of the multiplication of two conditional probability distributions [REFER TO EQUATION IN EARLIER CHAPTER]. Therefore, 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.

Q: Is Bayesian optimality the same as the ideal observer? A: Not necessarily. The term "ideal observer" can be used in different meanings, but one common meaning is that of an observer who can extract all information that is present in a physical stimulus. Such an observer is optimal in an "absolute" sense. Bayesian optimality means that the observer manipulates probability distributions over the stimulus in a way that is required by the task. This means that *during a particular computation*, such as cue integration, no information is lost. However, it is possible that the probability distributions that *enter* the computations are broader than the ones that could be extracted from the stimulus. A Bayesian observer is optimal in a "relative" sense. As a consequence, it is possible that 50% of the available information is lost, but that the observer is still Bayes-optimal. There can be many causes for information loss, but that is a different topic.

Neural variability

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, etc. 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 chracterized 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. 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

$$p\left(r \mid s\right)=\frac{e^{-\lambda}\lambda^{r}}{r!}. \tag{1}$$

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 1a.

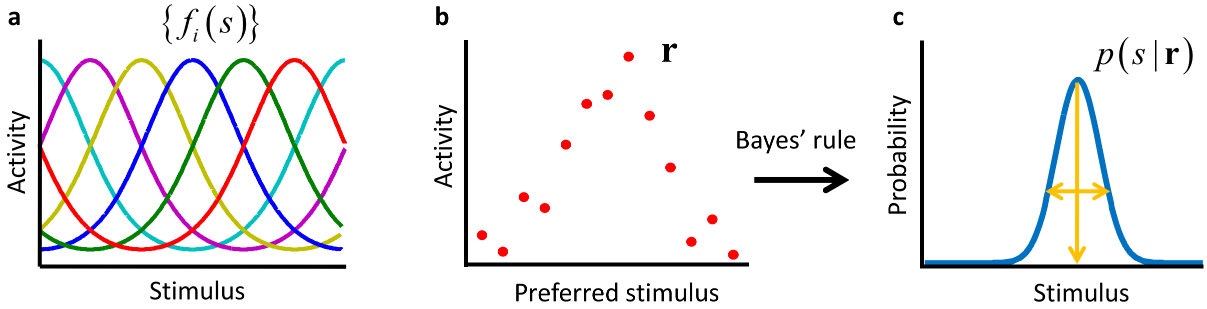

Figure 1: Schematic illustration of probabilistic population coding. a. Bell-shaped tuning curves of 6 neurons in a hypothetical population. (In real data, these do not look nearly as smooth and identical.) **b**. Population pattern of activity elicited by a stimulus (e.g. the orientation of a line segment) on a single trial. Neurons are ordered by their preferred stimuli. **c.** Based on this pattern of activity, 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).

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, …, r_N), where *N* is the number of neurons in the population. An example is shown in Figure 1b. 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, …, r_N) is equal to the product of the probabilities of all individual responses r_i under their respective distributions $p(r_i|s)$:

$$p\left(r_1,r_2...,r_N\,|\,s\right)=p\left(r_1\,|\,s\right)\cdot p\left(r_2\,|\,s\right)\cdots p\left(r_N\,|\,s\right). \tag{2}$$

From now on, we will use shorthand notation **r** for the vector (r_1, r_2, …, r_N), and Π for a product. Combining Equations (1) and (2), and $\lambda_i=\langle r_i\rangle=gf_i\left(s\right)$, we find for neural population variability:

$$\begin{aligned}p\left(\mathbf{r}\,|\,s\right)&=\prod_{i=1}^{N}p\left(r_i\,|\,s\right)\\&=\prod_{i=1}^{N}\frac{e^{-gf_i(s)}\left(gf_i\left(s\right)\right)^{r_i}}{r_i\,!}.\end{aligned} \tag{3}$$

If you know *g* and $f_i(s)$, this equation allows you to calculate for each possible pattern of activity **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 **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 **r**, in other words, to decode *s* from **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. That may be sufficient in some situations, but not in ours, since we are interested in the encoding of a *probability distribution* over the stimulus on a single trial, not just a best guess about the stimulus.

