Trial-to-trial, uncertainty-based adjustment of decision boundaries in visual categorization

Ahmad T. Qamar[a,1], R. James Cotton[a,1], Ryan G. George[a,1], Jeffrey M. Beck[b], Eugenia Prezhdo[a], Allison Laudano[a], Andreas S. Tolias[a], and Wei Ji Ma[a,2,3]

[a]Department of Neuroscience, Baylor College of Medicine, Houston, TX 77030; and [b]Department of Brain and Cognitive Science, University of Rochester, Rochester, NY 14627

Edited by Barbara Anne Dosher, University of California, Irvine, CA, and approved October 15, 2013 (received for review November 13, 2012)

Categorization is a cornerstone of perception and cognition. Computationally, categorization amounts to applying decision boundaries in the space of stimulus features. We designed a visual categorization task in which optimal performance requires observers to incorporate trial-to-trial knowledge of the level of sensory uncertainty when setting their decision boundaries. We found that humans and monkeys did adjust their decision boundaries from trial to trial as the level of sensory noise varied, with some subjects performing near optimally. We constructed a neural network that implements uncertainty-based, near-optimal adjustment of decision boundaries. Divisive normalization emerges automatically as a key neural operation in this network. Our results offer an integrated computational and mechanistic framework for categorization under uncertainty.

Bayesian inference | vision | decision-making | optimality

Imagine a woman is approaching you from a distance and you are trying to determine whether or not she is the friend you are waiting for. Because of various sources of noise, your observations of her facial features, hair color, etc. will be uncertain. A sensible strategy would be to be more tolerant to deviations between your observations and your knowledge of your friend's looks when she is far away than when she is close by and your observations are less uncertain. In this categorization problem, you are determining whether the image of the approaching woman falls into the narrow category of images of your friend or the wide category of images of all other people. Categorization can be modeled as a process of applying one or more decision boundaries to a noisy measurement in a space of stimulus features (1–7). The example suggests that adjusting such decision boundaries based on the current level of sensory uncertainty might be a better strategy than using uncertainty-independent decision boundaries.

Previous studies have not addressed whether organisms adjust their decision boundaries from trial to trial according to the level of sensory uncertainty. Perceptual studies of categorization under sensory uncertainty have typically used category distributions for which the level of uncertainty was irrelevant for optimal behavior (2, 3, 6, 8). For example, in a classic task, observers categorize the direction of motion of a set of dots coherently moving to the left or to the right, in the presence of distractor dots moving in random directions (8). Regardless of the level of sensory noise corrupting the brain's measurement of the net motion direction, the optimal decision is simply to report whether this measurement was to the right or to the left. In other words, applying a fixed decision boundary to a scalar estimate is optimal in this task; no knowledge of uncertainty about motion direction is needed. In cognitive models of categorization, dynamic decision boundaries have been invoked to explain a broad range of phenomenona, including sequential effects (9, 10), context effects (11), and generalization (12). However, these studies limited themselves to fixed levels of sensory noise and were not able to demonstrate optimality of behavior. Thus, a dichotomy exists: perceptual models are often normative and describe behavior in tasks with variable sensory uncertainty but trivial category distributions, whereas cognitive models examine more complex forms of categorization but are typically nonnormative and ignore the role of sensory uncertainty.

Here, we attempt to connect these domains using a visual categorization task in which sensory noise is varied unpredictably from trial to trial. Our simple experimental design allows us to determine how observers should adjust their decision boundaries to achieve optimal performance; thus, our approach is normative. We found that humans and monkeys do adjust their decision boundaries from trial to trial according to sensory uncertainty. We also constructed a biologically inspired neural network model that can perform near-optimal, uncertainty-based adjustment of decision boundaries. Thus, we offer both a computational and a mechanistic account of brain function in a task in which trial-to-trial sensory uncertainty drives decision boundary dynamics.

Results

Task. Human and monkey observers categorized the orientation of a drifting grating. The two categories ($C = 1, 2$) were defined by Gaussian probability distributions with the same mean (defined as 0°) and different SDs, denoted by σ_1 and σ_2 for categories 1 and 2, respectively (Fig. 1 *A* and *B*). As in related tasks (13, 14), the overlap of these distributions introduces ambiguity: a given orientation can come from either category, and therefore, categorization performance cannot be perfect even in the absence of sensory noise. Observers were trained using high-contrast stimuli and trial-to-trial feedback. During testing, contrast was varied from trial to trial to manipulate sensory uncertainty. Human observers did not receive trial-to-trial feedback during testing.

Theory. The statistical structure of the task, also called the generative model, contains three variables (Fig. 1*C*): category

Significance

Categorization is an important part of perception and cognition. For example, an animal must successfully categorize a disturbing sound as being due to the wind or to a predator. Computationally, categorization amounts to applying decision boundaries to noisy stimulus measurements. Here, we examine how these decision boundaries change as the quality of the sensory evidence varies unpredictably from trial to trial. We show that both humans and monkeys adjust their decision boundaries from trial to trial, often near-optimally. We further show how a neural network can perform this computation near-optimally. Our results might lead to a better understanding of categorization.

Author contributions: R.J.C., A.S.T., and W.J.M. designed research; A.T.Q., R.J.C., R.G.G., J.M.B., E.P., A.L., A.S.T., and W.J.M. performed research; A.T.Q., R.G.G., E.P., and W.J.M. analyzed data; and W.J.M. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]A.Q., R.J.C., and R.G. contributed equally to this work.

[2]Present address: Center for Neural Science and Department of Psychology, New York University, New York, NY 10003.

[3]To whom correspondence should be addressed. E-mail: weijima@nyu.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1219756110/-/DCSupplemental.

