Vision Research 50 (2010) 2308–2319

Contents lists available at ScienceDirect

Vision Research

journal homepage: www.elsevier.com/locate/visres

Review

# Signal detection theory, uncertainty, and Poisson-like population codes

Wei Ji Ma

*Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA*

ARTICLE INFO

*Article history:*
Received 19 October 2009
Received in revised form 20 August 2010

*Keywords:*
Signal detection theory
Population coding
Bayesian inference
Single neurons

ABSTRACT

The juxtaposition of established signal detection theory models of perception and more recent claims about the encoding of uncertainty in perception is a rich source of confusion. Are the latter simply a rehash of the former? Here, we make an attempt to distinguish precisely between optimal and probabilistic computation. In optimal computation, the observer minimizes the expected cost under a posterior probability distribution. In probabilistic computation, the observer uses higher moments of the likelihood function of the stimulus on a trial-by-trial basis. Computation can be optimal without being probabilistic, and vice versa. Most signal detection theory models describe *optimal* computation. Behavioral data only provide evidence for a neural representation of uncertainty if they are best described by a model of *probabilistic* computation. We argue that single-neuron activity sometimes suffices for optimal computation, but never for probabilistic computation. A population code is needed instead. Not every population code is equally suitable, because nuisance parameters have to be marginalized out. This problem is solved by Poisson-like, but not by Gaussian variability. Finally, we build a dictionary between signal detection theory quantities and Poisson-like population quantities.



## 1. Introduction

> *"The [authors state] that the encoding of an animal's uncertainty is the key point distinguishing models encoding probability distributions from those encoding estimates. Wrong. Signal detection theory has been used for decades to correlate an animal's performance with the degree to which neurons discriminate between alternative choices. [...] The idea is that the animal can no better discriminate between alternative choices than can an ideal observer observing the neurons representing those choices. The neural responses are often well correlated by this measure with the animal's performance, including when the animal gets it less than 100% right, which represents uncertainty. So the encoding of uncertainty, either explicitly or implicitly in the less-than-perfect discriminability of the activities of populations representing alternatives, provides no evidence for [computation with probability distributions]."* – *Anonymous reviewer, 2009*

In the past 50 years, signal detection theory has been used extensively to model human and animal perception (e.g., Burgess, 1985; Green & Swets, 1966; Macmillan & Creelman, 2005; Peterson, Birdsall, & Fox, 1954). Typically applied to binary (two-alternative) decisions, the central idea is that an observer computes the log posterior ratio of two alternatives, and compares its value to a criterion to make a decision. Observer errors arise from the variability (also called noise) in the observations used to compute the log posterior ratio. As the quoted reviewer points out, it is now standard practice to use signal detection theory to quantify the discrimination performance of a neuron (Bradley, Skottun, Ohzawa, Sclar, & Freeman, 1987; Britten, Shadlen, Newsome, & Movshon, 1992; Newsome, Britten, & Movshon, 1989; Parker & Newsome, 1998).

A more recent development is the application of Bayesian modeling to cue combination (Clark & Yuille, 1990; Knill & Richards, 1996). While signal detection theory is also Bayesian, the novelty here is that the optimal observer weights observations (cues) by their respective reliabilities when combining them. Humans (Alais & Burr, 2004; Battaglia, Jacobs, & Aslin, 2003; Ernst & Banks, 2002; Jacobs, 1999; Knill, 1998a, 1998b; Knill & Richards, 1996; Knill & Saunders, 2003; Landy, Maloney, Johnston, & Young, 1995; Ma, Zhou, Ross, Foxe, & Parra, 2009; Reuschel, Drewing, Henriques, Roesler, & Fiehler, 2010; van Beers, Sittig, & Denier van der Gon, 1996; van Beers, Sittig, & Gon, 1999) and monkeys (Gu, Angelaki, & DeAngelis, 2008) follow the predictions of the optimal cue combination model rather closely. Human near-optimality has also been found in other perceptual tasks, including visual speed discrimination (Stocker & Simoncelli, 2006), the cueing task (Shimozaki, Eckstein, & Abbey, 2003), visuomotor learning (Kording & Wolpert, 2004), causal inference (Kording et al., 2007), visual-memory integration (Brouwer & Knill, 2007), oddity detection (Hospedales & Vijayakumar, 2009), tactile trajectory perception (Goldreich, 2007), combining sensory information with reward (Whiteley & Sahani, 2008), and visual search (Ma, Navalpakkam, Beck, & Pouget, 2008b; Palmer, Verghese, & Pavel, 2000; Vincent, Baddeley, Troscianko, & Gilchrist, 2009).

*E-mail address:* wjma@bcm.edu

0042-6989/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.visres.2010.08.035

Weighting observations by reliability requires knowledge of reliability, which can be regarded as a property of a probability distribution over the stimulus. Therefore, the above psychophysical findings are often regarded as support for theoretical (Anastasio, Patton, & Belkacem-Boussaid, 2000; Anderson, 1994; Barlow, 1969; Deneve, 2008; Fiser, Berkes, Orban, & Lengyel, 2010; Foldiak, 1993; Hoyer & Hyvarinen, 2000; Jazayeri & Movshon, 2006; Ma, Beck, Latham, & Pouget, 2006; Pouget, Dayan, & Zemel, 2003; Sanger, 1996; Zemel, Dayan, & Pouget, 1998) and neurophysiological (Gu et al., 2008; Kepecs, Uchida, Zariwala, & Mainen, 2008; Kiani & Shadlen, 2009; Morgan, DeAngelis, & Angelaki, 2008; Yang & Shadlen, 2007) investigations of the neural encoding of probability or uncertainty in perceptual decision-making.

Here, we offer a cautionary tale. In many instances, Bayesian near-optimality does not provide evidence for a neural representation of probability, reliability, or uncertainty. Only when a psychophysics experiments satisfies a number of requirements does such evidence exist. While the majority of signal detection theory studies do not provide evidence for computation with probability distributions, the quoted reviewer was incorrect in stating that observers in those models still possess an implicit encoding of uncertainty by virtue of not performing perfectly. Conversely, it is possible that an observer's behavior is best described by a *suboptimal* model and still provides evidence for a neural representation of uncertainty.

## 2. Neural representation of uncertainty: psychophysical evidence

We first review the probabilistic approach to perceptual modeling. For simplicity, we assume that only one feature dimension is relevant, e.g., orientation or motion direction. The variables specifying this dimension across all stimuli are denoted by $s$. This could be a vector or a scalar, depending on how many stimuli there are; we will mostly use the singular "stimulus" for convenience. Sensory input is generated from $s$, but also depends on other variables that determine stimulus reliability, such as luminance contrast or size in an orientation judgment task. We denote these nuisance parameters by $\theta$. In natural environments, they typically change across time. Sensory input, also called the proximal stimulus or the observation, is denoted $I$; in a visual task, it could be the retinal image. $I$ is generated from $s$ and $\theta$ through a stochastic process, reflecting various sources of noise. Therefore, it varies from trial to trial even when $s$ and $\theta$ are held fixed. This process can be formalized as a conditional probability distribution, $p(I|s, \theta)$.

### 2.1. The likelihood function of the stimulus

All information that an observer receives about the stimulus and the nuisance parameters on a single trial is contained in the likelihood function, defined as

$$L_I(s, \theta) = p(I|s, \theta).$$

The most important thing to keep in mind about this likelihood function is that its arguments, $s$ and $\theta$, are not physical variables controlled by the experimenter, but variables that describe hypotheses in the mind of the observer. The probability assigned to them is also called *degree of belief* (Ramsey, 1926). To obtain the likelihood function of the stimulus alone, we have to *marginalize out* the nuisance parameters, since they are unknown (Kersten, Mamassian, & Yuille, 2004; Peterson et al., 1954)[1]:

$$L_I(s) = p(I|s) = \int p(I|s, \theta)p(\theta)d\theta, \quad (1)$$

where $p(\theta)$ is the prior over nuisance parameters (which could in principle depend on $s$, but usually is assumed not to). The likelihood is not a probability distribution, since it is not normalized. We denote its normalized version by $L_I^*(s)$. Because $I$ varies from trial to trial, the (normalized) likelihood does as well (Fig. 1a). It can be parametrized through its moments. For example, if it is a one-dimensional Gaussian, then it is uniquely specified by its mean $x$ and variance $\sigma^2$. We can then write

$$L_I^*(s) = L(s; x(I), \sigma^2(I)) = \frac{1}{\sqrt{2\pi\sigma^2(I)}} e^{-\frac{(s-x(I))^2}{2\sigma^2(I)}}, \quad (2)$$

where the notation indicates that the "summary statistics" $x$ and $\sigma^2$ are functions of $I$. When there are multiple stimuli, $s$ and $x$ are vectors and $\sigma^2$ is replaced by a covariance matrix. The mean $x$ is a point estimate of the stimulus, sometimes also called the internal representation of the stimulus, the observation, or the measurement. Since for a Gaussian, the mean equals the mode, $x(I)$ is also the maximum-likelihood estimate of the stimulus based on $I$. The variance $\sigma^2(I)$ is a measure of the uncertainty about the stimulus (which we will also call "stimulus uncertainty", not to be confused with overlapping class distributions in classification). Like $x$, it varies from trial to trial because $I$ does. $I$ in turn varies because the underlying true stimulus and nuisance parameters, $s_{\text{true}}$ and $\theta_{\text{true}}$, vary, but also because for given $s_{\text{true}}$ and $\theta_{\text{true}}$, $I$ is drawn from a probability distribution. A common approximation to the effect of this stochasticity is that it only affects the mean, $x$, but not the variance $\sigma^2(I)$ (see Fig. 1b). In that case, variance is simply a function of $s_{\text{true}}$ and $\theta_{\text{true}}$, which we denote $\sigma^2(s_{\text{true}}, \theta_{\text{true}})$. We use the label "true" to distinguish experimentally controlled variables from the unlabeled variables $s$ and $\theta$, which indicate hypotheses considered by the observer, as in Eq. (1).

### 2.2. Estimation

Suppose an observer has to estimate a world-state variable $C$; this could be a physical feature of the stimulus or a more abstract variable, such as target presence in visual search. $C$ can be binary, multi-valued, or continuous. For given input $I$, the observer's estimate of $C$ is denoted $\widehat{C}$. The estimate follows a probability distribution, $p(\widehat{C}|I)$. If the estimator is deterministic, the estimate can be written as a function of $I$, $\widehat{C} = F(I)$, and the distribution $p(\widehat{C}|I)$ is given by a delta function. If the Gaussian form of the stimulus likelihood, Eq. (2), is used, then the estimate is a function of $x$ and $\sigma^2$, $\widehat{C} = F(x, \sigma^2)$.

Since $I$ is the noisy sensory input on a single trial, it is not directly accessible to the experimenter. Therefore, of interest is commonly the probability of $\widehat{C}$ predicted by the model over many trials, in a particular experimental condition given by $s$ and $\theta$. This is obtained by marginalizing over $I$:

$$p(\widehat{C}|s, \theta) = \int p(\widehat{C}|I)p(I|s, \theta)dI. \quad (3)$$

All measures of model observer performance can be computed from this estimate distribution.

### 2.3. Optimal inference

An optimal observer estimates $C$ based on the posterior probability distribution over $C$ given sensory input $I$ on a single trial. This distribution, $p(C|I)$, quantifies the observer's degree of belief in all possible values of $C$. On each trial, the posterior is computed by combining the likelihood function, $L_I(s)$, with other, task-specific probability distributions. This is done according to the rules of

[1] Here is an everyday example of marginalization: suppose you know for each province $P$ in a country the proportion of farmers, $p(F|P)$, and the proportion of the country's population living in the province, $p(P)$. Then the proportion of farmers in the entire country, $p(F)$, is a weighted average of the province-specific proportions, with the weights given by the proportions of the country's population in each province: $p(F) = \Sigma_P\, p(F|P)p(P)$.