Probabilistic population codes

This is where Bayes' rule comes in. In Eq. (3), we fixed *s* and considered the probabilities of different **r**. Now, we do the reverse: we fix **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, $p(\mathbf{r}|s)$:

$$p(s\,|\,\mathbf{r})=\frac{p(\mathbf{r}\,|\,s)\,p(s)}{p(\mathbf{r})}. \qquad (4)$$

The left-hand side is called the posterior distribution, while $p(\mathbf{r}|s)$ is called the likelihood function when considered as a function of *s*. (This is somewhat confusing: even though **r** is the first argument in $p(\mathbf{r}|s)$, one can still regard it as a function of *s* by fixing **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 (**r**) are observed. In this chapter, we will choose this prior distribution to be uniform (flat), i.e. the observer is completely agnostic. 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

$$p(s\,|\,\mathbf{r})\propto p(\mathbf{r}\,|\,s). \qquad (5)$$

Q: I thought that the prior distribution was the heart and soul of Bayesian inference. How can you ignore it? A: There is a long-standing debate between so-called frequentists and Bayesians in statistics, but this debate has more to do with the very definition of probability than with the prior distribution. According to frequentists, probability is only defined as the expected frequency of occurrence of events. They would consider $p(\mathbf{r}|s)$ a real probability distribution, because one can – in principle – present *s* many times and count how often each **r** occurs. However, there is no stochastic process by which **r** gives rise to *s*, and therefore, $p(s|\mathbf{r})$ would be meaningless in their minds. Based on a single **r**, frequentists would only reconstruct a single value of *s*, the "best guess" we mentioned above. On the other hand, Bayesians interpret probability as *degree of belief*. Given a single observation **r**, a Bayesian can ask to what extent one believes that *s* caused **r**. This view allows for the 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 **r**, you not only can reconstruct the value of s most likely to have caused **r** (which is what a maximum-likelihood decoder would do), but also the probability that *any* value of s caused **r**. This was first proposed in the 1990s in two far-sighted papers (Foldiak, 1993; Sanger, 1996). An example is shown in Figure 1c. The width of this probability distribution over s is interpreted as the uncertainty about the stimulus (words like fidelity and reliability are sometimes also used). Note that this width is in general different from the width of the tuning curve. Even though both the probability distribution and the population pattern of activity are related and can both have a bell shape, their meaning is completely different in this type of coding.

As a consequence, if the task is to estimate s from **r**, the observer has knowledge of the *confidence* of his decision without the need of a confidence estimation mechanism separate from **r**, the stimulus representation. However, the merits of this type of coding (often called probabilistic population coding) do not only in the representation of confidence, but also in the propagation of uncertainty through multiple stages of computation. What we mean by this is that in many computations, such as cue integration, it is important to take into account the uncertainty of the variables involved in the computation – otherwise behavior will be suboptimal, sometimes severely so. In such computations, the representation **r** which encodes a probability distribution s, 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.

Q: Is this the only way to encode probabilities in neural activity? A: No, many other ways have been proposed. 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 codes (PPCs) are 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 PPCs and "explicit" coding becomes very clear. However, PPCs are closely related to so-called convolution codes. For a review, see (Ma, Beck, & Pouget, 2008).

Combining Eqs. (3) and (5) allows us to write down the posterior distribution for a population pattern of activity drawn from an independent Poisson distribution:

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

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

Optimal cue integration – independent Poisson case

Now we can turn to the problem of cue integration. Suppose there are two cues about the same stimulus, which we will call auditory (A) and visual (V) for convenience. Each cue is represented in a neural population of *N* neurons; we will denote their patterns of activity by $\mathbf{r}_A$ and $\mathbf{r}_V$. 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 the gain: 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 a more narrow $p(s|\mathbf{r})$ and less uncertainty – we will examine this in detail later. Choosing different gains for the auditory and the visual cues reflects that they come with different degrees of uncertainty.

We are now interested in the optimal posterior distribution that is encoded by $\mathbf{r}_A$ and $\mathbf{r}_V$ together, because the psychophysics of cue integration suggests that this distribution is computed in the brain. Using Bayes' rule, it takes the following form:

$$\begin{aligned}p\left(s\,|\,\mathbf{r}_A,\mathbf{r}_V\right)&\propto p\left(\mathbf{r}_A,\mathbf{r}_V\,|\,s\right)\\&=p\left(\mathbf{r}_A\,|\,s\right)p\left(\mathbf{r}_V\,|\,s\right)\end{aligned}\tag{6}$$