C, stimulus orientation s, and measurement x. On each trial, C is drawn randomly and determines whether s is drawn from the narrow Gaussian, $p(s \mid C = 1)$ with SD σ_1 or the wide Gaussian, $p(s \mid C = 2)$ with SD σ_2. We assume that on each trial, the true orientation is corrupted by sensory noise to give rise to the observer's measurement of orientation, x. We denote by $p(x \mid s)$ the probability distribution over x for a given stimulus orientation s. We assume it to be Gaussian with mean s and SD σ (6). This SD is experimentally manipulated through contrast.

On a given trial, the observer uses the measurement x to infer category C. An optimal observer would do this by computing the posterior probability distribution over category, denoted $p(C \mid x)$, which indicates the degree of belief that the category was C, based on x. It is convenient to express the posterior in terms of the log posterior ratio, d, which is, using Bayes' rule, the sum of the log likelihood ratio and the log prior ratio

$$d = \log\frac{p(C=1 \mid x)}{p(C=2 \mid x)} = \log\frac{p(x \mid C=1)}{p(x \mid C=2)} + \log\frac{p_1}{1-p_1}, \qquad [1]$$

where p_1 is the observer's prior belief that $C = 1$. The absolute value of d is one possible measure of decision confidence. The difficulty in computing the category likelihood $p(x \mid C)$ is that the stimulus s that caused x is unknown. The optimal observer deals with this by multiplying, for every possible s separately, the probability that the observed x came from this s with the probability of this s under the hypothesized category, and finally summing over all s

$$p(x \mid C) = \int p(x \mid s)p(s \mid C)ds = \int L_x(s)p(s \mid C)ds. \qquad [2]$$

Here, we use the notation $L_x(s)$ to denote the likelihood of a stimulus value s based on a measurement x, $L_x(s) = p(x|s)$. The width of this function, also σ, measures sensory uncertainty (Fig. 1*E*). Thus, Eq. **2** reflects the combination of sensory information, $L_x(s)$, with category information, $p(s \mid C)$. An integral over a hidden variable, as in Eq. **2**, is known as marginalization and is a central operation in Bayesian computation (15–18). A straightforward calculation gives (*SI Text*)

$$d = k_1 - k_2 x^2, \qquad [3]$$

where $k_1 = \frac{1}{2}\log\frac{\sigma^2+\sigma_2^2}{\sigma^2+\sigma_1^2} + \log\frac{p_1}{1-p_1}$ and $k_2 = \frac{\sigma_2^2-\sigma_1^2}{2(\sigma^2+\sigma_1^2)(\sigma^2+\sigma_2^2)}$. The decision strategy that maximizes accuracy is the maximum-a-posteriori (MAP) read-out, i.e., to report the value of C for which $p(C \mid x)$ is larger. This rule is equivalent to reporting category 1 when d is positive, or in other words, when $|x| < \sqrt{\frac{k_1}{k_2}} \equiv k$. Thus, the optimal observer reports category 1 when the measurement lies within the interval from $-k$ to k and category 2 otherwise. Critically, the optimal boundary or criterion k depends on the sensory uncertainty σ: when evidence is uncertain, the optimal observer is more willing to attribute measurements far away from zero to category 1 (Fig. 1*D*). This effect, which reflects the intuition of the friend recognition example, is a direct consequence of the shape of the category distributions. We consider two variants of the optimal model: one in which $p_1 = 0.5$, reflecting the experimental statistics (which we call the Opt model), and one in which p_1 is a free parameter (Opt-P model).

The main alternative to the optimal model is one in which the observer uses a fixed decision boundary (Fixed model). Then, the decision rule is $|x| < k_0$, with k_0 a constant.

Of course, the optimal model is not the only possible model in which the observer takes into account uncertainty on a trial-to-trial basis, even when we restrict ourselves to decision rules of the form $|x|$ < function of σ. As a first step in exploring this model space, we test all linear functions of σ (we call this model Lin-σ): the observer uses the rule $|x| < k_0\left(1+\frac{\sigma}{\sigma_p}\right)$. We also test a model in which the observer applies a fixed boundary not to the measurement x but to the MAP estimate of the stimulus, $\frac{\sigma_p^2}{\sigma_p^2+\sigma^2}x$, obtained under a Gaussian prior with mean 0 and SD σ_p. The decision rule is then $|x| < k_0\left(1+\frac{\sigma^2}{\sigma_p^2}\right)$. We call this the Quad-$\sigma$ model (Fig. 1*E*). In all models except for the Fixed model, the observer takes into account the trial-to-trial level of sensory uncertainty; these models therefore describe probabilistic, but not necessarily optimal, computation (19).

In each model, we describe the relationship between noise variance σ^2 and contrast c (expressed as a proportion, not as a percentage) as a power law with a baseline, $\sigma^2(c) = (\alpha c)^{-\beta} + \gamma$, and include a lapse rate λ to account for random guesses and unintended responses (20). The Opt model has four free parameters (α, β, γ, and λ), Opt-P and Fixed each have one more (p_1 and k, respectively), and Lin-σ and Quad-σ each have two more (k and σ_p). Thus, Lin-σ and Quad-σ can be considered more flexible models than Opt and Opt-P. Models are summarized in Table S1.