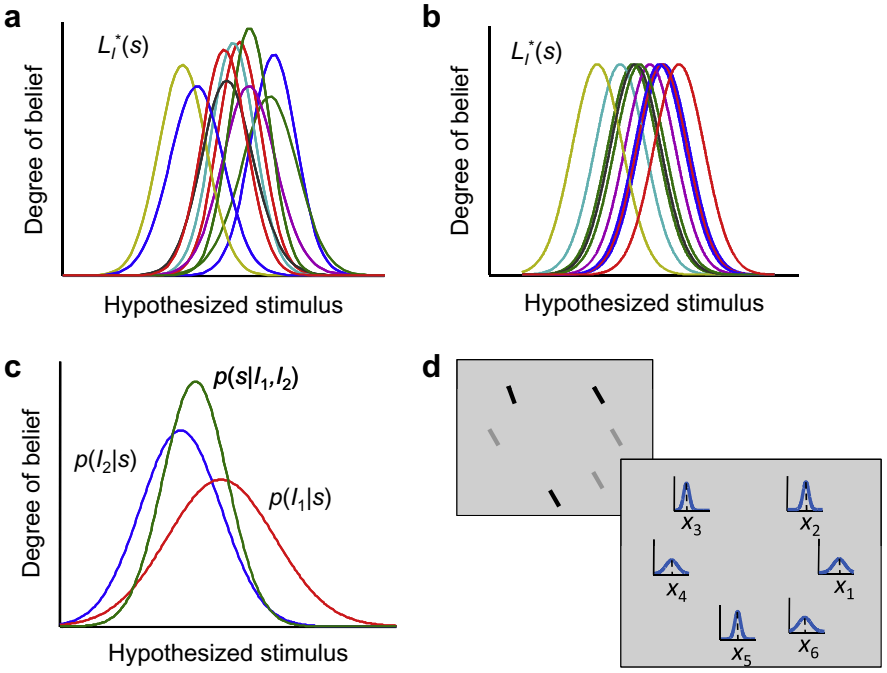


**Fig. 1.** The likelihood function of the stimulus in probabilistic models of perception. (a) Normalized likelihood functions of the stimulus obtained by repeating the same true stimulus value many times. (b) Common approximation to the likelihood function: only the mean varies from trial to trial, but the variance is constant. (c) Cue combination: single-cue likelihoods, $p(x_1|s)$ and $p(x_2|s)$, and normalized optimal combined likelihood, $L_I^*(s) \propto p(I_1|s)p(I_2|s)$. The prior was chosen flat. (d) Visual search: example display (top) and local normalized stimulus likelihood functions (bottom). The likelihood function contains information both about the most likely value of the stimulus, labeled $x_i$, and about uncertainty about the stimulus (width). In this example, some bars have lower contrast, which leads to wider likelihoods.

probability calculus and the structure of the generative model, i.e., the set of probability distributions through which observations are generated (see the examples below). A typical generative model is shown in Fig. 2a. The optimal way to read out the posterior depends on the cost function; one way is the maximum-a-posteriori (MAP) estimate:

$$\hat{C} = \underset{C}{\operatorname{argmax}}\, p(C|I),$$

where argmax is the operation of finding the value (here of $C$) for which the expression following it is maximized.

**Example 1** (*Classification*). In classification, a stimulus $s$ is drawn from one of several classes, characterized by probability distributions $p(s|C)$, where $C$ is the class label (Fig. 2a). The observer has learned these distributions and has to decide on each trial which class the stimulus was drawn from. Then the posterior probability distribution over $C$ is calculated as

$$p(C|I) = \frac{p(I|C)p(C)}{p(I)} = \frac{p(I|C)p(C)}{\sum_C p(I|C)p(C)} = \frac{p(C)\int p(I|s)p(s|C)ds}{\sum_C p(C)\int p(I|s)p(s|C)ds}$$
$$= \frac{p(C)\int L_I(s)p(s|C)ds}{\sum_C p(C)\int L_I(s)p(s|C)ds} = \frac{p(C)\int L_I^*(s)p(s|C)ds}{\sum_C p(C)\int L_I^*(s)p(s|C)ds}. \quad (4)$$

Most signal detection theory models describe binary classification. If we choose the values of $C$ to be ±1, then the MAP estimate is the sign of the log posterior ratio (Green & Swets, 1966),

$$\hat{C} = \operatorname{sgn}(d), \quad d = \log\frac{p(C=1|I)}{p(C=-1|I)}. \quad (5)$$

Detection and discrimination are special cases of classification; in some classification tasks, such as same-different judgment, more than one stimulus might be present at the same time.

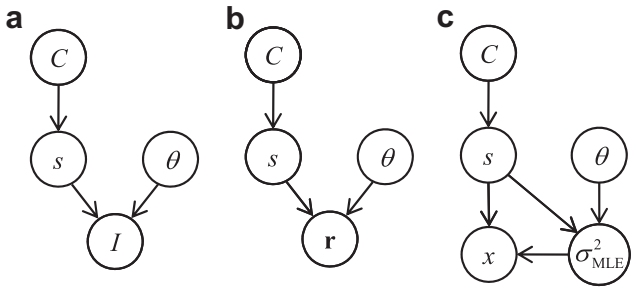


**Fig. 2.** Generative models of a classification task. Stimuli, collectively denoted $s$, and nuisance parameters $\theta$ generate an observation (internal representation) through a noisy process. The observer has to infer $C$ from the observation. Each arrow corresponds to a conditional probability distribution. The diagrams differ in their description of the observations: (a) Abstract sensory input $I$. (b) Neural population activity r. (c) Simplified model using summary statistics: the maximum-likelihood estimate $x$ and its variance.

**Example 2** (*Cue combination, estimation task*). In cue combination, an observer receives two conditionally independent sensory inputs, $I_1$ and $I_2$, generated by the same stimulus $s$. The world-state variable $C$ is $s$ itself. The posterior is (Clark & Yuille, 1990)

$$p(s|I_1,I_2) = \frac{p(I_1,I_2|s)p(s)}{p(I_1,I_2)} = \frac{p(I_1|s)p(I_2|s)p(s)}{\int p(I_1|s)p(I_2|s)p(s)ds}$$
$$= \frac{L_{I_1}^*(s)L_{I_2}^*(s)p(s)}{\int L_{I_1}^*(s)L_{I_2}^*(s)p(s)ds}.$$

(Fig. 1c). When likelihoods are Gaussian (Eq. (2)) and the prior is flat, the MAP estimate is $\hat{s} = \frac{x_1\sigma_1^{-2}+x_2\sigma_2^{-2}}{\sigma_1^{-2}+\sigma_2^{-2}}$. This expression has been tested in many psychophysical studies.

**Example 3** (*Homogeneous visual search, yes/no task*). In visual search with homogeneous distractors with a fixed value, an observer detects whether a target stimulus with value $s_T$ is present ($C = 1$) or absent ($C = -1$) among $N$ stimuli $s_1, \ldots, s_N$, with corresponding sensory inputs $I_1, \ldots, I_N$. Distracters have a value $s_D$. We make the assumption that nuisance parameters are uncorrelated across locations, which is unrealistic in natural scenes. At each location, a likelihood $L_{I,i}(s_i) = p(I_i|s_i)$ is computed (Fig. 1d). When the prior probability of target presence equals ½, the log posterior ratio is ((Peterson et al., 1954), Eq. 162):

$$\frac{p(C=1|I_1,\ldots,I_N)}{p(C=-1|I_1,\ldots,I_N)} = \frac{1}{N}\sum_{i=1}^{N}\frac{p(I_i|s_i=s_T)}{p(I_i|s_i=s_D)} = \frac{1}{N}\sum_{i=1}^{N}\frac{L^*_{I_i}(s_T)}{L^*_{I_i}(s_D)} = \frac{1}{N}\sum_{i=1}^{N}\exp\left(\frac{s_T-s_D}{\sigma_i^2}\left(x_i-\frac{s_T+s_D}{2}\right)\right), \quad (6)$$

where the last equality holds when likelihoods are Gaussian (Eq. (2)). Here too, $x_i$ is weighted by $1/\sigma_i^2$. When distractors are heterogeneous, integrals like the one in Eq. (4) appear.

In all examples, we see that the posterior over $C$ is a functional of $L^*_I(s)$:

$$p(C|I) = \Phi_C[L^*_I(s)] \quad (7)$$

(a functional is a function of a function and its argument is put in square brackets). The form of $\Phi_C$ depends on the task, as seen in the examples. Eq. (7) expresses that the likelihoods over the stimuli form the building blocks from which the posterior is constructed. When the normalized likelihood is parametrized by its moments, for example $x$ and $\sigma^2$ when it is a one-dimensional Gaussian, then the posterior over class, $p(C|I)$, will be a function of those moments, $p(C|I) = \Phi(C;x(I),\sigma^2(I))$. In the Gaussian case, since $L^*_I(s)$ depends on $x$ and $\sigma^2$ only through the form $(s-x)^2/\sigma^2$ (Eq. (2)), those variables will appear in the posterior and MAP estimate only in the combinations $x/\sigma^2$ and $x^2/\sigma^2$. This explains why weighting is always by $1/\sigma^2$, as in the examples.

### 2.4. Model concepts

We can now formulate the following model concepts (Fig. 3):

- A *probabilistic model* is a model in which the trial-to-trial observations are stochastic when the presented stimulus and nuisance parameters are fixed, i.e., $I$ is a random variable for given $s$ and $\theta$. Due to variability in $I$, the estimate $\widehat{C}$ will also be a random variable, with a distribution given by Eq. (3). All signal detection theory and Bayesian models are probabilistic, but limited-capacity or high-threshold models are not (Green & Swets, 1966).
- A *Bayesian or optimal (observer) model* (or model of optimal computation) is a model in which the observer's estimate of $C$ minimizes the expected cost under the posterior distribution, $p(C|I)$.
- A *model of probabilistic computation* is a probabilistic model in which the observer, on every trial, utilizes not only on a point estimate $x$ of the stimulus to estimate $C$ or decision confidence, but also at least one other parameter (higher moment) of the normalized stimulus likelihood $L^*_I(s)$ that is independent of that point estimate. If the normalized stimulus likelihood is Gaussian, this means that the observer utilizes trial-to-trial knowledge of (co)variance. A model in which the observer's estimates are functions of $x$ only is not a model of probabilistic computation. The higher moments of the likelihood are collectively called uncertainty about the stimulus. Therefore, probabilistic computation requires the neural representation of uncertainty. It does not necessarily require the neural representation of the exact values of the normalized likelihood, which is why we avoid the term "neural representation of probability".

These notions differ in the following ways. Many probabilistic models are neither optimal nor describe probabilistic computation, for example, a model with sensory noise in which the observer guesses randomly. It is possible for optimal computation to be non-probabilistic, and for probabilistic computation to be suboptimal. If an optimal observer performs MAP estimation but the MAP estimate only depends on $x$ (and no decision confidence is estimated), the model is non-probabilistic, even though the posterior which the MAP estimate is derived from always depends on the full likelihood function (Eq. (7)). This happens in numerous tasks, including discrimination between two stimuli $s_A$ and $s_B$ (optimally done by comparing $x$ to $(s_A + s_B)/2$), detection (similar to discrimination), and cue combination under equal likelihood widths (optimally done by an unweighted average of observations). In general, equal likelihood widths do not guarantee that the optimal estimator is non-probabilistic: in Eq. (6), even when $\sigma_i$ is independent of $i$, the MAP estimate still depends on $\sigma_i$. On the other hand, not all probabilistic computation is optimal. For example, in cue combination, the estimator $\hat{s} = \frac{x_1\sigma_1^{-1}+x_2\sigma_2^{-1}}{\sigma_1^{-1}+\sigma_2^{-1}}$ is probabilistic because it depends on $\sigma_i$, but it is not optimal (for a given cost function, there can only be one optimal estimator). To our knowledge, all models of probabilistic computation that have been found to describe human data best have also been models of optimal computation. However, this might be because of the tasks that have been studied so far and difficulty in finding appropriate suboptimal models.