In going from the first to the second line, we have assumed that the cues are conditionally independent given the stimulus, just as in the behavioral theory. We substitute Eq. (3) and absorb all factors independent of *s* into the proportionality sign. The gives

$$p\left(s\,|\,\mathbf{r}_A,\mathbf{r}_V\right)\propto\left(\prod_{i=1}^{N}e^{-g_Af_i(s)}f_i\left(s\right)^{r_{Ai}}\right)\left(\prod_{i=1}^{N}e^{-g_Af_i(s)}f_i\left(s\right)^{r_{Ai}}\right).\tag{7}$$

This can be rewritten as:

$$p\left(s\,|\,\mathbf{r}_A,\mathbf{r}_V\right)\propto\exp\sum_{i=1}^{N}\left(-\left(g_A+g_V\right)f_i\left(s\right)+\left(r_{Ai}+r_{Vi}\right)\log f_i\left(s\right)\right).\tag{8}$$

This is the optimal posterior distribution encoded by $\mathbf{r}_A$ and $\mathbf{r}_V$ together. However, $\mathbf{r}_A$ and $\mathbf{r}_V$ are separate populations – what we want instead is a single multisensory population. To implement optimal cue integration neurally means to ask what we can do to $\mathbf{r}_A$ and $\mathbf{r}_V$ so that the resulting multisensory population encodes the optimal posterior distribution. In that way, we would not lose any information about *s*. Eq. (8) immediately suggests the answer: addition. Indeed, the population pattern $\mathbf{r}_A$ + $\mathbf{r}_V$ would still be independent Poisson (the sum of two Poisson processes is again Poisson), with mean activities $(g_A + g_V) f_i(s)$. Therefore, it encodes a posterior distribution that is given by:

$$p\left(s \mid \mathbf{r}_A + \mathbf{r}_V\right) \propto \exp \sum_{i=1}^{N} \left(-\left(g_A + g_V\right) f_i\left(s\right) + \left(r_{Ai} + r_{Vi}\right) \log f_i\left(s\right)\right). \tag{9}$$

This distribution is identical to the one in Eq. (8). We conclude that adding independent Poisson population patterns of activity implements a multiplication of the probability distributions over the stimulus that are encoded in those patterns. This is depicted in Figure 2 when the weights $\mathbf{W}_A$ and $\mathbf{W}_V$ are equal to 1. This is the simplest demonstration of optimal cue integration using probabilistic population codes, but there is much more to the story.

Q: Are these posterior distributions the same as behavioral response distributions (such as the ones underlying psychometric functions)? A: No. This is a tempting and common mistake . The posterior distribution, $p(s|\mathbf{r})$ 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 your favorite cost function). This value is the modelled 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, $p\left(\hat{s} \mid s_0\right)$ (sampled discretely). *There is no reason why* $p\left(\hat{s} \mid s_0\right)$ *should have the same shape or functional form as* $p(s|\mathbf{r})$. The reason that this mistake is so common is that in behavioral modeling, the internal representation is taken to be μ, not **r,** and from the outset, a Gaussian distribution $p(\mu|s)$ is assumed. Under uniform priors, this choice makes makes both $p(s|\mu)$ and $p\left(\hat{s} \mid s_0\right)$ Gaussian with the same variance. Operations performed on $p(s|\mu)$ are then directly mirrored in operations on $p\left(\hat{s} \mid s_0\right)$, just like we will see in this chapter. However, in the presence of non-uniform priors (Stocker & Simoncelli, 2006), or when the posterior is non-Gaussian (Kording et al., 2007), this is no longer true and identifying the posterior with the response distribution leads to wrong predictions.