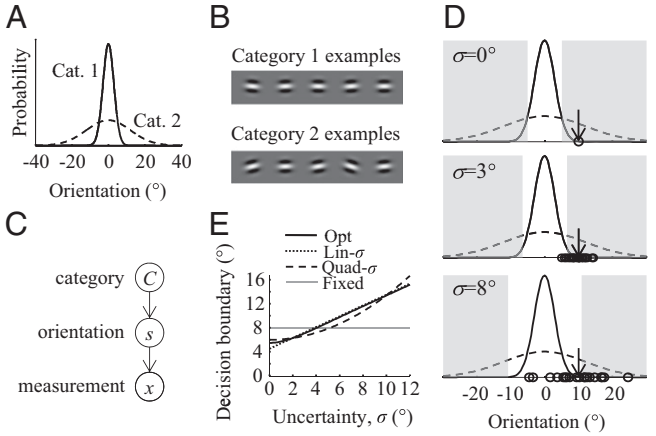

Fig. 1. Task and models. (*A*) Probability distributions over orientation for Categories 1 and 2. The distributions have the same mean (0°) but different SDs ($\sigma_1 = 3°$ and $\sigma_2 = 12°$; for Monkey L, $\sigma_2 = 15°$). (*B*) Sample stimuli drawn from each category. (*C*) Generative model of the task (see text). (*D*) A completely certain ($\sigma = 0°$) optimal observer would set the decision boundaries $\pm k$ at the intersection points of the two category distributions (black curves). The shaded areas indicate where the optimal observer would respond Category 2. When sensory noise is larger, not only will the measurements (open circles) be more variable for a given true orientation (arrow), but the optimal observer will also move the decision boundaries to larger values. In the experiment, noise levels are interleaved. (*E*) Decision boundary as a function of uncertainty level under four models: optimal, linear (example with $k_0 = 4.5°$ and $\sigma_p = 5°$), quadratic (example with $k_0 = 6°$ and $\sigma_p = 9°$), and Fixed (example with $k_0 = 8°$).

Behavioral Results. We first obtained maximum-likelihood estimates of the model parameters (Table S2). To compare models, we then computed both the marginal log likelihood (using the Laplace approximation) and the Akaike information criterion (AIC) for each model and each subject (*Materials and Methods*).

Our goal was to determine whether observers take into account sensory uncertainty in setting their decision boundaries. Thus, we are interested in whether or not the Fixed model accounts better for the data than all probabilistic models. It did not for any human subject or either monkey, according to either measure (Table 1 and Table S3). Moreover, the differences between the best probabilistic model and the Fixed model were large; to illustrate, Jeffreys (21) considered a marginal log likelihood difference of more than log(30) = 3.4 very strong evidence. Subjects differed in which probabilistic model described their data best, with Opt, Opt-P, Lin-σ, and Quad-σ each winning for at least

Table 1. Model comparison using log marginal likelihood for main experiment

Subject	Opt	Opt-P	Lin-σ	Quad-σ	Fixed	DIFF
Human 1	−1,479.0	−1,429.1	−1,422.9	−1,425.9	−1,452.0	−29.1
Human 2	−1,642.7	−1,646.2	−1,648.6	−1,660.6	−1,718.6	−75.9
Human 3	−891.7	−893.0	−897.6	−897.7	−919.6	−27.9
Human 4	−1,396.6	−1,383.5	−1,353.7	−1,364.8	−1,500.6	−146.8
Human 5	−1,280.6	−1,272.4	−1,241.3	−1,245.4	−1,318.0	−76.7
Human 6	−1,152.0	−1,131.0	−1,134.9	−1,135.6	−1,162.5	−31.5
Monkey A	−33,875.2	−32,862.9	−32,954.4	−32,762.6	−35,354.8	−2,592.2
Monkey L	−71,973.6	−71,787.5	−71,464.0	−71,501.7	−73,167.5	−1,703.5

Numbers are marginal log likelihoods obtained using the Laplace approximation. Shaded in green are the models whose values fall within log(30) of the value of the best model. The Fixed model is never among them. DIFF, difference between the Fixed model and the best probabilistic model.

one human subject, according to either method. The data of monkey A were best described by Quad-σ and those of monkey L best by Lin-σ. These results suggest that observers use different and sometimes suboptimal strategies to incorporate sensory uncertainty information on a trial-by-trial basis, but no observer ignores this information.

Fig. 2 shows three types of psychometric curves with corresponding model fits: proportion correct as a function of contrast (Fig. 2*A*), proportion of category 1 reports as a function of contrast and true category (Fig. 2*B*), and as a function of contrast and orientation (Fig. 2*C*). For each curve, the Fixed model provides a much worse fit than each of the probabilistic models (measured by RMSE), providing further indication that observers compute their decision boundaries in an uncertainty-dependent manner. In Fig. 2*C*, we observe that the curve widens as contrast decreases; because the Fixed model has a fixed decision boundary, it cannot account for this widening. Among the probabilistic models, the Lin-σ model fits best overall, in accordance with Table 1 and Table S3, but the differences are small. Although the Opt model provides a slightly worse fit, it should be kept in mind that in this model, the decision boundary is rigidly specified as a function of uncertainty σ: in contrast to the Lin-σ and Quad-σ models, the Opt model does not introduce any free parameters in the decision boundary function. In this light, the good fit of the Opt model is remarkable.

To agnostically estimate the uncertainty-dependent decision boundary humans use, we fitted an additional, flexible model, in which a separate boundary is fitted at each contrast level (*SI Text*). The decision rule is then $|x| < k_c$, where k_c is the boundary at contrast c. The resulting estimates of k_c are plotted in Fig. 2*D*. We find a significant effect of contrast on k_c [repeated-measures ANOVA: $F(5,25) = 12.4$, $P < 10^{-5}$]. The probabilistic models account for the trend in k_c much better than the Fixed model (in RMSE).

To confirm that the fitted values of sensory uncertainty σ in the probabilistic models are meaningful, we measured them in an independent experiment. The same six human observers performed a left/right orientation discrimination task (*Materials and Methods*) under the same contrasts and other experimental settings as in the categorization experiment. The estimates of σ obtained from the categorization task were strongly correlated with those obtained from the discrimination task (Fig. 2*E*; Pearson $r = 0.89$, $P < 10^{-3}$). Using the estimates of σ from the discrimination task instead of those obtained from the estimates of α, β, and γ in the flexible model, although worse, still produces reasonable probabilistic model fits to the decision boundary function (Fig. 2*F*). Together, these results are evidence that the fitted values of σ in the main experiment are meaningful.

To ascertain that our results did not depend on the choice of vertical as the central orientation, we repeated the experiment using 45° clockwise from vertical instead. Results were consistent with the main experiment (*SI Text* and Fig. S1).