An observer who uses a non-probabilistic estimator of $C$ still performs probabilistic computation if a higher moment of the stimulus likelihood is utilized on a trial-to-trial basis to estimate confidence or certainty about the estimate of $C$. Decision confidence is relevant when the observer has to report it, as in a signal detection theory rating experiment (Green & Swets, 1966; Macmillan & Creelman, 2005), or use it for behavioral output (Kepecs et al.,

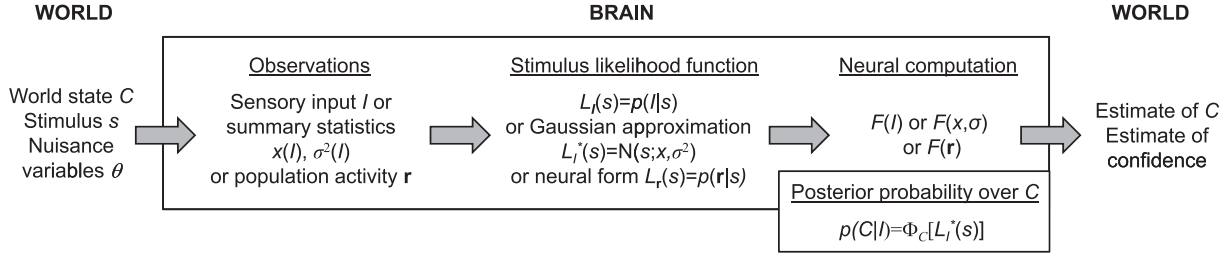


**Fig. 3.** Schematic of perceptual computation. The observer estimates a world-state variable $C$ from observations $I$ (abstract form) or $\mathbf{r}$ (neural form). All computation is based on stimulus likelihood functions $L_I(s)$ or $L_\mathbf{r}(s)$. The posterior over $C$ is a functional of the likelihood function. When computation is optimal, the estimate of $C$ is based on this posterior $C$ but may be independent of higher moments of the stimulus likelihood. When computation is probabilistic, those higher moments are used. $N(s; x, \sigma^2)$ denotes the Gaussian probability density function over $s$ with mean $x$ and variance $\sigma^2$.

2008; Kiani & Shadlen, 2009). In optimal inference, decision confidence can, for instance, be quantified using the variance or the entropy of the posterior distribution, $p(C|I)$ (for a continuous variable $C$) or as the absolute value of the log posterior ratio (for a binary variable $C$). Since the posterior distribution depends on uncertainty about the stimulus, decision confidence does as well. When the observer's decision process does not use the optimal posterior distribution, it might still be possible to define decision confidence. For example, when he uses a wrong generative model, there is still a (wrong) posterior; when a decision variable is compared to a criterion, the distance between the two is a measure of confidence. However, in suboptimal observer models, confidence does not necessarily depend on stimulus uncertainty. For confidence estimation to be regarded as probabilistic computation, it must depend on stimulus uncertainty, and stimulus uncertainty must be independent of the stimulus estimate and unknown in advance to the observer.

In some cases, it is possible to approximate an optimal estimator that depends on stimulus uncertainty by a suboptimal estimator that does not. Several applications of signal detection theory to perception have considered such non-probabilistic, suboptimal estimators (Pelli, 1985). A common example is the maximum-of-outputs (or max) model of visual search (Eckstein, 1998; Eckstein, Thomas, Palmer, & Shimozaki, 2000; Green & Swets, 1966; Nolte & Jaarsma, 1966; Palmer et al., 2000), which uses, in visual search with homogeneous distractors, the estimator $\widehat{C} = \mathrm{sgn}(\max_i x_i - c)$, where $c$ is an arbitrary criterion. This max estimator is a reasonable approximation of the optimal estimator when the variances $\sigma_i^2$ are equal and we are only interested in the receiver operating characteristic (ROC), rather than the specific value of the criterion $c$. (The ROC is swept out by varying the criterion along a continuum.) Although the max rule is not an optimal estimator, it has the advantage that the observer does not need to know $\sigma$. Similarly, the signed-max rule (Baldassi & Verghese, 2002), the sum rule (Baldassi & Burr, 2000; Graham, Kramer, & Yager, 1987; Green & Swets, 1966), and the maximum-of-differences rule (Palmer et al., 2000) are suboptimal and non-probabilistic.

### 2.5. Evidence for a neural representation of uncertainty

The distinctions made in the previous section allow us to outline criteria which a psychophysics experiment should satisfy in order to provide evidence for probabilistic computation, and therefore for a neural representation of uncertainty. These criteria are:

- Behavioral data are best described by a model in which the observer makes his decision using independent higher moments of the normalized likelihood function (stimulus uncertainty), for example $\sigma^2(I)$ in the Gaussian case. "Best described by" can be made precise using Bayesian model comparison (MacKay, 2003). The model does not need to be optimal.
- In the experiment, these higher moments are not known to the observer in advance. In practice, this means that (a) feedback during testing is absent or limited; (b) the observer is not trained extensively with feedback on the same values of nuisance parameters as are used in testing (Whiteley & Sahani, 2008); (c) preferably, the values of the nuisance parameters vary randomly from trial to trial instead of per block or not at all.

Consider an experiment in which the nuisance parameters $\theta$ are fixed throughout an experiment and feedback is given on every trial during testing. Then, the observer would be able to gradually learn the values of the nuisance parameters and thus the higher moments of the normalized likelihood function $L_I^*(s)$ (this would be harder if the higher moments also depend on the stimulus itself, and the stimulus varies from trial to trial). Once this is accomplished, there is no need for the brain to maintain a trial-to-trial representation of $L_I^*(s)$. Any subsequent computation would by definition be non-probabilistic. Therefore, a finding of near-optimality of a human observer in such an experiment would not provide evidence for a representation of uncertainty. In some signal detection theory and Bayesian studies (such as Jacobs, 1999; Kording et al., 2007; Palmer et al., 2000; Shimozaki et al., 2003; van Beers et al., 1996; Vincent et al., 2009), nuisance parameters were fixed and could be learned by the observer. These studies might demonstrate that humans perform optimal computation, but not that uncertainty is encoded on each individual trial.

Although it is best to minimize feedback when the goal is to demonstrate probabilistic computation, a complete lack of feedback is not always feasible. Training is often necessary to familiarize the subject with the task or to make the subjects learn prior distributions (such as the class distributions $p(s|C)$ in Example 1). Then, longer presentation times, higher contrasts etc. can be used in training than in testing. If feedback during testing is necessary, like when testing non-human primates, then it is advisable to use many different values of the nuisance parameters, so that it becomes harder to train on each individual one.

## 3. Neural representation of uncertainty: beyond single neurons

So far, we have described the observations, $I$, as unspecified sensory input. Now, we would like to represent $I$ by a neural quantity, as a first step towards understanding how neurons implement probabilistic computation (Fig. 2b). We limit ourselves to codes in which neural activity is measured by spike count or firing rate. The traditional view in systems neuroscience is that the firing rate of a single neuron represents the decision variable in a perceptual task. We argue that this view is limited to non-probabilistic computation (whether optimal or not), and that population codes are needed for probabilistic computation (whether optimal or not).

### 3.1. Standard view: one neuron fits all

The notion that single neurons are the key to the neural basis of perception has guided systems neuroscience at least since Barlow (Barlow, 1972). As indicated in the opening quote, by using the firing rate of a single neuron as the decision variable, signal detection theory methods can be used to compare the discrimination performances of a single neuron and of the animal (Bradley et al., 1987; Britten et al., 1992; Newsome et al., 1989; Parker & Newsome, 1998). The idea is that in binary decisions, there exists a one-to-one correspondence between the activity of a single neuron on an individual trial, $r$, and the value of the of the log posterior ratio of the two alternatives, $d$ from Eq. (5), on that trial. Neural activity is variable even when the same stimulus is presented repeatedly (Dean, 1981; Tolhurst, Movshon, & Dean, 1982), and this variability would correspond to the variability in $d$ induced by variability in sensory input, $I$. We examine two applications of this idea, to discrimination and visual search.

**Example 4** (*Discrimination with a single Gaussian neuron*). An observer discriminates between two stimulus values, $s_A$ and $s_B$. We denote the neuron's mean responses to these values by $f(s_A, \theta)$ and $f(s_B, \theta)$. Previous authors have modeled neural variability, for given nuisance parameters $\theta$, as Gaussian (Britten et al., 1992; Gold & Shadlen, 2001; Newsome et al., 1989):

$$p(r|s,\theta) = \frac{1}{\sqrt{2\pi\sigma^2_{\mathrm{neural}}(\theta)}} e^{-\frac{(r-f(s,\theta))^2}{2\sigma^2_{\mathrm{neural}}(\theta)}}, \tag{8}$$

where $\sigma^2_{\text{neutral}}(\theta)$ is the variance of the neuron's firing rate (see Fig. 4a). The neuron's response on a single trial is taken as the internal representation of the stimulus. Then, the log likelihood ratio is

$$\log\frac{L_r(s_A,\theta)}{L_r(s_B,\theta)} = \log\frac{p(r|s_A,\theta)}{p(r|s_B,\theta)} = \frac{f(s_A,\theta)-f(s_B,\theta)}{\sigma^2_{\text{neural}}(\theta)}\left(r-\frac{f(s_A,\theta)+f(s_B,\theta)}{2}\right). \quad (9)$$

Optimal inference amounts to estimating that the stimulus is $s_A$ when $r > (f(s_A,\theta)+f(s_B,\theta))/2$. If $\theta$ is fixed and known, or $f$ does not depend on $\theta$, this single neuron will discriminate as well as the optimal observer, regardless of $\sigma_{\text{neural}}$.

**Example 5** (*Homogeneous visual search with single Gaussian neurons*). In visual search with homogeneous distractors with a fixed value (Example 3), the observer has to report whether a target is present among $N$ stimuli. If each stimulus is encoded by one neuron, the observer makes a decision on the basis of the spike counts $\{r_i\}$ of $N$ neurons. All neurons are tuned identically, with their mean activity in response to the target higher than to the distractor. It has been proposed that, as in Example 4, $r_i$ corresponds to the noisy internal representation of a stimulus, and that the decision variable $d = \max_i x_i$, discussed in Section 2.3, corresponds to taking the maximum activity across all neurons: $d = \max_i r_i$ (Verghese, 2001). As mentioned there, this decision rule is close to optimal with respect to the ROC as long as nuisance parameters are identical for all stimuli and fixed across trials.

These examples show that a single neuron per stimulus is capable of optimal or near-optimal inference, as long as the computation is non-probabilistic. Then, the only information the observer needs on a given trial is the maximum-likelihood estimate of the stimulus. Since this is a single number, it can be encoded by the firing rate of a single neuron.