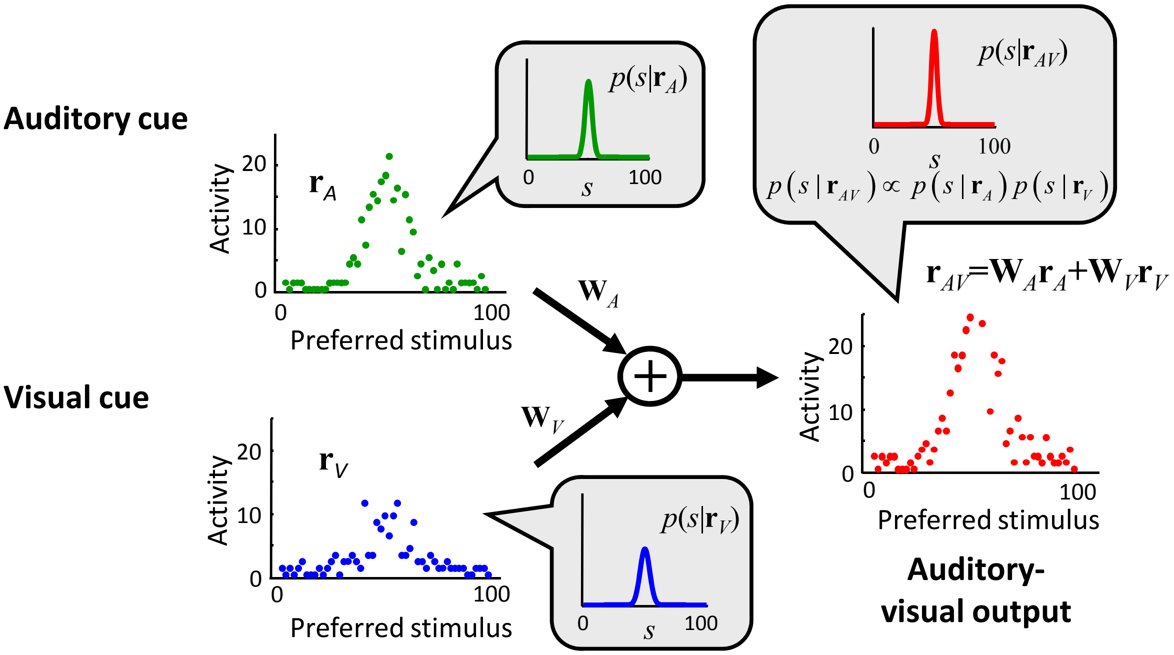

Figure 2: Optimal cue integration with probabilistic population codes. The cues elicit activity in input populations $\mathbf{r}_A$ and $\mathbf{r}_V$, indicated by green and blue dots. The dialogue boxes show the probability distributions over the stimulus encoded in each population on a single trial. A simple linear combination of the population patterns of activity, $\mathbf{r}_{AV} = \mathbf{W}_A\mathbf{r}_A + \mathbf{W}_V\mathbf{r}_V$, guarantees optimal cue integration, if neural variability is Poisson-like. Optimal cue integration means that the probability distribution over the stimulus encoded in the multisensory population is a product of the distributions encoded in the unisensory populations, *i.e.*, $p(s|\mathbf{r}_{AV})$ α $p(s|\mathbf{r}_A)p(s|\mathbf{r}_V)$. The synaptic weight matrices $\mathbf{W}_A$ and $\mathbf{W}_V$ depend on the tuning curves and covariance matrices of the input populations, but do not have to be adjusted over trials.

Poisson-like variability

First of all, the assumption that neurons follow Poisson statistics is often violated. 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, 2005; Tolhurst, Movshon, & Dean, 1982). Moreover, neurons are not independent (when conditioned on the stimulus) but exhibit correlations. 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:

$$p(\mathbf{r}\,|\,s) = \frac{\Phi(\mathbf{r})}{\eta(s)} e^{\mathbf{h}(s)\cdot\mathbf{r}}, \tag{10}$$

where Φ(**r**) is an arbitrary function of **r**, **h**(*s*) is a vector-valued function of *s* that we will specify later, and *η*(*s*) serves as a normalization (since this is a probability distribution over **r**). The exponent contains the inner product of **h**(*s*) with the population pattern of activity. To gain some intuition for Eq. (10), it helps to see what Φ, **h**, and *η* are for independent Poisson variability:

$$\Phi(\mathbf{r}) = \frac{1}{\prod_i r_i}, \quad \eta(s) = e^{g\sum_i f_i(s)}, \quad h_i(s) = \log f_i(s). \tag{11}$$

Thus, **h** and *η* both depend on the tuning curve, whereas Φ contains all factors that only depend on **r** and therefore do not affect the posterior distribution. Indeed, the posterior distribution encoded ina pattern **r** drawn from this family is

$$p(s \mid \mathbf{r}) \propto \frac{e^{\mathbf{h}(s)\cdot\mathbf{r}}}{\eta(s)}. \tag{12}$$