Monkey summary statistics are shown in Fig. 3 and Fig. S2. Again, the Fixed model did not account well for the data. For

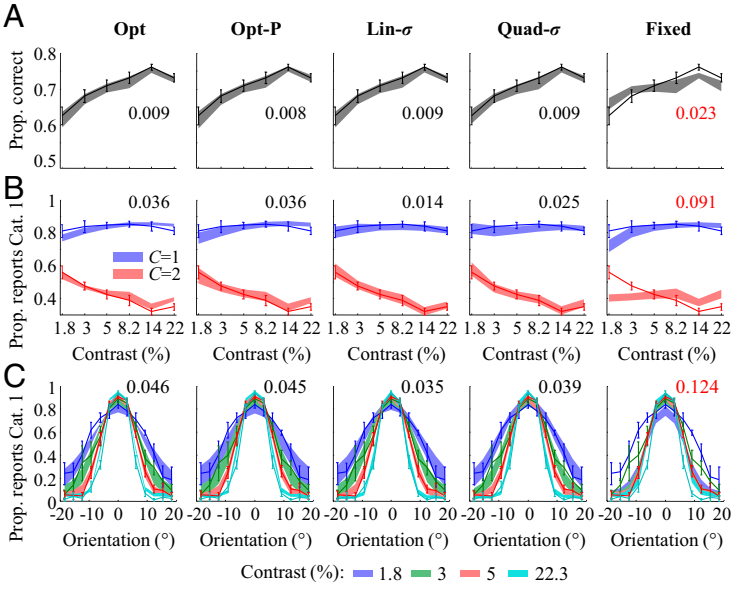

Fig. 2. (*A*) Proportion correct versus contrast, with model fits. Model fits often do not look smooth because they were computed based on the orientations actually presented in the experiment, and those were drawn randomly (see *SI Text*). Here and elsewhere, error bars and shaded areas indicate 1 SEM, and numbers indicate root mean squared error between data and model means (red is worst). (*B*) Proportion "Category 1" responses versus contrast, separated by true category, with model fits. (*C*) Proportion "Category 1" responses versus orientation and contrast, with model fits. For visibility, not all contrasts are plotted. (*D*) Decision boundaries fitted separately at each contrast level (error bars: data; shaded areas: models). (*E*) Estimates of sensory uncertainty σ estimated from the categorization task against ones obtained from an independent orientation discrimination task; color labels contrast. (*F*) Decision boundaries predicted by three probabilistic models based on the uncertainty estimates from the discrimination task.

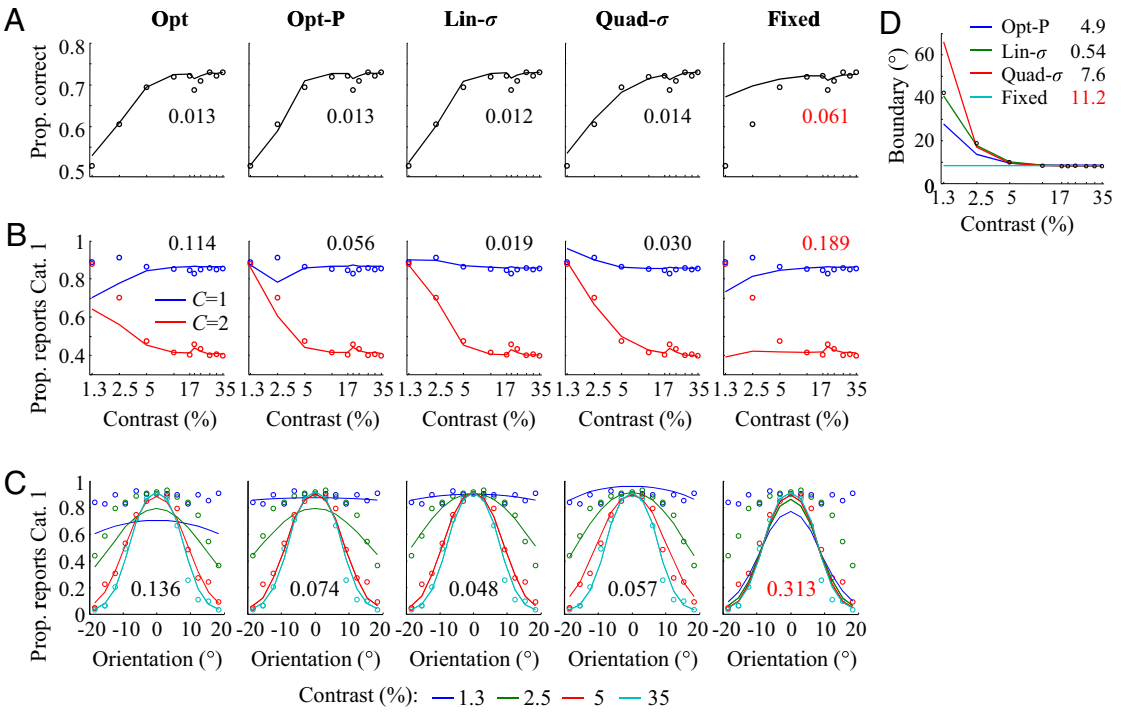

Fig. 3. As in Fig. 2, but for Monkey L.

monkey L, Lin-σ provided the best fit (Fig. 3), and for monkey A, Quad-σ (Fig. S2). Taken together, our results indicate that both humans and monkeys adjust their decision boundaries from trial to trial based on sensory uncertainty, but use diverse and sometimes suboptimal strategies to do so.