### 3.2. Single neurons cannot perform probabilistic computation

When computation is probabilistic, the observer uses higher moments of the stimulus likelihood that are independent of a point estimate of the stimulus. Can this be realized with a single Gaussian neuron, as described in the previous section? We consider two cases for the nuisance parameters $\theta$: fixed or variable. If $\theta$ is fixed and known with value $\theta_{\text{true}}$, the stimulus likelihood is $L_r(s;\theta_{\text{true}}) = p(r|s;\theta_{\text{true}})$. This will in general not be a Gaussian function (Fig. 4b and c). The maximum-likelihood estimate is the solution of the equation $f(s) = r$, whereas higher moments depend on $r$ and $\theta_{\text{true}}$. Since the only free parameter is $r$, higher moments are either constant or functions of the maximum-likelihood estimate (Fig. 4d). Therefore, it is impossible to encode stimulus uncertainty on a trial-to-trial basis as an independent variable, and computation cannot be probabilistic.

If, on the other hand, $\theta$ varies from trial to trial (for example, the contrast of a bar whose orientation is of interest is chosen randomly from multiple values), or is fixed but unknown, the likelihood has two free parameters. Although in principle, two independent moments can be encoded, the value of one of the parameters, $\theta$, is not known so cannot be utilized. One could propose two solutions to this problem. The first is to estimate $\theta$ on each trial through a separate mechanism, but this would require at least one additional neuron to encode that estimate. For example, in a study of decision-making, it has been proposed that time elapsed until the activity of an accumulator neuron reaches a decision termination bound can serve as a proxy of $\theta$ (Kiani & Shadlen, 2009). This strategy seems only reliable in tasks that allow the integration of information over hundreds of milliseconds. Moreover, in general, substituting a point estimate of $\theta$ is suboptimal, since $\theta$ cannot be estimated with infinite precision. The other solution, which is optimal, is to marginalize $\theta$ out in a way analogous to Eq. (1):

$$L_r(s) = p(r|s) = \int \frac{1}{\sqrt{2\pi\sigma^2_{\text{neural}}(\theta)}} e^{\frac{(r-f(s,\theta))^2}{2\sigma^2_{\text{neural}}(\theta)}} p(\theta)\,d\theta. \quad (10)$$

In this expression however, regardless of the form of $p(\theta)$, the only free parameter left is $r$ and the same argument can be made as when $\theta$ is fixed and known. It follows that in neither solution, a single neuron is sufficient for probabilistic computation.

This extends to estimation of decision confidence. In the discrimination task of Example 4, decision confidence can be quantified as the absolute value of the log likelihood ratio, and can therefore be computed from the activity of a single neuron. However, this is under the assumption that $\theta$ is fixed and known.

This problem is not limited to neurons with Gaussian variability. For example, if a neuron's variability is described by a Poisson distribution,

$$p(r|s,\theta) = \frac{e^{-f(s,\theta)}(f(s,\theta))^r}{r!},$$

then we find for the likelihood function:

$$L_r(s,\theta) = p(r|s,\theta) \propto e^{-f(s,\theta)} f(s,\theta)^r.$$

Both the maximum-likelihood estimate of $s$ and the variance (or any higher moment) of the normalized likelihood function will have $r$ as the only parameter, and are therefore not independent. Marginalization over $\theta$, similar to Eq. (10), does not help.

We conclude that regardless of whether nuisance variables are marginalized out, a single neuron cannot perform probabilistic

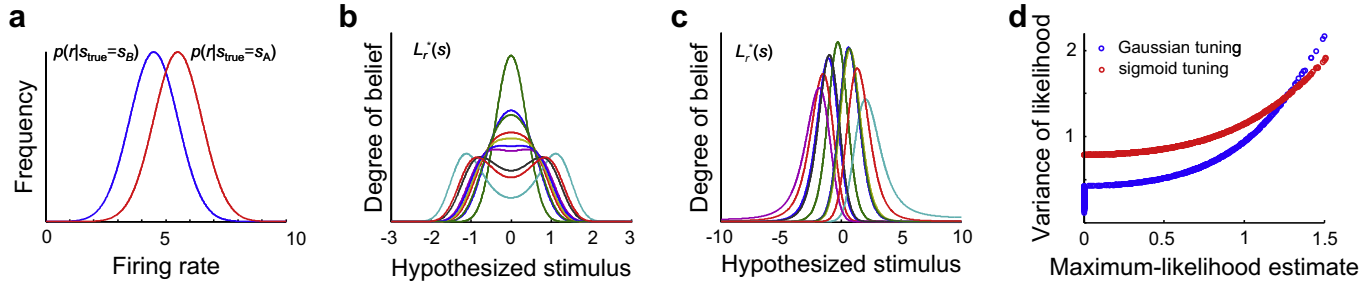


**Fig. 4.** Computation with single neurons. (a) Gaussian distribution of a single neuron firing rate $r$ conditioned on the two possible stimulus values, $s_A$ and $s_B$, in a discrimination task. (b) Normalized likelihood functions of the stimulus for a single neuron with Gaussian variability (with $\sigma_{\text{neural}} = 1$) and a Gaussian tuning curve $f(s) = g\cdot\exp(-s^2/(2\sigma^2_{\text{tc}}))$ (with $g = 10$ and $\sigma_{\text{tc}} = 1$), responding to a true stimulus value 0. (c) Normalized likelihood functions of the stimulus for a single neuron with Gaussian variability (with $\sigma_{\text{neural}} = 1$) and a sigmoid tuning curve $f(s) = g/(1+\exp(-\alpha s))$ (with $g = 10$ and $\alpha = 0.5$), responding to a true stimulus value 0. (d) Scatter plot of likelihood variance versus the maximum-likelihood estimate, for the neurons in (b) and (c). When the likelihood is bimodal (in (b)), the higher mode is chosen. The two quantities are completely correlated with each other.

computation, because higher moments are either constant and do not need to be encoded, or are functions of the point estimate and are therefore not independent measures of stimulus uncertainty.

### 3.3. Marginalization is a problem for Gaussian population codes

Since a single neuron is inadequate for probabilistic computation, i.e., to encode and utilize uncertainty about the stimulus on a trial-by-trial basis, a code involving multiple neurons is called for. In cortex, many stimuli elicit activity in large populations of neurons, and multiple theoretical schemes have been proposed for how a population could encode uncertainty (Ma, Beck, & Pouget, 2008a; Pouget et al., 2003). Here, we use the framework of probabilistic population coding, in which the stimulus likelihood is directly derived from neural variability in the same way as in Section 2, where we described the observations as sensory input.

Given the prominence of the single Gaussian neuron model for discrimination, a reasonable first try would be to consider a population of Gaussian neurons. We denote by $f_1(s,\theta),\ldots,f_n(s,\theta)$ the tuning curves of the neurons in a population, and by $\mathbf{r}=(r_1,\ldots,r_n)$ a specific pattern of activity. The simplest assumption for the relationship between the neurons is that they are independent, although that is not necessary for the following argument. A population of independent neurons with Gaussian variability is described by the following conditional probability distribution:

$$p(\mathbf{r}|s,\theta)=\prod_{j=1}^{n}p(r_j|s,\theta)=\prod_{j=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2_{\text{neural}}(\theta)}}e^{-\frac{(r_j-f_j(s,\theta))^2}{2\sigma^2_{\text{neural}}(\theta)}}.$$

The likelihood is obtained by marginalizing over $\theta$:

$$L_{\mathbf{r}}(s)=p(\mathbf{r}|s)=\int\left(\prod_{j=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2_{\text{neural}}(\theta)}}e^{-\frac{(r_j-f_j(s,\theta))^2}{2\sigma^2_{\text{neural}}(\theta)}}\right)p(\theta)d\theta, \tag{11}$$

We think of this likelihood function as being encoded in the population.[2] The function has $n$ parameters, $r_1$ to $r_n$, thereby allowing for the encoding of stimulus uncertainty on a trial-by-trial basis. The number of neurons equals the number of independent moments that can be encoded: with two neurons, one can encode a mode or mean and a variance, with three neurons also skewness, with four neurons also kurtosis, etc. Therefore, a population of Gaussian neurons can, in principle, be used for probabilistic computation. However, there are several other problems with Eq. (11), that already exist in the single-neuron case of Eq. (10). First, the likelihood and therefore the uncertainty about the stimulus depend on the parameter prior, $p(\theta)$. Second, a neural circuit must be able to compute the integral in Eq. (11). Finally, apart from any probabilistic considerations, the domain of a normally distributed variable is the entire real line, while neural activity is non-negative; this leads to inconsistency when the mean firing rate is low. These problems are reason to consider other forms of neural variability.

### 3.4. Gaussian neuron–antineuron system

In the context of a motion direction discrimination task, it has been proposed that a system of two uncorrelated neurons with Gaussian variability (and equal variances) implements the log likelihood ratio (Gold & Shadlen, 2001). This is a special case of the population of Gaussian neurons in the previous section. Let the stimulus values to be discriminated be $s_A$ and $s_B$. The first neuron responds to these stimuli with mean activities $f_1(s_A)$ and $f_1(s_B)$. The mean responses of the other neuron are reversed, $f_2(s_A)=f_1(s_B)$ and $f_2(s_B)=f_1(s_A)$. If the nuisance parameters are known, the log likelihood ratio is

$$\begin{aligned}\log\frac{L_{r_1,r_2}(s_A,\theta)}{L_{r_1,r_2}(s_B,\theta)}&=\log\frac{p(r_1,r_2|s_A,\theta)}{p(r_1,r_2|s_B,\theta)}=\log\frac{p(r_1|s_A,\theta)p(r_2|s_A,\theta)}{p(r_1|s_B,\theta)p(r_2|s_A,\theta)}\\&=\frac{1}{2\sigma^2_{\text{neural}}(\theta)}\left(-(r_1-f_1(s_A))^2-(r_2-f_2(s_A))^2\right.\\&\quad\left.+(r_1-f_1(s_B))^2+(r_2-f_2(s_B))^2\right)\\&=\frac{f_1(s_A)-f_1(s_B)}{\sigma^2_{\text{neural}}(\theta)}(r_1-r_2).\end{aligned} \tag{12}$$

When confidence is measured as the absolute value of the log likelihood ratio, it is then directly proportional to the absolute difference between the activities of the neuron and the antineuron. However, this is not probabilistic computation, since $\theta$ is assumed known. If $\theta$ is not known, it has to be marginalized over, as in Eq. (11). This leads to the expression

$$\begin{aligned}\log\frac{L_{r_1,r_2}(s_A)}{L_{r_1,r_2}(s_B)}&=\log\frac{\int p(r_1,r_2|s_A,\theta)p(\theta)d\theta}{\int p(r_1,r_2|s_B,\theta)p(\theta)d\theta}=\log\frac{\int p(r_1|s_A,\theta)p(r_2|s_A,\theta)p(\theta)d\theta}{\int p(r_1|s_B,\theta)p(r_2|s_A,\theta)p(\theta)d\theta}\\&=\log\frac{\int\frac{1}{\sigma^2_{\text{neural}}(\theta)}\exp\left(-\frac{1}{2\sigma^2_{\text{neural}}(\theta)}((r_1-f_1(s_A))^2+(r_2-f_2(s_A))^2)\right)p(\theta)d\theta\cdot}{\int\frac{1}{\sigma^2_{\text{neural}}(\theta)}\exp\left(-\frac{1}{2\sigma^2_{\text{neural}}(\theta)}((r_1-f_1(s_B))^2+(r_2-f_2(s_B))^2)\right)p(\theta)d\theta}\end{aligned}$$

This expression is a lot less palatable than Eq. (12). This illustrates the problem with Gaussian (and in fact many other types of) variability: as long as the population contains at least two neurons, probabilistic computation is possible in principle, but difficult in practice because of the marginalization over $\theta$.