The important aspect of this family is that, just like for independent Poisson variability, optimal cue integration is achieved by addition of population patterns:

$$p(s \mid \mathbf{r}_A + \mathbf{r}_V) \propto p(\mathbf{r}_A \mid s)\, p(\mathbf{r}_V \mid s), \tag{13}$$

where we assumed that **h**(*s*) is the same for auditory and visual input. This can be verified by first calculating the distribution of the new random variable $\mathbf{r}_{AV} = \mathbf{r}_A + \mathbf{r}_V$. 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:

$$\begin{aligned} p(\mathbf{r}_{AV} \mid s) &= \int p(\mathbf{r}_V \mid s)\, p(\mathbf{r}_A = \mathbf{r}_{AV} - \mathbf{r}_V \mid s)\, d\mathbf{r}_V \\ &= \int \frac{\Phi_V(\mathbf{r}_V)}{\eta_V(s)} e^{\mathbf{h}(s)\cdot\mathbf{r}_V} \frac{\Phi_A(\mathbf{r}_{AV} - \mathbf{r}_V)}{\eta_A(s)} e^{\mathbf{h}(s)\cdot(\mathbf{r}_{AV} - \mathbf{r}_V)} d\mathbf{r}_V \\ &= \int \frac{\Phi_V(\mathbf{r}_V)}{\eta_V(s)} \frac{\Phi_A(\mathbf{r}_{AV} - \mathbf{r}_V)}{\eta_A(s)} e^{\mathbf{h}(s)\cdot\mathbf{r}_{AV}} d\mathbf{r}_V \\ &= \frac{\int \Phi_V(\mathbf{r}_V)\Phi_A(\mathbf{r}_{AV} - \mathbf{r}_V)\, d\mathbf{r}_V}{\eta_V(s)\eta_A(s)} e^{\mathbf{h}(s)\cdot\mathbf{r}_{AV}}. \end{aligned} \tag{14}$$

Note that in the transition from the second to the third line, it is essential that **h**(*s*) is the same in both populations. As a consequence of Eq. (14), the posterior encoded in a multisensory population pattern of activity is

$$p\left(s\,|\,\mathbf{r}_{AV}\right)\propto\frac{e^{\mathbf{h}(s)\cdot\mathbf{r}_{AV}}}{\eta_V\left(s\right)\eta_A\left(s\right)}. \tag{15}$$

When we substitute $\mathbf{r}_{AV}=\mathbf{r}_A+\mathbf{r}_V$ (i.e. giving the *random variable* $\mathbf{r}_{AV}$ the *value* $\mathbf{r}_A+\mathbf{r}_V$) this has the same dependence on *s* as the right-hand side of Eq. (13), thereby proving Eq. (13).

We saw that for independent Poisson variability, **h**(*s*) is equal to the log of the tuning curve. However, as Poisson-like variability is much more general, what is the new relationship between **h**(*s*) and the tuning curve. By using the definition, Eq. (10), it is possible to show that the derivative of **h** satisfies:

$$\mathbf{h}'\left(s\right)=\mathbf{\Sigma}^{-1}\left(s\right)\mathbf{f}'\left(s\right), \tag{16}$$

where $\mathbf{\Sigma}^{-1}\left(s\right)$ is the inverse of the covariance matrix of the population and **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

$$\mathbf{\Sigma}\left(s\right)=\begin{pmatrix} f_1\left(s\right) & 0 & \cdots & 0 \\ 0 & f_2\left(s\right) & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & f_N\left(s\right) \end{pmatrix}. \tag{17}$$

Therefore, Eq. (16) implies that $h_i'\left(s\right)=\frac{f_i'\left(s\right)}{f_i\left(s\right)}$,which is, up to a constant, equivalent to the second part of Eq. (11). This confirms that independent Poisson variability is indeed a special case of Eq. (16).

Gain independence

In the formulation of the Poisson-like family, Eq. (10), we did not include the gain *g*. Still, in principle, each of the factors in Eq. (10) could depend on the gain. For example, in Eq. (11), the expression for $\eta\left(s\right)$ contains the gain. This does not pose a problem if the gain is known, but in the inference problem the brain is trying to solve, which is obtaining a probability distribution over *s* from **r**, gain is not known. For example, if **r** is a visual population, then *g* could be determined by the contrast of the stimulus. 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:

$$p(s\,|\,\mathbf{r}) \propto p(\mathbf{r}\,|\,s) = \int p(\mathbf{r}\,|\,s,g)\,p(g)\,dg\,, \tag{18}$$

where *p*(*g*) is some prior distribution over *g*. In writing this integral, we have assumed that *g* does not depend on *s*. Eq. (18) severely complicates our framework for optimal cue integration, as can be checked by trying to repeat the above steps with the integral of Eq. (18) instead of *p*(**r**|*s*). Fortunately, there is a situation in which these complications can be avoided, namely if the *s*- and *g*-dependences of *p*(**r**|*s,g*) can be separated, i.e. if *p*(**r**|*s,g*) can be written as the product of a factor that only depends on *s* and **r** and one that only depends on *g* and **r**. Looking back at Eq. (10), this means that it would take the following form:

$$p(\mathbf{r}\,|\,s,g) = \frac{\Phi(\mathbf{r},g)}{\eta(s)}e^{\mathbf{h}(s)\cdot\mathbf{r}}. \tag{19}$$