Neural Network Results. Our next goal is to provide a proof of concept that a biologically plausible neural network can adjust the decision boundary from trial to trial based on sensory uncertainty. We do this for the optimal model, because in this model, on each trial, the observer does not simply evaluate the decision rule, but also has a representation of the probability that the decision is correct. Although we did not explore it experimentally, knowing this probability is important for combining category information with information about the consequences of actions. For example, an animal might categorize a sound as being caused by the wind, but if the posterior probability of wind is 55% and that of predator is 45%, the best course of action would still be to run. Thus, for this animal, it is important to know whether the posterior probability of wind is 55% or 99%. In many forms of decision-making, observers must possess knowledge of posterior probability to maximize reward (3, 22–24). Although such a posterior probability is naturally given by the optimal model, it might also be possible to associate (suboptimal) posterior probabilities with the decision rules in the Lin-σ and Quad-σ models. However, this is not obvious and we leave it for future work.

To obtain a neural code, we replace the scalar observation x by the vector of spike counts in a population of orientation-tuned neurons, denoted $\mathbf{r}$ (Fig. 4*A*). We assume variability in the exponential family with linear sufficient statistics (Poisson-like variability) (25), which is consistent with the physiology of primary visual cortex (26, 27). Then, the distribution of $\mathbf{r}$ across trials for a given stimulus s can be described by

$$p(\mathbf{r} \mid s,g) = \varphi(\mathbf{r},g)e^{\mathbf{h}(s)\cdot\mathbf{r}}, \qquad [4]$$

where g denotes the gain (mean amplitude) of the population, which is affected by contrast, and φ is an arbitrary function. Using the framework of probabilistic population coding (25), all available information about the stimulus is contained in the neural likelihood function, $L_{\mathbf{r}}(s) = p(\mathbf{r} \mid s) = \int p(\mathbf{r} \mid s,g)p(g)dg = \Phi(\mathbf{r})e^{\mathbf{h}(s)\cdot\mathbf{r}}$ (Fig. 4*B*), where Φ is easily expressed in terms of φ. We assume that $\mathbf{h}(s)$ is a quadratic function of s and thus can be written as $\mathbf{h}(s) = -\frac{1}{2}s^2\mathbf{a} + s\mathbf{b}$, where $\mathbf{a}$ and $\mathbf{b}$ are constant vectors, so that the likelihood $L_{\mathbf{r}}(s)$ is an (unnormalized) Gaussian. The mean $\frac{\mathbf{b}\cdot\mathbf{r}}{\mathbf{a}\cdot\mathbf{r}}$ and variance $\frac{1}{\mathbf{a}\cdot\mathbf{r}}$ of this likelihood function correspond to x and σ^2 in the behavioral model, respectively. In the special case of independent Poisson variability with Gaussian tuning curves, $\mathbf{a}$ is a vector whose entries are equal to the inverse squared widths of the neurons' tuning curves, and $\frac{\mathbf{b}\cdot\mathbf{r}}{\mathbf{a}\cdot\mathbf{r}}$ is the center-of-mass decoder (28), the real-line analog of the population vector decoder (29) (*SI Text*). The log likelihood ratio over category can then be found from Eq. **3** by substituting the neural quantities for x and σ^2

$$d = \log\frac{p(\mathbf{r} \mid C=1)}{p(\mathbf{r} \mid C=2)} = \frac{1}{2}\log\frac{1+\sigma_2^2\mathbf{a}\cdot\mathbf{r}}{1+\sigma_1^2\mathbf{a}\cdot\mathbf{r}} - \frac{(\sigma_2^2-\sigma_1^2)(\mathbf{b}\cdot\mathbf{r})^2}{2(1+\sigma_1^2\mathbf{a}\cdot\mathbf{r})(1+\sigma_2^2\mathbf{a}\cdot\mathbf{r})}. \qquad [5]$$

Compared with Eq. **3**, we have left out the log prior ratio; this is a constant shift and therefore easily implemented. A key aspect of Eq. **5** is that the log posterior ratio is a nonlinear function of $\mathbf{r}$, in line with probabilistic population code implementations of other Bayesian computations that require marginalization (17, 18).

We propose that categorization is performed by a feedforward neural network that is Poisson-like both in input and output (Fig. 4*C*). For concreteness, one could think of the input layer as primary visual cortex, encoding orientation, and of the output layer as prefrontal cortex, encoding category; however, the computation does not depend on these labels. The assumption that category is encoded by a Poisson-like output population $\mathbf{z}$ is supported by extant physiological findings in decision-making areas (30). The problem is then to find the mapping from $\mathbf{r}$ to $\mathbf{z}$ such that $\mathbf{z}$ encodes not just category, but the optimal likelihood of category. Poisson-like variability in the output can be described by a probability distribution analogous to Eq. **4**, namely $p(\mathbf{z} \mid C, g_{\mathbf{z}}) = \varphi_{\mathbf{z}}(\mathbf{z},g_{\mathbf{z}})e^{\mathbf{H}(C)\cdot\mathbf{z}}$.

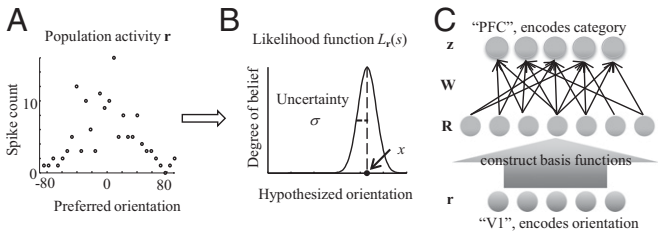

Fig. 4. Network design for uncertainty-dependent categorization. (*A*) In probabilistic population coding, a population of neurons selective for orientation (single-trial activity shown on left) encodes a likelihood function over orientation on each individual trial (right). (*B*). The peak value of the likelihood function is the stimulus measurement (x), and its width represents sensory uncertainty (σ). The higher the gain (mean amplitude) of the population, the lower sensory uncertainty. (*C*) The network is feedforward and consists of three layers. The input layer represents orientation-tuned neurons. The intermediate layer computes a set of nonlinear basis functions of the activities in the first layer. The output layer computes linear combinations of these basis functions; its goal is to veridically encode the posterior over category on each trial, regardless of the level of sensory uncertainty.