### 3.5. Poisson-like population codes solve the marginalization problem

The marginalization problem is solved if population variability is Poisson-like (Ma et al., 2006), which is defined as

$$p(\mathbf{r}|s,\theta)=\varphi(\mathbf{r},\theta)e^{\mathbf{h}(s)\cdot\mathbf{r}}, \tag{13}$$

where $\varphi$ is an arbitrary function of $\mathbf{r}$ and $\theta$. This is a family of distributions characterized by a set of functions $\mathbf{h}(s)$, with one $h_i(s)$ belonging to each neuron in the population. These functions can be computed from the tuning curves of the neurons, $\mathbf{f}(s,\theta)$, and the covariance matrix of the population (Ma et al., 2006). Poisson-like variability solves the marginalization problem, because the likelihood function of the stimulus is

$$\begin{aligned}L_{\mathbf{r}}(s)=p(\mathbf{r}|s)&=\int p(\mathbf{r}|s,\theta)p(\theta)d\theta=\int\varphi(\mathbf{r},\theta)e^{\mathbf{h}(s)\cdot\mathbf{r}}p(\theta)d\theta\\&=\left(\int\varphi(\mathbf{r},\theta)p(\theta)d\theta\right)e^{\mathbf{h}(s)\cdot\mathbf{r}}\propto e^{\mathbf{h}(s)\cdot\mathbf{r}}.\end{aligned} \tag{14}$$

The key feature of this equation is that unlike in Eq. (11), the $\theta$-dependent factors in the integral are separable from the $s$-dependent ones. As a consequence, the normalized likelihood when the nuisance parameters are marginalized out, $L^*_{\mathbf{r}}(s)$, is identical to the normalized likelihood when they are known, $L^*_{\mathbf{r}}(s,\theta)$. Specifically, the normalized likelihood is independent of the prior over nuisance parameters, $p(\theta)$. Therefore, uncertainty about the stimulus can be estimated from the population activity $\mathbf{r}$ without knowledge of $\theta$. This is the most important benefit of Poisson-like variability.

Poisson-like variability is not the only form of variability that allows a separation of the $s$- and $\theta$-dependent factors in marginalizing over $\theta$. One could replace the neural activity $\mathbf{r}$ in the exponent in Eq. (13) by any function of $\mathbf{r}$ to achieve the same result. However, there are good reasons to use a linear function. First, it turns

[2] It is often stated that a neural population encodes a *posterior distribution* over the stimulus (Ma et al., 2006; Pouget et al., 2003). The more accurate statement is that it encodes a *normalized likelihood function* of the stimulus. The stimulus prior is not encoded without further assumptions.

out to be consistent with the observed property of neural firing that its variance is approximately proportional to its mean (Ma et al., 2006). Second, for the special case of independent Poisson variability, the function is linear (see Section 3.7). Third, the linearity allows for simple neural implementations of optimal cue combination (Ma et al., 2006) and optimal decision-making (Beck et al., 2008; Ma et al., 2006).

### 3.6. Laplace approximation to the Poisson-like likelihood function

After normalization, the neural likelihood, Eq. (14), is close to a Gaussian distribution when the product of the number of neurons and the gain is sufficiently large (Figs. 1a and 5b). When this product is lower, the likelihood may cease to look Gaussian but will typically still have a dominant peak that can be reasonably approximated by a Gaussian (Fig. 5b). This is called the Laplace approximation (MacKay, 2003). The maximum-likelihood estimate is

$$x = \underset{s}{\operatorname{argmax}}\, p(\mathbf{r}|s) = \underset{s}{\operatorname{argmax}}(\mathbf{h}(s)\cdot\mathbf{r}). \tag{15}$$

We expand the exponent in Eq. (14) about the maximum-likelihood estimate:

$$\begin{aligned}\mathbf{h}(s)\cdot\mathbf{r} &= \mathbf{h}(x)\cdot\mathbf{r} + (\mathbf{h}'(x)\cdot\mathbf{r})(s-x) + \tfrac{1}{2}(\mathbf{h}''(x)\cdot\mathbf{r})(s-x)^2 + \ldots \\ &= \mathbf{h}(x)\cdot\mathbf{r} + \tfrac{1}{2}(\mathbf{h}''(x)\cdot\mathbf{r})(s-x)^2 + \ldots\end{aligned} \tag{16}$$

Under this approximation, we can write the normalized likelihood as a Gaussian distribution,

$$L^*_{\mathbf{r}}(s) \approx \frac{1}{\sqrt{2\pi\sigma^2(\mathbf{r})}} e^{-\frac{(s-x(\mathbf{r}))^2}{2\sigma^2(\mathbf{r})}}, \tag{17}$$

where the variance

$$\sigma^2(\mathbf{r}) = -\frac{1}{\mathbf{h}''(x)\cdot\mathbf{r}} \tag{18}$$

measures the uncertainty about the stimulus. It varies from trial to trial even when the nuisance parameters are fixed and known (Fig. 1a). The approximate likelihood function, Eq. (17), is the neural equivalent of the more abstract expression in Eq. (2); the latter can be regarded as a Laplace approximation to a general likelihood function. If $\mathbf{h}(s)$ is quadratic, Eq. (17) is exact because the higher-order terms in the Taylor series, Eq. (16), vanish. Then, the neural and abstract likelihoods can be equated exactly; this gives

$$\mathbf{h}(s)\cdot\mathbf{r} = (\mathbf{a}\cdot\mathbf{r})s^2 + (\mathbf{b}\cdot\mathbf{r})s, \tag{19}$$

with $\mathbf{a}\cdot\mathbf{r} = -1/(2\sigma^2)$ (consistent with Eq. (18)) and $\mathbf{b}\cdot\mathbf{r} = -x/\sigma^2$. The Laplace approximation will break down under high uncertainty, i.e., when the product of number of neurons and gain is very low.

### 3.7. Example: independent Poisson population

Independent Poisson variability is a special case of Poisson-like variability, in which the neurons are assumed conditionally independent. The distribution of activity is given by

$$p(\mathbf{r}|s,\theta) = \prod_{j=1}^{n} p(r_j|s,\theta) = \prod_{j=1}^{n} \frac{e^{-f_j(s,\theta)} f_j(s,\theta)^{r_j}}{r_j!}.$$

We now assume that the $\theta$-dependence is separable, i.e., $f_j(s,\theta) = g(\theta)f_j(s)$. This would not hold if the nuisance parameters affected the width of the tuning curve. Under this assumption,

$$\begin{aligned}L_{\mathbf{r}}(s) = p(\mathbf{r}|s) &= \int \left(\prod_{j=1}^{n} \frac{e^{-g(\theta)f_j(s)} g(\theta)^{r_j} f_j(s)^{r_j}}{r_j!}\right) p(\theta)\,d\theta \\ &= \exp\left(\sum_{j=1}^{n} r_j \log f_j(s)\right) \int \exp\left(-g(\theta)\sum_{j=1}^{n} f_j(s)\right) \left(\prod_{j=1}^{n} \frac{g(\theta)^{r_j}}{r_j!}\right) p(\theta)\,d\theta \\ &\propto \exp\left(\sum_{j=1}^{n} r_j \log f_j(s)\right),\end{aligned}$$

where in the last step we have assumed that $\sum_j f_j(s) = \text{constant}$. This is approximately satisfied when the tuning curves are shifted versions of each other, and preferred orientations are equally and closely spaced. Again, $L_{\mathbf{r}}(s)$ does not depend on the form of $p(\theta)$. Here, $h_j(s) = \log f_j(s)$.

We further work out the case of equal-width Gaussian tuning curves. The tuning curve of the $j$th neuron is

$$f_j(s) = e^{-\frac{(s-s_j^{\text{pref}})^2}{2\sigma^2_{\text{tuning}}}},$$

where $s_j^{\text{pref}}$ is the preferred stimulus of the $j$th neuron and $\sigma_{\text{tuning}}$ is the width of the tuning curve. Then the likelihood function is

$$L_{\mathbf{r}}(s) \propto \exp\left(-\frac{1}{2\sigma^2_{\text{tuning}}} \sum_{j=1}^{n} r_j \left(s - s_j^{\text{pref}}\right)^2\right).$$

After rewriting, the maximum-likelihood estimate is found to be $x(\mathbf{r}) = \frac{\sum_j r_j s_j^{\text{pref}}}{\sum_j r_j}$; this is known as the center-of-mass or (on a periodic space) the population vector decoder (Georgopoulos, Kalaska, Caminiti, & Massey, 1982). The variance of the likelihood function is

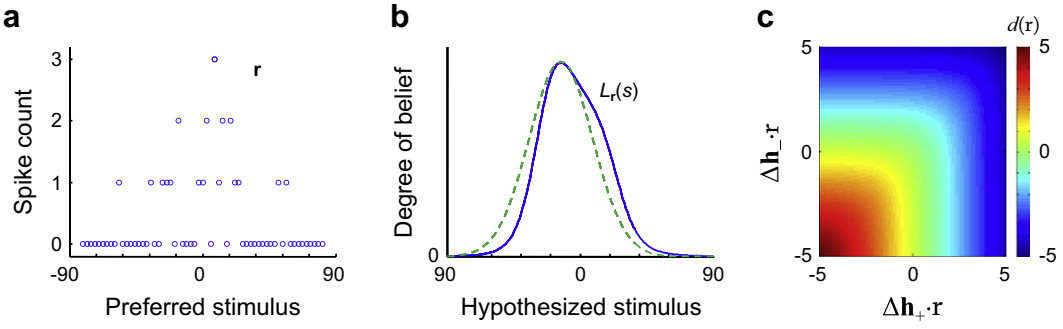


**Fig. 5.** Computation with populations. The generative model is in Fig. 2b. (a) A population pattern of activity. Neurons are ordered by their preferred stimulus, which ranges from −90 to 90 in steps of 3. Tuning curves are Gaussian with width 20, gain 1, and baseline 0.1. Variability is independent Poisson. The normalized likelihood functions of the stimulus in Fig. 1a were obtained from activity patterns in this population. In spite of the low gain, these likelihood functions are close to Gaussian. (b) Laplace approximation (green) to a non-Gaussian likelihood function (blue). (c) Example of an optimal decision variable $d(\mathbf{r})$ that is nonlinear in neural activity $\mathbf{r}$. The task is to discriminate a stimulus value $s_A$ from two flanking values, $s_A - \Delta s$ and $s_A + \Delta s$. We use the notation $\Delta\mathbf{h}_\pm = \mathbf{h}(s_A \pm \Delta s) - \mathbf{h}(s_A)$. The value of the decision variable is color-coded. A linear function would have linear equi-color lines.

$\frac{\sigma^2_{\text{tuning}}}{\sum_j r_j}$, i.e., the uncertainty about the stimulus is determined by the width of the tuning curve and the total activity in the population. The variance is independent of the maximum-likelihood estimate, so the population activity encodes two independent numbers that vary from trial to trial and together specify the stimulus likelihood completely. Therefore, this population can be used for probabilistic computation. Since $\mathbf{h}(s)$ is quadratic, Eq. (17) is exact.

### 3.8. Optimal neural computation

In optimal computation, the posterior over $C$ is a functional of the normalized stimulus likelihood (Eq. (7)). This posterior is now based on neural activity and we denote it $p(C|\mathbf{r})$. If we adopt Poisson-like probabilistic population codes, Eq. (14), the neural equivalent of this expression is

$$p(C|\mathbf{r}) = \Phi_C[\mathbf{h}(s)\cdot\mathbf{r}],$$

with $\Phi_C$ a task-dependent functional. When $\mathbf{h}(s)$ is quadratic, as in Eq. (19), $p(C|\mathbf{r})$ reduces to a function of $\mathbf{a}\cdot\mathbf{r}$ and $\mathbf{b}\cdot\mathbf{r}$.