How does this help? If we substitute Eq. (19) into the integral of Eq. (18), then we find:

$$\begin{aligned} p(s\,|\,\mathbf{r}) &\propto \int \frac{\Phi(\mathbf{r},g)}{\eta(s)}e^{\mathbf{h}(s)\cdot\mathbf{r}}\,p(g)\,dg \\ &= \frac{e^{\mathbf{h}(s)\cdot\mathbf{r}}}{\eta(s)}\int \Phi(\mathbf{r},g)\,p(g)\,dg \\ &\propto \frac{e^{\mathbf{h}(s)\cdot\mathbf{r}}}{\eta(s)}, \end{aligned} \tag{20}$$

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. Then the posterior is exactly the same as in Eq. (12), and everything goes through as before.

Starting from Eq. (19), here is another constraint that is imposed by the definition of the gain. To obtain it, we first compute the derivative of *η*(*s*) with respect to *s*, keeping in mind that *η*(*s*) is a normalization factor:

$$\begin{aligned} \frac{d}{ds}\eta(s) &= \int \left(\mathbf{h}'(s)\cdot\mathbf{r}\right)\Phi(\mathbf{r},g)\,e^{\mathbf{h}(s)\cdot\mathbf{r}}\,d\mathbf{r} \\ &= \mathbf{h}'(s)\cdot\int \mathbf{r}\;\Phi(\mathbf{r},g)\,e^{\mathbf{h}(s)\cdot\mathbf{r}}\,d\mathbf{r} \\ &= \mathbf{h}'(s)\cdot\langle\mathbf{r}\rangle = \mathbf{h}'(s)\cdot g\mathbf{f}(s). \end{aligned} \tag{21}$$

Now, differentiating both sides with respect to *g* gives $0=\mathbf{h}'(s)\cdot\mathbf{f}(s)$. Substituting this back into Eq. (21) implies that $\frac{d\eta}{ds}=0$. Surprisingly, this condition is not very hard to meet. For example, in the independent Poisson case, Eq. (11), it seems as if *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 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 $\Phi(\mathbf{r},g)$.

To allow Eq. (20) to be true, it is also important that **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 $\mathbf{\Sigma}^{-1}(s,g)\mathbf{f}'(s,g)$ (from Eq. (16)) cannot. Since $\mathbf{f}(s,g)=g\mathbf{f}(s)$, this means that the covariance matrix must be of the form $\mathbf{\Sigma}(s,g)=g\mathbf{\Sigma}(s)$. On the diagonal, this means that the variance scales with the gain, in other words, that the Fano factor is constant but not necessarily equal to 1 (as in a true Poisson process). Off-diagonal, it means that the entries $\langle r_i r_j\rangle-g^2 f_i(s)f_j(s)$ should be proportional to *g* as well. These conditions seem to be roughly satisfied in cortical neurons (Gur & Snodderly, 2005; Tolhurst et al., 1982), but further study is needed. In conclusion, we find that neural variability must be of the form

$$p(\mathbf{r}\,|\,s)=\Phi(\mathbf{r},g)e^{\mathbf{h}(s)\cdot\mathbf{r}}, \tag{22}$$

and the posterior distribution over *s* is

$$p(s\,|\,\mathbf{r})\propto e^{\mathbf{h}(s)\cdot\mathbf{r}}. \tag{23}$$

Instead of *g*, the same arguments can be made for other so-called *nuisance parameters* – these are parameters that influence neural variability but are not relevant to the stimulus. For example, when *s* is direction of motion, then speed, size, and contrast are nuisance parameters. As long as these variables only affect Φ, the framework for optimal cue integration stays the same.

Q: How can I check if my physiological data are Poisson-like? A: When population activity has been recorded over many trials of the same stimulus in a fine discrimination task, one can check whether the data are consistent with Poisson-like variability. First, check if the Fano factor is approximately constant. Then, compute the locally optimal linear decoder (Series, Latham, & Pouget, 2004). If variability is Poisson-like, this decoder should extract all available information. This can be checked using other decoders, such as a support vector machine.