The log likelihood ratio over C encoded in $\mathbf{z}$ is then $\log\frac{p(\mathbf{z}|C=1)}{p(\mathbf{z}|C=2)} = (\mathbf{H}(C=1)-\mathbf{H}(C=2))\cdot\mathbf{z}$, which we write shorthand as $\Delta\mathbf{H}\cdot\mathbf{z}$. To encode the optimal likelihood over category, this linear combination of output activity must equal the right side of Eq. **5**.

We now approximate the optimal log likelihood ratio d by constructing a neural network consisting of a basis of nonlinear functions of input activity that are linearly combined to produce $\mathbf{z}$ (Fig. 4*C*). Our goal is to find a minimal and biologically plausible set of basis functions that can achieve near optimality. We allow operations that have been found in electrophysiology: linear, quadratic (i.e., multiplicative) (31–33), and divisive normalization (34, 35). The second term in Eq. **5** can already be constructed using these operations; we now study whether the first term can be approximated using functions of the same general form. Therefore, we consider a basis of functions that are quadratic in neural activity in the numerator and denominator: $\mathbf{R}^{\mathrm{QDN}} = \left[\frac{r_i r_j}{1+\mathbf{v}_2\cdot\mathbf{r}+\mathbf{r}^{\mathrm{T}}\mathbf{V}\mathbf{r}}\right]$, where QDN stands for "quadratic and divisive normalization." The approximate decision variable is a linear combination of these basis functions.

We simulated independent Poisson input with six values of gain. The network was trained using stochastic gradient descent (*SI Text*). After training, the network closely approximated the optimal posterior over category across a range of gains (Fig. 5*A*) and accurately reproduced the proportion reports of category 1 as a function of orientation of the optimal observer (Fig. 5*B*). Shannon information loss (*Materials and Methods*) was less than 0.5%. By comparison, we were unable to veridically approximate the posterior (Fig. 5*A*) or the psychometric curves (Fig. 5*B*) using a linear (LIN) network and a quadratic (QUAD) network. These networks had information losses of 42% and 21%, respectively, suggesting that the crucial operation for this categorization task is a quadratic operation combined with a divisive normalization that is itself also quadratic. We visualized the ability of the QDN network to approximate a highly nonlinear decision surface (Fig. S6).

Discussion

We showed in a simple but nontrivial categorization task that both humans and monkeys adjust their decision boundaries according to the current level of sensory uncertainty. Despite its simplicity, the task contains both ambiguity and variable sensory uncertainty—essential elements of natural forms of categorization. Evidence that organisms possess and use trial-to-trial knowledge of sensory uncertainty is abundant in perception (18, 36, 37), but rare in tasks with nontrivial stimulus categories (38, 39). It might be worth studying the effects of sensory uncertainty on decision boundaries in more complex or more cognitive forms of categorization.

We found that some observers' behavior is better described by simple suboptimal (but still probabilistic) decision rules than by the optimal rule. This finding raises the interesting question of whether these suboptimal rules are still associated with posterior distributions over category (and thus with a marginalization operation), but ones based on incorrect assumptions about the generative model.

We constructed a simple neural network that not only performs near-optimal categorization under varying levels of sensory noise, but also returns a good approximation of posterior probability. Divisive normalization, found in many species and brain areas, emerges automatically from our construction. This provides a normative justification for the operation (for other justifications, see ref. 35). In our framework, divisive normalization has an intuitive origin: it arises from the fact that in a Poisson-like

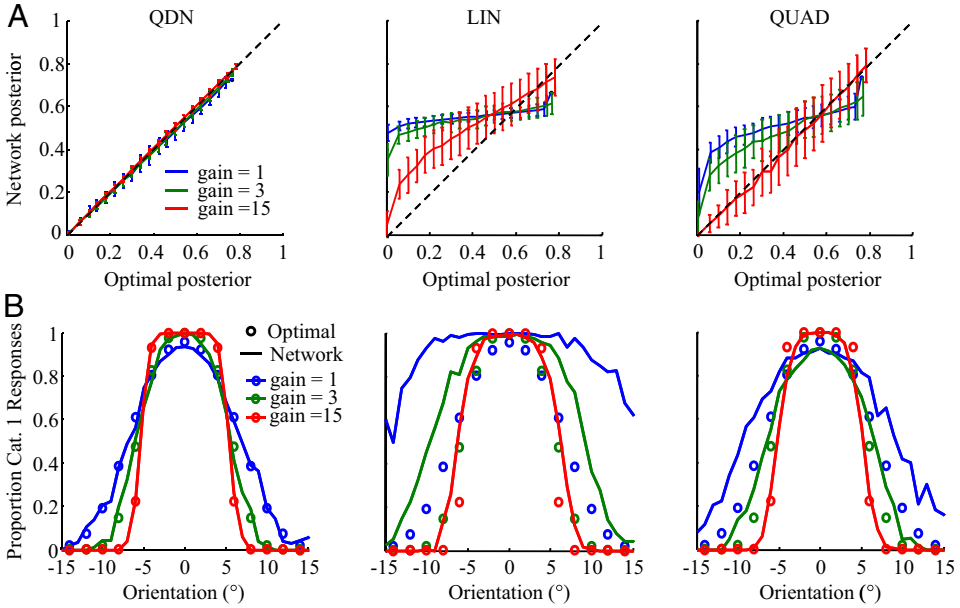

Fig. 5. Network results. (*A*) The posterior probability of Category 1 as decoded from the network versus the optimal posterior (binned), at different levels of input gain (sensory uncertainty), for a QDN network (trained for 10,000 iterations), a linear network (trained for 100,000 iterations), and a quadratic network (trained for 100,000 iterations). Error bars are SDs. Network information loss was 0.39%, 41.7%, and 21.0%, respectively; we verified that these values were close to the asymptotic ones. Note that the optimal posterior is bounded from above (see SI Text). (*B*) Proportion of Category 1 responses as a function of orientation for the optimal observer (dots) and the three networks (lines). The QDN network is the only of the three networks that performs well at each level of sensory uncertainty.

probabilistic population code, the amplitude of the population pattern of activity encodes certainty. As a result, the stimulus estimate, x, must be computed using an amplitude-invariant operation. A representative example is the center-of-mass (or population vector) decoder, which computes the weighted sum of the neurons' preferred orientations, with weights equal to the neurons' spike counts; because the sum of the spike counts appears in the denominator, the center-of-mass decoder automatically contains a divisive normalization. Bayesian decision-making is performed by applying an operation to x and σ^2. In the present task, the decision variable is a quadratic function of x (Eq. **3**), explaining why quadratic operations and quadratic divisive normalization appear in the neural implementation (Eq. **5**).