**Example 6** (*Discrimination*). We return to Example 4, discrimination between $s_A$ and $s_B$, but now based on Poisson-like population activity $\mathbf{r}$. The log likelihood ratio is

$$d = \log\frac{p(\mathbf{r}|s=s_A)}{p(\mathbf{r}|s=s_B)} = \mathbf{h}(s_A)\cdot\mathbf{r} - \mathbf{h}(s_B)\cdot\mathbf{r} \equiv \Delta\mathbf{h}\cdot\mathbf{r}, \tag{20}$$

where the last equality is just a notation. This shows that the optimal decision variable after marginalizing out the nuisance parameters is a linear combination of population activity. Confidence can be measured by the absolute value of $\Delta\mathbf{h}\cdot\mathbf{r}$. In contrast to Example 4, no separate estimation of nuisance parameters is needed for either reporting decision confidence or optimal downstream computation.

**Example 7** (*Classification*). We return to Example 1. Expressed in terms of neural activity, the likelihood of class is

$$p(\mathbf{r}|C) = \int p(\mathbf{r}|s)p(s|C)ds = \int L_{\mathbf{r}}(s)p(s|C)ds.$$

The optimal decision variable is based on the log posterior ratio of the two classes, as in Eq. (4):

$$\begin{aligned} d &= \log\frac{p(C=1)}{p(C=-1)} + \log\frac{p(\mathbf{r}|C=1)}{p(\mathbf{r}|C=-1)} \\ &= \log\frac{p(C=1)}{p(C=-1)} + \log\frac{\int L_{\mathbf{r}}(s)p(s|C=1)ds}{\int L_{\mathbf{r}}(s)p(s|C=-1)ds}. \end{aligned} \tag{21}$$

Thus, the optimal neural decision variable depends on the class distributions $p(s|C)$. Discrimination (Example 6) is a trivial special case, with $p(s|C=1)=\delta(s-s_A)$ and $p(s|C=-1)=\delta(s-s_B)$, where $\delta$ denotes the Dirac delta function. A more complex case is when a $C=1$ stimulus takes the value $s_A$, whereas a $C=-1$ stimulus takes one of two "flanking" values, $s_A-\Delta s$ and $s_A+\Delta s$, with equal probability. Then we have $p(s|C=1)=\delta(s-s_A)$ and $p(s|C=-1)=[\delta(s-s_A-\Delta s)+\delta(s-s_A+\Delta s)]$, and the log likelihood ratio is

$$\begin{aligned} \log\frac{p(\mathbf{r}|C=1)}{p(\mathbf{r}|C=-1)} &= \log\frac{\int e^{\mathbf{h}(s)\cdot\mathbf{r}}\delta(s-s_A)ds}{\int e^{\mathbf{h}(s)\cdot\mathbf{r}}\frac{1}{2}(\delta(s-s_A-\Delta s)+\delta(s-s_A+\Delta s))ds} \\ &= -\log\frac{e^{\Delta\mathbf{h}_+\cdot\mathbf{r}}+e^{\Delta\mathbf{h}_-\cdot\mathbf{r}}}{2}, \end{aligned} \tag{22}$$

where $\Delta\mathbf{h}_\pm = \mathbf{h}(s_A\pm\Delta s)-\mathbf{h}(s_A)$. In contrast to Example 6, this is a nonlinear function of population activity (Fig. 5c).

**Example 8** (*Homogeneous visual search*). In visual search with homogeneous distractors with a fixed value, the log likelihood ratio of target presence at the $i$th location is identical to the log likelihood ratio in discrimination (Example 6), $\Delta\mathbf{h}_i\cdot\mathbf{r}_i$. The global likelihood ratio of target presence is obtained by combining these quantities nonlinearly across locations according to Eq. (6):

$$\frac{p(\mathbf{r}|C=1)}{p(\mathbf{r}|C=-1)} = \frac{1}{N}\sum_{i=1}^{N} e^{\Delta\mathbf{h}}_i\cdot\mathbf{r}_i.$$

This solves several problems associated with the single-neuron proposal (Example 5). Like the local log likelihood ratio, $\Delta\mathbf{h}_i\cdot\mathbf{r}_i$ can take both positive and negative values. By contrast, a single neuron's spike count is only defined for positive values. Moreover, the local log likelihood ratio is linear in $x_i$ (Eq. (6)) and therefore follows an approximately Gaussian distribution (see Section 4.1). However, a single neuron has a spike count distribution that is closer to Poisson, with variance being proportional to the mean. A Poisson distribution is similar to a Gaussian only for large means. By contrast, $\Delta\mathbf{h}_i\cdot\mathbf{r}_i$ is close to normally distributed when there are many neurons, in view of the central limit theorem, and its variance is not necessarily proportional to its mean.

A neural network is said to perform optimal computation if it maps Poisson-like input activity $\mathbf{r}$ to output activity $\mathbf{z}$, such that $p(C|\mathbf{z}) = p(C|\mathbf{r}) = \Phi_C(\mathbf{h}(s)\cdot\mathbf{r})$. Trivially, one might take $\mathbf{z}=\mathbf{r}$ and the equation holds: every posterior distribution that the brain computes from a visual scene is already encoded in the retina. However, the purpose of computation is to reformat the input into a form that is easier to decode. Therefore, one has to impose a constraint on $\mathbf{z}$. In previous work, we have required that the output activity is again a Poisson-like population code (Ma et al., 2006; Ma et al., 2008b). For a binary variable $C$, this is equivalent to stating that the log likelihood ratio of $C$ is linear in $\mathbf{z}$:

$$d = \log\frac{p(\mathbf{z}|C=1)}{p(\mathbf{z}|C=-1)} = (\mathbf{H}(C=1)-\mathbf{H}(C=-1))\cdot\mathbf{z}$$

(compare Eq. (20)). Therefore, finding an optimal network amounts to finding $\mathbf{z}$ such that for some $\mathbf{H}(C)$,

$$(\mathbf{H}(C=1)-\mathbf{H}(C=-1))\cdot\mathbf{z} = \log\frac{\Phi_{C=1}[\mathbf{h}(s)\cdot\mathbf{r}]}{\Phi_{C=-1}[\mathbf{h}(s)\cdot\mathbf{r}]}. \tag{23}$$

One can find the types of operations $\mathbf{r}\to\mathbf{z}$ needed in an optimal network by expanding the right-hand side for the task at hand under the assumption that $\mathbf{h}(s)$ is quadratic.

## 4. Signal detection theory with Poisson-like population codes

To conclude, we establish correspondences between signal detection variables and Poisson-like population quantities (see Table 1). Therefore, this section deals with binary variables $C$ and optimal, but not necessarily probabilistic computation.

### 4.1. The MLE distribution

In Section 2, we described sensory input as an abstract quantity $I$, and noted that the maximum-likelihood of the stimulus and the variance of the normalized stimulus likelihood are functions of $I$ (Eq. (2)). In the limit that $I$ is highly informative about the stimulus, the maximum-likelihood estimate itself has an approximately Gaussian distribution across many trials,

$$p(x|s_{\text{true}},\theta_{\text{true}}) = \frac{1}{\sqrt{2\pi\sigma^2_{\text{MLE}}(s_{\text{true}},\theta_{\text{true}})}} e^{-\frac{(x-s_{\text{true}})^2}{2\sigma^2_{\text{MLE}}(s_{\text{true}},\theta_{\text{true}})}}. \tag{24}$$

**Table 1**
Correspondence between signal detection theory and Poisson-like population quantities.

| Signal detection theory quantity | Poisson-like population code quantity |
|---|---|
| Sensory input (observation) $I$ | Population activity $\mathbf{r}$ |
| Sensory variability, $p(I\|s,\theta)$ | Poisson-like neural variability, $p(\mathbf{r}\|s,\theta) = \varphi(\mathbf{r},\theta)\exp(\mathbf{h}(s)\cdot\mathbf{r})$ |
| Marginalized likelihood function, $L_I(s) = p(I\|s) = \int p(I\|s,\theta)p(\theta)d\theta$ | Neural likelihood function independent of $p(\theta)$, $L_{\mathbf{r}}(s) \propto e^{\mathbf{h}(s)\cdot\mathbf{r}}$ |
| Gaussian likelihood assumption (1D) $L_I^*(s) = \frac{1}{\sqrt{2\pi\sigma^2(I)}} e^{-\frac{(s-x(I))^2}{2\sigma^2(I)}}$ | Laplace approximation to likelihood (1D) $L_{\mathbf{r}}^*(s) = \frac{1}{\sqrt{2\pi\sigma^2(\mathbf{r})}} e^{-\frac{(s-x(\mathbf{r}))^2}{2\sigma^2(\mathbf{r})}}$ Exact when $\mathbf{h}(s)$ quadratic |
| Maximum-likelihood estimate ("internal representation") $x = \operatorname{argmax}_s p(I\|s,\theta)$ | $x = \operatorname{argmax}_s p(\mathbf{r}\|s) = \operatorname{argmax}_s(\mathbf{h}(s)\cdot\mathbf{r})$ Approximation: $x \approx s_{\text{true}} - \frac{\mathbf{h}'(s_{\text{true}})\cdot\mathbf{r}}{\mathbf{h}''(s_{\text{true}})\cdot\mathbf{r}}$ |
| Variance of likelihood function $\sigma^2(I)$ | $\sigma^2(\mathbf{r}) \approx -(\mathbf{h}''(x)\cdot\mathbf{r})^{-1}$ |
| Variance of $x$ given $s_{\text{true}}$ and $\theta_{\text{true}}$ $\frac{1}{\sigma^2_{\text{MLE}}(s_{\text{true}},\theta_{\text{true}})} \approx \langle\frac{1}{\sigma^2(I)}\rangle_{p(I\|s_{\text{true}},\theta_{\text{true}})}$ | Inverse Fisher information $\frac{1}{\sigma^2_{\text{MLE}}(s_{\text{true}},\theta_{\text{true}})} = -\mathbf{h}''(s_{\text{true}})\cdot\mathbf{f}(s_{\text{true}},\theta_{\text{true}})$ |
| Signal detection theory decision variable (log posterior ratio over $C$) $d(I) = \log\frac{p(C=1\|I)}{p(C=-1\|I)}$ | Functional of $\mathbf{h}(s)\cdot\mathbf{r}$ $d(\mathbf{r}) = \log\frac{p(C=1\|\mathbf{r})}{p(C=-1\|\mathbf{r})} = \log\frac{\Phi_{C=1}[\mathbf{h}(s)\cdot\mathbf{r}]}{\Phi_{C=-1}[\mathbf{h}(s)\cdot\mathbf{r}]} + \log\frac{p(C=1)}{p(C=-1)}$ |
| Confidence rating: $\|d(I)\|$ | $\|d(\mathbf{r})\|$ |
| Decision variable in discrimination (Gaussian likelihood, flat prior) $d(I) = \log\frac{p(I\|s_A)}{p(I\|s_B)} = \frac{s_A-s_B}{\sigma^2}\left(x - \frac{s_A+s_B}{2}\right)$ | Linear combination of population activity, $d(\mathbf{r}) = (\mathbf{h}(s_A) - \mathbf{h}(s_B))\cdot\mathbf{r} \equiv \Delta\mathbf{h}\cdot\mathbf{r}$ |
| Sensitivity $d'$ in discrimination $d' = \frac{s_A-s_B}{\sigma}$ | $d' = \frac{\Delta\mathbf{h}\cdot(\mathbf{f}(s_A,\theta)-\mathbf{f}(s_B,\theta))}{\sqrt{\frac{1}{2}\Delta\mathbf{h}^{\mathrm{T}}(\sum(s_A,\theta)+\sum(s_B,\theta))\Delta\mathbf{h}}}$ |

This is the well-known Gaussian assumption (Green & Swets, 1966). It is related to but different from the assumption that the normalized likelihood of the stimulus is Gaussian, Eq. (2).