Moreover, this decoder turns out to be equal to **h**'(*s*) (Beck et al., 2008) and therefore it should be independent of gain.

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. In behavioral modeling discussed in this book, it is assumed that the posterior distribution *p*(*s*|**r**) is Gaussian. Therefore, it is exponential with a quadratic function of *s* in the exponent:

$$p\left(s\,|\,\mathbf{r}\right) \propto e^{-\frac{1}{2}s^2 a(\mathbf{r}) + s b(\mathbf{r})}, \tag{24}$$

where *a*(**r**) and *b*(**r**) are functions of **r**. Comparing with Eq. (23), we see that these functions must be of the form $a\left(\mathbf{r}\right) = \mathbf{a}\cdot\mathbf{r}$ and $b\left(\mathbf{r}\right) = \mathbf{b}\cdot\mathbf{r}$, where now **a** and **b** are constant vectors. From Eq. (24), we can find the mean μ and variance σ^2 of the Gaussian, since the exponent of a Gaussian is of the form $-\frac{\left(s-\mu\right)^2}{2\sigma^2} = -\frac{s^2}{2\sigma^2} + \frac{s\mu}{\sigma^2} + \text{constant}$. They are given by

$$\frac{1}{\sigma^2} = a\left(\mathbf{r}\right) = \mathbf{a}\cdot\mathbf{r} \tag{25}$$

and

$$\frac{\mu}{\sigma^2} = b\left(\mathbf{r}\right) = \mathbf{b}\cdot\mathbf{r}\,. \tag{26}$$

Since $\mathbf{r}_{AV} = \mathbf{r}_A + \mathbf{r}_V$ for optimal cue integration, applying the inner product with **a** gives (from Eq. (25)):

$$\frac{1}{\sigma_{AV}^2} = \frac{1}{\sigma_A^2} + \frac{1}{\sigma_V^2}, \tag{27}$$

which is the *single-trial* version of our well-known equation for optimal combination of variances. For the mean, Eq. (26) gives

$$\frac{\mu_{AV}}{\sigma_{AV}^2} = \frac{\mu_A}{\sigma_A^2} + \frac{\mu_V}{\sigma_V^2} \tag{28}$$

(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*(*s*|**r**) into a single estimate of the stimulus, $\hat{s}(\mathbf{r})$, on each trial (this step is called decoding, estimation, or read-out of *s*). The optimal (so-called "efficient") estimator is the maximum-likelihood estimator, which chooses the *s* that makes *p*(**r**|*s*) maximal. Since we have assumed that the prior is uniform, this is also the estimator that maximizes *p*(*s*|**r**). For a Gaussian *p*(*s*|**r**), 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 (across many trials!) of an optimal estimator is given by the *Fisher information*:

$$\frac{1}{\sigma^2_{\text{estimator}}} = I(s) \equiv -\left\langle \frac{\partial^2}{\partial s^2} \log p(\mathbf{r} \mid s) \right\rangle, \tag{29}$$

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

$$I(s) = -\mathbf{h}''(s) \cdot g\mathbf{f}(s) = g\mathbf{h}'(s) \cdot \mathbf{\Sigma}(s)\mathbf{h}'(s) = g\mathbf{f}'(s) \cdot \mathbf{\Sigma}^{-1}(s)\mathbf{f}'(s). \tag{30}$$

This offers a particularly easy way to check optimality of $\mathbf{r}_{AV} = \mathbf{r}_A + \mathbf{r}_V$. Taking the average on both sides, we find $g_{AV} = g_A + g_V$, and since *I*(*s*) is proportional to *g* according to Eq. (30), it follows from Eq. (29) the estimate's inverse variances sum:

$$\frac{1}{\sigma^2_{\text{estimator},AV}} = \frac{1}{\sigma^2_{\text{estimator},A}} + \frac{1}{\sigma^2_{\text{estimator},V}}. \tag{31}$$

This is the relationship found in behavior. Similarly, for the mean estimate, we have from Eq. (26):

$$\langle \hat{s} \rangle = \sigma^2 \mathbf{a} \cdot \langle \mathbf{r} \rangle = \sigma^2_{\text{estimator}} \mathbf{a} \cdot g\mathbf{f}(s), \tag{32}$$

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

$$\frac{\langle \hat{s} \rangle_{AV}}{\sigma^2_{\text{estimator},AV}} = \frac{\langle \hat{s} \rangle_A}{\sigma^2_{\text{estimator},AV}} + \frac{\langle \hat{s} \rangle_V}{\sigma^2_{\text{estimator},AV}}, \tag{33}$$

which is the behavioral result for the mean multisensory estimate.