Quadratic operations and divisive normalization have proven crucial for implementing a number of seemingly disparate forms of Bayesian inference (19, 20). At first sight, these frequent appearances might be surprising, because in the probabilistic population code implementation of cue combination, neither quadratic operations nor divisive normalization are needed (27). However, this is because cue combination is special in that optimal inference only involves the ratio x/σ^2, in which the divisive normalization is cancelled out to produce an operation linear in neural activity. However, this type of cancellation is the exception rather than the rule. In virtually every other form of Bayesian decision-making, the optimal decision variable will involve quadratic operations and divisive normalization when implemented with probabilistic population codes.

We make several predictions for physiology. First, we predict that populations of orientation-sensitive neurons in early visual cortex not only encode an estimate of orientation but also trial-to-trial information about sensory uncertainty and that this encoded sensory uncertainty correlates with the animal's decision boundary as obtained from its behavior in the categorization task. Second, we predict that the activity of category-sensitive neurons, which might be found in prefrontal cortex (2) or lateral intraparietal cortex (LIP) (1, 5), is linearly related to the logarithm of the posterior probability ratio over category. This prediction is consistent with patterns found in LIP, where category probability can be reconstructed from a logistic mapping on neural activity (33). Third, we predict that sensory uncertainty decoded on each trial from early visual areas is propagated (along with the best estimate of orientation) to categorization neurons, so that including this decoded uncertainty as an explanatory factor should help to predict the activity of those neurons. Finally, we predict that if divisive normalization is selectively removed from a neural circuit involved in computing category, then the observer will become severely suboptimal in a way predicted by the QUAD network (Fig. 5). These predictions illustrate how a probabilistic, computationally [in the sense of Marr (40)] driven approach to categorization can guide the generation of hypotheses about neural mechanisms.

Materials and Methods

Details of all methods are provided in *SI Text*. Categories were defined by normal distributions with means 0° and SDs $\sigma_1 = 3°$ and $\sigma_2 = 12°$ (except for monkey L, for whom $\sigma_2 = 15°$). On each trial, a category was selected with equal probability. The stimulus was a drifting Gabor (50 ms for humans; 500 ms for monkeys) whose orientation s was drawn from the distribution of the selected category. Contrast was varied randomly from trial to trial. For monkeys, through training, the narrow distributon was associated with a red target and the wide distribution with a green target; the monkey reported category through a saccade to the red or the green target. Humans responded using a key press. Stimuli were delivered using Psychophysics Toolbox for Matlab (Mathworks). Models were fitted using maximum-likelihood estimation, implemented through a conjugate gradient method. Bayes' factors were computed using the Laplace approximation. Neural networks were trained using a stochastic gradient descent on the Kullback-Leibler divergence between the true and the approximated posterior.

ACKNOWLEDGMENTS. This work was supported by National Science Foundation Grant IIS-1132009 (Collaborative Research in Computational Neuroscience) (to W.J.M. and A.S.T.).