The variance $\sigma^2_{\text{MLE}}$ is computed as the inverse of Fisher information (Cover & Thomas, 1991; Paradiso, Carney, & Freeman, 1989; Seung & Sompolinsky, 1993). Fisher information is defined as

$$J(s_{\text{true}},\theta_{\text{true}}) = -\left\langle \frac{\partial^2}{\partial s^2}\log p(I|s)\right\rangle_{p(I|s_{\text{true}},\theta_{\text{true}})} = \left\langle\frac{1}{\sigma^2(I)}\right\rangle_{p(I|s_{\text{true}},\theta_{\text{true}})}, \tag{25}$$

where the last equality follows from Eq. (2). If we ignore trial-to-trial fluctuations in $\sigma^2(I)$, then the variance of the maximum-likelihood estimate equals the variance of the normalized likelihood function. One can think of the pair $(x, \sigma^2_{\text{MLE}})$ as a summary representation of $I$. An optimal model of any psychophysical task can be built self-consistently if one uses those quantities as observations instead of $I$ (Fig. 2c).

At the neural level, our starting point is Eq. (15) for the maximum-likelihood estimate. In general, it is not possible to obtain a closed-form expression for $x$. However, a decent approximation is obtained by expanding the function $\mathbf{h}'(s)\cdot\mathbf{r}$ about the true stimulus $s_{\text{true}}$ (Wu, Nakahara, & Amari, 2001):

$$\mathbf{h}'(s)\cdot\mathbf{r} = \mathbf{h}'(s_{\text{true}})\cdot\mathbf{r} + (\mathbf{h}''(s_{\text{true}})\cdot\mathbf{r})(s-s_{\text{true}}) + \ldots, \tag{26}$$

and substituting $s = x$ so that the left-hand side vanishes (since $x$ is a maximum). The result is

$$x \approx s_{\text{true}} - \frac{\mathbf{h}'(s_{\text{true}})\cdot\mathbf{r}}{\mathbf{h}''(s_{\text{true}})\cdot\mathbf{r}}. \tag{27}$$

This approximation is exact when $\mathbf{h}(s)$ is quadratic, since then the higher-order terms in Eq. (26) vanish. As in Eq. (25), the variance of $x$ is approximately given by the inverse of Fisher information, which is here

$$\sigma^2(s_{\text{true}},\theta_{\text{true}}) = -(\mathbf{h}''(s_{\text{true}})\cdot\mathbf{f}(s_{\text{true}},\theta_{\text{true}}))^{-1}.$$

### 4.2. Sensitivity

The optimal choice between the alternatives $C = 1$ and $C = -1$ is based on the log posterior ratio $d$ of Eqs. (5) and (21). The observer's performance can be derived from the distributions of $d$ when the observations ($I$ or $\mathbf{r}$) are drawn from either class across many trials. These distributions are (in the neural version) given by the integral

$$p(d|C) = \int\int\int p(d|\mathbf{r})p(\mathbf{r}|s,\theta)p(s|C)p(\theta)d\mathbf{r}\,ds\,d\theta, \tag{28}$$

(compare Eq. (3)), where

$$p(d|\mathbf{r}) = \delta\left(d - \log\frac{p(C=1|\mathbf{r})}{p(C=-1|\mathbf{r})}\right),$$

which expresses that the observer performs maximum-likelihood estimation (an alternative is sampling from the posterior (Mamassian & Landy, 1998)). An ROC is obtained from the distributions of $d$ conditioned on $I$ being drawn from one of both classes. Sensitivity, denoted $d'$, is well-defined if the distributions $p(d|C)$ are approximately Gaussian for both values of $C$. This is true in the simple case of discrimination, Eq. (20). When well-defined, sensitivity is the difference between the mean values of $d$ under both alternatives divided by the standard deviation of $d$. For discrimination at particular true values of the nuisance parameters, $\theta_{\text{true}}$, the difference in means is $\Delta\mathbf{h}\cdot(\mathbf{f}(s_A,\theta_{\text{true}}) - \mathbf{f}(s_B,\theta_{\text{true}}))$. The average variance of $\Delta\mathbf{h}\cdot\mathbf{r}$ is $\Delta\mathbf{h}^{\mathrm{T}}(\mathbf{\Sigma}(s_A,\theta_{\text{true}}) + \mathbf{\Sigma}(s_B,\theta_{\text{true}}))\Delta\mathbf{h}/2$, where "T" denotes transpose and $\mathbf{\Sigma}$ is the covariance matrix of the population. We find for sensitivity (compare (Johnson, 1980)):

$$d' = \frac{\Delta\mathbf{h}\cdot(\mathbf{f}(s_A,\theta_{\text{true}}) - \mathbf{f}(s_B,\theta_{\text{true}}))}{\sqrt{\frac{1}{2}\Delta\mathbf{h}^{\mathrm{T}}(\mathbf{\Sigma}(s_A,\theta_{\text{true}}) + \mathbf{\Sigma}(s_B,\theta_{\text{true}}))\Delta\mathbf{h}}}.$$

For tasks more complex than discrimination, the distributions in Eq. (28) often have to be computed numerically, for example through Monte Carlo simulation.

### 4.3. Estimating $\Delta\mathbf{h}$

Identifying the optimal decision variable with a neural population quantity is only of practical value if the latter can be computed from a population recording. If the recorded population activity on a given trial is $\mathbf{r}$ and variability is Poisson-like, we find from Eq. (20):

$$p(s_A|\mathbf{r}) = \frac{1}{1+\exp(-\Delta\mathbf{h}\cdot\mathbf{r})} = \Lambda(\Delta\mathbf{h}\cdot\mathbf{r}), \tag{29}$$

where $\Lambda(x) = 1/(1+\exp(-x))$ is the logistic function. Estimating $\Delta\mathbf{h}$ from samples $(s,\mathbf{r})$ is a logistic regression problem, which can be solved using standard techniques. Here, we use variational Bayesian logistic regression (VBLR) (Bishop, 2006). We simulated 2500 trials by drawing population patterns of activity from an independent Poisson distribution. Fig. 6a shows the theoretical weights $\Delta\mathbf{h}$ (blue) along with the best VBLR estimate. We did the same for a simple correlated population (Fig. 6b). Starting with a population

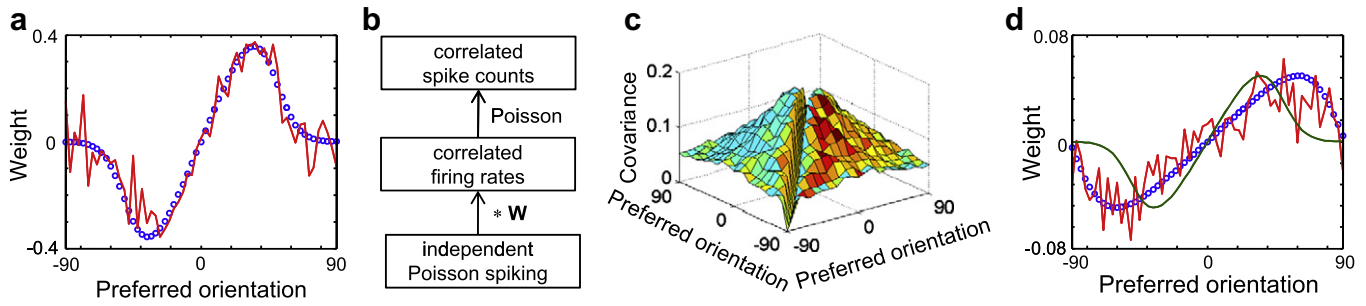


**Fig. 6.** Estimating $\Delta\mathbf{h}$ from simulated neural data. (a) Weights estimated from 2500 trials using variational Bayesian logistic regression in the independent Poisson case (red); optimal weights are in blue. (b) Simple procedure for creating correlated, approximately Poisson-like populations. For the weights $\mathbf{W}$, we used a translation-invariant Gaussian profile of width 50. (c) Resulting covariance structure, with the diagonal (variance) removed for visibility. (d) As (a), but for the correlated population. The green line shows the (scaled) weights under the (wrong) assumption of independence.

of independent Poisson neurons, we applied a symmetric weight matrix $\mathbf{W}$ with a Gaussian profile of connection strengths; this was simply chosen as an illustration. The weighted sums $\mathbf{Wr}$ were used as rates for another set of Poisson processes. Neurons in the output layer became correlated due to shared input. The covariance structure is shown in Fig. 6c. The resulting population is approximately Poisson-like, with mean $\mathbf{Wf}(s)$ and covariance matrix $\mathbf{\Sigma} = \mathrm{diag}(\mathbf{Wf}(s)) + \mathbf{W}\mathrm{diag}(\mathbf{f}(s))\mathbf{W}$, where $\mathrm{diag}(\cdot)$ indicates a diagonal matrix with the argument vector on the diagonal. Using this expression, we computed the optimal decision weights as $\Delta\mathbf{h} = \mathbf{\Sigma}^{-1}\mathbf{W}\Delta\mathbf{f}$. These are shown along with the VBLR estimate in Fig. 6d. The weights differ from the ones obtained if one would wrongly assume the neurons to be independent.

Once the weights $\Delta\mathbf{h}$ have been estimated, checks can be performed to determine how close the population is to being Poisson-like: generalization performance on a new data set (reconstructing $s$ from $\mathbf{r}$) should be high; $\Delta\mathbf{h}$ should not depend strongly on nuisance parameters like contrast; no nonlinear decoder should perform significantly better than the linear decoder $\Delta\mathbf{h}$; and the relationship $\mathbf{h}'(s) = \mathbf{\Sigma}^{-1}(s,\theta)\mathbf{f}'(s,\theta)$ should hold, where $\mathbf{\Sigma}$ is the covariance matrix of the population (though this might be difficult to estimate in practice). Moreover, a population ROC can be obtained similarly to a single-neuron ROC, except that $\Delta\mathbf{h}\cdot\mathbf{r}$ is used instead of single-neuron activity. To do this, the proportion of activity patterns $\mathbf{r}$ elicited by $s_A$ for which $\Delta\mathbf{h}\cdot\mathbf{r}$ exceeds a given criterion $c$ is plotted against the proportion of patterns elicited by $s_B$ that do so. Then, $c$ is varied along the real line to sweep out a curve. Finally, once $\Delta\mathbf{h}$ has been estimated, Eq. (29) provides the posterior probability distribution over class on a trial-to-trial basis. This can be correlated with the value of a nuisance parameter that controls uncertainty, or with a behavioral measure of decision confidence.

## 5. Summary

In binary decisions usually studied in signal detection theory models, using a population of neurons with different tuning clearly leads to increased sensitivity compared to a single neuron (Eckstein, Peterson, Pham, & Droll, 2009). Here, we argue that population coding is necessary for a more important reason: to allow for any form of probabilistic computation. Taking the likelihood function over the stimulus as the central object, we made distinctions between probabilistic models, models of optimal computation, and models of probabilistic computation. Every model with sensory noise is a probabilistic model, but only in models of probabilistic computation, the observer uses a neural representation of uncertainty on a trial-by-trial basis. Referring to the opening quote, it is possible for an observer to be less than 100% correct on a task, due to noise, and still never encode uncertainty. While all signal detection theory models are probabilistic, some are not optimal and many are not models of probabilistic computation (including models for detection and discrimination, and max and sum models of visual search). In optimal computation, a posterior is computed, which always depends on the full likelihood function over the stimulus; however, for the MAP estimate, only a point estimate of the stimulus might suffice. Some signal detection theory models make suboptimal, non-probabilistic approximations to the optimal, probabilistic computation. We have proposed requirements for a psychophysics experiment to provide evidence for probabilistic computation.