Q: What is optimal about cue integration? Would an optimal strategy not lead to separate, veridical percepts of the auditory and the visual stimulus? A: In cue integration experiments, small conflicts between the cues are introduced by the researcher. However, small conflicts are always present even in normal perception, when no artificial conflict is introduced. 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. (28)). Thus, even when the true auditory and visual stimuli are completely in agreement to the experimenter, the brain still has to solve the cue integration problem. Of course, it is important that in experimental settings, the artificial conflicts are kept small enough to be mistaken for naturally occurring ones. If they are too large, subjects will start noticing that they might have different sources, and move towards perceiving separate, visual percepts of the stimuli. Such perception can be modeled using Bayesian causal inference models (Kording et al., 2007) (see also this book, chapter [REF]).

Extensions and predictions

While the above framework is already quite general, it still assumes that $\mathbf{h}(s)$ is the same for both the auditory and the visual population. In general, this might not be the case, since tuning curves and covariance matrices can easily 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, i.e. $\mathbf{h}_A(s) = \mathbf{W}_A\mathbf{H}(s)$ and $\mathbf{h}_V(s) = \mathbf{W}_V\mathbf{H}(s)$, where $\mathbf{W}_A$ and $\mathbf{W}_V$ are stimulus-independent matrices and $\mathbf{H}(s)$ is the common basis. Then, it can be shown that the linear combination $\mathbf{r}_{AV} = \mathbf{W}_A^{\mathrm{T}}\mathbf{r}_A + \mathbf{W}_V^{\mathrm{T}}\mathbf{r}_V$ (where the superscript “T” denotes a transpose) implements optimal cue integration (for details, see the Supplement of (Ma et al., 2006)). This is shown in Figure 2. Importantly, 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 can now be learnt once and for all.

We have so far used very abstract, firing-rate neurons 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. This was done in (Ma et al., 2006); a network of 1008 conductance-based integrate-and-fire neurons was tuned so that it would act as an optimal cue integrator. It is not known how to capture the network behavior of such neurons in simple equations, and therefore a direct mapping between this network and the firing-rate neurons cannot be made. However, this provides an example of neural modeling that is driven top-down, where the computational model (at a 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.

The theory predicts that in a multisensory area, the activity due to both an auditory and a visual cue is approximately equal to the sum of the activity due to only the auditory cue and that due to only the visual cue. Recent physiological studies have begun to test this (see Angelaki's chapter in this book [REF]). Earlier ideas that claim superadditivity of multisensory responses (Stein & Meredith, 1993) are now largely discredited (see (Ma & Pouget, 2008) for a review).

In greater generality, the PPC framework also predicts that any neural population that encodes a stimulus variable, simultaneously also encodes uncertainty about this variable. No separate population is needed for this. Moreover, the form of neural variability is of great importance for inference. These concepts can be applied to other computations where combing pieces of uncertain information is key, such as decision-making (Beck et al., 2008), visual search (Ma, Navalpakkam, Beck, & Pouget, 2008), and causal inference.

Beck, J. M., Ma, W. J., Kiani, R., Hanks, T. D., Churchland, A. K., Roitman, J. D., et al. (2008). Bayesian decision-making with probabilistic population codes. *Neuron, 60*(6), 1142-1145.

Foldiak, 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. (2005). High Response Reliability of Neurons in Primary Visual Cortex (V1) of Alert, Trained Monkeys. *Cereb Cortex*.

Kording, 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.

Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nat Neurosci, 9*(11), 1432-1438.

Ma, W. J., Beck, J. M., & Pouget, A. (2008). Spiking networks for Bayesian inference and choice. *Curr Opin Neurobiol, 18*, 217-222.

Ma, W. J., Navalpakkam, V., Beck, J. M., & Pouget, A. (2008). Bayesian theory of visual search, *Society for Neuroscience*. Washington, DC.

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

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

Series, P., Latham, P., & Pouget, A. (2004). Tuning curve sharpening for orientation selectivity: coding efficiency and the impact of correlations. *Nature Neuroscience, 10*(7), 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. *Nat Neurosci, 9*(4), 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.