Single-neuron coding is inadequate for probabilistic computation, because the latter requires that the brain encodes, on a trial-by-trial basis, at least two numbers for each stimulus, for example the maximum-likelihood estimate and the variance of the normalized likelihood function. Population coding allows to encode uncertainty on a single trial, and Poisson-like variability additionally solves the problem of marginalization over nuisance parameters, while being broadly consistent with observed statistics of neural firing.

Since generative models can be made arbitrarily complex, it is likely that perceptual tasks exist in which the brain has to make crude approximations to perform inference; this could lead to a suboptimal model describing behavioral data better than the optimal one (Landy, Goutcher, Trommershauser, & Mamassian, 2007). However, we speculate that no task exists in which a suboptimal, *non-probabilistic* model outperforms an optimal, probabilistic one, because in realistic situations, taking into account uncertainty is a crucial element of good behavioral performance. This is a potential direction for future work. Another active field is the search for neural implementations of probabilistic optimal computation of the form of Eq. (23), both theoretically and experimentally.

## Acknowledgments

We are very grateful to Miguel Eckstein, Ronald van den Berg, Jeff Beck, Preeti Verghese, and Alex Pouget for useful discussions and comments, and to Jan Drugowitsch for making his VBLR code publicly available (http://www.bcs.rochester.edu/people/jdrugowitsch/code.html).

## References

Alais, D., & Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current Biology, 14*(3), 257–262.

Anastasio, T. J., Patton, P. E., & Belkacem-Boussaid, K. (2000). Using Bayes' rule to model multisensory enhancement in the superior colliculus. *Neural Computation, 12*(5), 1165–1187.

Anderson, C. (1994). Neurobiological computational systems. In *Computational intelligence imitating life* (pp. 213–222). New York: IEEE Press.

Baldassi, S., & Burr, D. C. (2000). Feature-based integration of orientation signals in visual search. *Vision Research, 40*, 1293–1300.

Baldassi, S., & Verghese, P. (2002). Comparing integration rules in visual search. *Journal of Vision, 2*(8), 559–570.

Barlow, H. B. (1969). Pattern recognition and the responses of sensory neurons. *Annals of the New York Academy of Sciences, 156*(2), 872–881.

Barlow, H. B. (1972). Single units and sensation: A neuron doctrine for perceptual psychology? *Perception, 1*(4), 371–394.

Battaglia, P. W., Jacobs, R. A., & Aslin, R. N. (2003). Bayesian integration of visual and auditory signals for spatial localization. *Journal of the Optical Society of America A – Optics Image Science and Vision, 20*(7), 1391–1397.

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.

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

Bradley, A., Skottun, B. C., Ohzawa, I., Sclar, G., & Freeman, R. D. (1987). Orientation and spatial frequency discrimination: Single cells and behavior. *Journal of Neurophysiology, 57*, 755–772.

Britten, K. H., Shadlen, M. N., Newsome, W. T., & Movshon, J. A. (1992). The analysis of visual motion: A comparison of neuronal and psychophysical performance. *Journal of Neuroscience, 12*(12), 4745–4765.

Brouwer, A.-M., & Knill, D. C. (2007). The role of memory in visually guided reaching. *Journal of Vision, 7*(5), 1–12.

Burgess, A. (1985). Visual signal detection III. On Bayesian use of prior knowledge and cross-correlation. *Journal of the Optical Society of America A. Optics and Image Science, 2*(9), 1498–1507.

Clark, J., & Yuille, A. L. (1990). *Data fusion for sensory information processing systems*. Norwell, MA: Kluwer.

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. New York: John Wiley & Sons.

Dean, A. F. (1981). The variability of discharge of simple cells in the cat striate cortex. *Experimental Brain Research, 44*, 437–440.

Deneve, S. (2008). Bayesian spiking neurons I: Inference. *Neural Computation, 20*(1), 91–117.

Eckstein, M. P. (1998). The lower visual search efficiency for conjunctions is due to noise and not serial attentional processing. *Psychological Science, 9*(2), 111–118.

Eckstein, M. P., Peterson, M. F., Pham, B. T., & Droll, J. A. (2009). Statistical decision theory to relate neurons to behavior in the study of covert visual attention. *Vision Research, 49*(10), 1097–1128.

Eckstein, M. P., Thomas, J. P., Palmer, J., & Shimozaki, S. S. (2000). A signal detection model predicts the effects of set size on visual search accuracy for feature, conjunction, triple conjunction, and disjunction displays. *Perception & Psychophysics, 62*(3), 425–451.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature, 415*(6870), 429–433.

Fiser, J., Berkes, P., Orban, G., & Lengyel, M. (2010). Statistically optimal perception and learning: From behavior to neural representations. *Trends in Cognitive Sciences, 14*(3), 119–130.

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.

Georgopoulos, A., Kalaska, J., Caminiti, R., & Massey, J. T. (1982). On the relations between the direction of two-dimensional arm movements and cell discharge in primate motor cortex. *Journal of Neuroscience, 2*(11), 1527–1537.

Gold, J. I., & Shadlen, M. N. (2001). Neural computations that underlie decisions about sensory stimuli. *Trends in Cognitive Sciences, 5*, 10–16.

Goldreich, D. (2007). A Bayesian perceptual model replicates the cutaneous rabbit and other tactile spatiotemporal illusions. *PLoS ONE, 2*(3).

Graham, N., Kramer, P., & Yager, D. (1987). Signal detection models for multidimensional stimuli: Probability distributions and combination rules. *Journal of Mathematical Psychology, 31*, 366–409.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Los Altos, CA: John Wiley & Sons.

Gu, Y., Angelaki, D. E., & DeAngelis, G. C. (2008). Neural correlates of multisensory cue integration in macaque MSTd. *Nature Neuroscience, 11*(10), 1201–1210.

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

Hoyer, P. O., & Hyvarinen, A. (2000). Independent component analysis applied to feature extraction from colour and stereo images. *Network, 11*(3), 191–210.

Jacobs, R. A. (1999). Optimal integration of texture and motion cues to depth. *Vision Research, 39*(21), 3621–3629.

Jazayeri, M., & Movshon, J. A. (2006). Optimal representation of sensory information by neural populations. *Nature Neuroscience, 9*(5), 690–696.

Johnson, K. O. (1980). Sensory discrimination: Neural processes preceding discrimination decision. *Journal of Neurophysiology, 43*, 1793–1815.

Kepecs, A., Uchida, N., Zariwala, H. A., & Mainen, Z. F. (2008). Neural correlates, computation and behavioural impact of decision confidence. *Nature, 455*, 227–233.

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

Kiani, R., & Shadlen, M. N. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. *Science, 324*, 759–764.

Knill, D. C. (1998a). Discrimination of planar surface slant from texture: Human and ideal observers compared. *Vision Research, 38*(11), 1683–1711.

Knill, D. C. (1998b). Surface orientation from texture: Ideal observers, generic observers and the information content of texture cues. *Vision Research, 38*, 1655–1682.

Knill, D. C., & Richards, W. (Eds.). (1996). *Perception as Bayesian inference*. New York: Cambridge University Press.

Knill, D. C., & Saunders, J. A. (2003). Do humans optimally integrate stereo and texture information for judgments of surface slant? *Vision Research, 43*(24), 2539–2558.

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.

Kording, K. P., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature, 427*(6971), 244–247.

Landy, M. S., Goutcher, R., Trommershauser, J., & Mamassian, P. (2007). Visual estimation under risk. *Journal of Vision*, 7(6), 4, 1–15.

Landy, M. S., Maloney, L. T., Johnston, E. B., & Young, M. (1995). Measurement and modeling of depth cue combination: In defense of weak fusion. *Vision Research, 35*(3), 389–412.

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

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

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

Ma, W. J., Zhou, X., Ross, L. A., Foxe, J. J., & Parra, L. (2009). Lip-reading aids word recognition most in moderate noise: A Bayesian explanation using high-dimensional feature space. *PLoS ONE, 4*(3), e4638.

MacKay, D. J. (2003). *Information theory, inference, and learning algorithms*. Cambridge, UK: Cambridge University Press.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Mamassian, P., & Landy, M. S. (1998). Observer biases in the 3D interpretation of line drawings. *Vision Research, 38*, 2817–2832.

Morgan, M. L., DeAngelis, G. C., & Angelaki, D. E. (2008). Multisensory integration in macaque visual cortex depends on cue reliability. *Neuron, 59*, 662–673.

Newsome, W. T., Britten, K. H., & Movshon, J. A. (1989). Neuronal correlates of a perceptual decision. *Nature, 341*(6237), 52–54.

Nolte, L. W., & Jaarsma, D. (1966). More on the detection of one of M orthogonal signals. *Journal of the Acoustical Society of America, 41*(2), 497–505.

Palmer, J., Verghese, P., & Pavel, M. (2000). The psychophysics of visual search. *Vision Research, 40*(10–12), 1227–1268.

Paradiso, M. A., Carney, T., & Freeman, R. D. (1989). Cortical processing of hyperacuity tasks. *Vision Research, 29*(2), 247–254.

Parker, A. J., & Newsome, W. T. (1998). Sense and the single neuron: Probing the physiology of perception. *Annual Review of Neuroscience, 21*, 227–277.

Pelli, D. G. (1985). Uncertainty explains many aspects of visual contrast detection and discrimination. *Journal of the Optical Society of America [A], 2*(9), 1508–1532.

Peterson, W. W., Birdsall, T. G., & Fox, W. C. (1954). The theory of signal detectability. *Transactions IRE Profession Group on Information Theory, PGIT-4*, 171–212.

Pouget, A., Dayan, P., & Zemel, R. S. (2003). Inference and computation with population codes. *Annual Review of Neuroscience, 26*, 381–410.

Ramsey, F. (1926). Truth and probability. In R. B. Braithwaite (Ed.), *The foundations of mathematics and other logical essays* (pp. 156–198). London: Kegan, Paul, Trench, Trubner & Co.

Reuschel, J., Drewing, K., Henriques, D. Y. P., Roesler, F., & Fiehler, K. (2010). Optimal integration of visual and proprioceptive movement information for the perception of trajectory geometry. *Experimental Brain Research, 201*, 853–862.

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

Seung, H., & Sompolinsky, H. (1993). Simple model for reading neuronal population codes. *Proceedings of National Academy of Sciences, USA, 90*, 10749–10753.

Shimozaki, S. S., Eckstein, M. P., & Abbey, C. K. (2003). Comparison of two weighted integration models for the cueing task: Linear and likelihood. *Journal of Vision, 3*(3), 209–229.

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

van Beers, R. J., Sittig, A. C., & Denier van der Gon, J. J. (1996). How humans combine simultaneous proprioceptive and visual position-information. *Experimental Brain Research, 111*(2), 253–261.

van Beers, R. J., Sittig, A. C., & Gon, J. J. (1999). Integration of proprioceptive and visual position-information: An experimentally supported model. *Journal of Neurophysiology, 81*(3), 1355–1364.

Verghese, P. (2001). Visual search and attention: A signal detection theory approach. *Neuron, 31*(4), 523–535.

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

Whiteley, L., & Sahani, M. (2008). Implicit knowledge of visual uncertainty guides decisions with asymmetric outcomes. *Journal of Vision, 8*(3), 1–15.

Wu, S., Nakahara, H., & Amari, S. (2001). Population coding with correlation and an unfaithful model. *Neural Computation, 13*(4), 775–797.

Yang, T., & Shadlen, M. N. (2007). Probabilistic reasoning by neurons. *Nature, 447*(7148), 1075–1080.

Zemel, R., Dayan, P., & Pouget, A. (1998). Probabilistic interpretation of population code. *Neural Computation, 10*, 403–430.