1. Freedman DJ, Assad JA (2006) Experience-dependent representation of visual categories in parietal cortex. *Nature* 443(7107):85–88.
2. Freedman DJ, Riesenhuber M, Poggio T, Miller EK (2001) Categorical representation of visual stimuli in the primate prefrontal cortex. *Science* 291(5502):312–316.
3. Kepecs A, Uchida N, Zariwala HA, Mainen ZF (2008) Neural correlates, computation and behavioural impact of decision confidence. *Nature* 455(7210):227–231.
4. Ashby FG, Lee WW (1991) Predicting similarity and categorization from identification. *J Exp Psychol Gen* 120(2):150–172.
5. Swaminathan SK, Freedman DJ (2012) Preferential encoding of visual categories in parietal cortex compared with prefrontal cortex. *Nat Neurosci* 15(2):315–320.
6. Green DM, Swets JA (1966) *Signal Detection Theory and Psychophysics* (John Wiley & Sons, Los Altos, CA).
7. Ashby FG, Maddox WT (2005) Human category learning. *Annu Rev Psychol* 56: 149–178.
8. Newsome WT, Britten KH, Movshon JA (1989) Neuronal correlates of a perceptual decision. *Nature* 341(6237):52–54.
9. Stewart N, Brown GD, Chater N (2002) Sequence effects in categorization of simple perceptual stimuli. *J Exp Psychol Learn Mem Cogn* 28(1):3–11.
10. Petrov AA, Anderson JR (2005) The dynamics of scaling: A memory-based anchor model of category rating and absolute identification. *Psychol Rev* 112(2):383–416.
11. Lewandowsky S, Kalish M, Ngang SK (2002) Simplified learning in complex situations: Knowledge partitioning in function learning. *J Exp Psychol Gen* 131(2):163–193.
12. Lamberts K (1994) Flexible tuning of similarity in exemplar-based categorization. *J Exp Psychol Learn Mem Cogn* 20(5):1003–1021.
13. Sanborn AN, Griffiths TL, Shiffrin RM (2010) Uncovering mental representations with Markov chain Monte Carlo. *Cognit Psychol* 60(2):63–106.
14. Liu Z, Knill DC, Kersten D (1995) Object classification for human and ideal observers. *Vision Res* 35(4):549–568.
15. Kersten D, Mamassian P, Yuille A (2004) Object perception as Bayesian inference. *Annu Rev Psychol* 55:271–304.
16. Knill DC (2003) Mixture models and the probabilistic structure of depth cues. *Vision Res* 43(7):831–854.
17. Ma WJ, Navalpakkam V, Beck JM, Berg Rv, Pouget A (2011) Behavior and neural basis of near-optimal visual search. *Nat Neurosci* 14(6):783–790.
18. Beck JM, Latham PE, Pouget A (2011) Marginalization in neural circuits with divisive normalization. *J Neurosci* 31(43):15310–15319.
19. Ma WJ (2012) Organizing probabilistic models of perception. *Trends Cogn Sci* 16(10): 511–518.
20. Wichmann FA, Hill NJ (2001) The psychometric function: I. Fitting, sampling, and goodness of fit. *Percept Psychophys* 63(8):1293–1313.
21. Jeffreys H (1961) *The Theory of Probability* (Oxford Univ Press, Oxford), 3rd Ed, p 470.
22. Whiteley L, Sahani M (2008) Implicit knowledge of visual uncertainty guides decisions with asymmetric outcomes. *J Vis* 8(3):1–15.
23. Maiworm M, König P, Röder B (2011) Integrative processing of perception and reward in an auditory localization paradigm. *Exp Psychol* 58(3):217–226.
24. Kiani R, Shadlen MN (2009) Representation of confidence associated with a decision by neurons in the parietal cortex. *Science* 324(5928):759–764.
25. Ma WJ, Beck JM, Latham PE, Pouget A (2006) Bayesian inference with probabilistic population codes. *Nat Neurosci* 9(11):1432–1438.
26. Graf ABA, Kohn A, Jazayeri M, Movshon JA (2011) Decoding the activity of neuronal populations in macaque primary visual cortex. *Nat Neurosci* 14(2):239–245.
27. Berens P, et al. (2012) A fast and simple population code for orientation in primate V1. *J Neurosci* 32(31):10618–10626.
28. Ma WJ, Pouget A (2009) *Population Coding: Theoretic Aspects. Encyclopedia of Neuroscience* (Elsevier, New York), Vol 7, pp 749–755.
29. Georgopoulos AP, Kalaska JF, Caminiti R, Massey JT (1982) On the relations between the direction of two-dimensional arm movements and cell discharge in primate motor cortex. *J Neurosci* 2(11):1527–1537.
30. Yang T, Shadlen MN (2007) Probabilistic reasoning by neurons. *Nature* 447(7148): 1075–1080.
31. Trotter Y, Celebrini S (1999) Gaze direction controls response gain in primary visual-cortex neurons. *Nature* 398(6724):239–242.
32. Galletti C, Battaglini PP (1989) Gaze-dependent visual neurons in area V3A of monkey prestriate cortex. *J Neurosci* 9(4):1112–1125.
33. Andersen RA, Essick GK, Siegel RM (1985) Encoding of spatial location by posterior parietal neurons. *Science* 230(4724):456–458.
34. Heeger DJ (1992) Normalization of cell responses in cat striate cortex. *Vis Neurosci* 9(2): 181–197.
35. Carandini M, Heeger DJ (2012) Normalization as a canonical neural computation. *Nat Rev Neurosci* 13(1):51–62.
36. Knill DC, Richards W, eds (1996) *Perception as Bayesian Inference* (Cambridge Univ Press, New York).
37. Trommershauser J, Kording K, Landy MS, eds (2011) *Sensory Cue Integration* (Oxford Univ Press, New York).
38. van den Berg R, Vogel M, Josic K, Ma WJ (2012) Optimal inference of sameness. *Proc Natl Acad Sci USA* 109(8):3178–3183.
39. Keshvari S, van den Berg R, Ma WJ (2012) Probabilistic computation in human perception under variability in encoding precision. *PLoS ONE* 7(6):e40216.
40. Marr D (1982) *Vision* (MIT Press, Cambridge, MA).

Supporting Information

Qamar et al. 10.1073/pnas.1219756110

SI Text

Human Psychophysics: Main Experiment. ***Stimuli.*** The stimulus was a drifting Gabor whose orientation s was drawn from one of two category distributions. On each trial, category 1 or category 2 was selected with equal probability. Categories distributions were normal with means 0° (horizontal, drifting to the right) and SDs $\sigma_1 = 3°$ and $\sigma_2 = 12°$, respectively (Fig. 1*A*). During training, the Gabor had 100% contrast. During testing, the contrast of the Gabor was 1.8%, 3.0%, 5.0%, 8.2%, 13.5%, or 22.3%. Stimuli were delivered using Psychophysics Toolbox for Matlab (Mathworks).

Procedure. Six human subjects participated (one female). Each subject completed five sessions, each consisting of 816 trials, organized as follows: 72 training, 216 testing, 48 training, 216 testing, 48 training, and 216 testing. The last two training blocks served to refresh observers' memories of the category distributions. In total, each subject completed 3,240 testing trials, equally divided among six contrast levels, for a total of 540 trials per contrast level. Contrast was chosen randomly on each trial. Exemplars of stimuli in each category were shown at the beginning of each session. A trial proceeded as follows (Fig. 1*C*). Subjects fixated on a central cross. The Gabor appeared at fixation for 300 ms during training and for 50 ms during testing. Immediately afterward, subjects indicated through a key press whether they believed the stimulus belonged to category 1 or category 2. During training, the fixation turned green if the response was correct and red if it was incorrect. During testing, no such feedback was given. After each block, the total score on that block was shown.

Human Psychophysics: Control Experiment. The control experiment was identical to the main experiment except for the following differences.

Stimuli. Stimuli were generated as in the main experiment but then rotated clockwise by 45°. An interrupted black diagonal line at the mean orientation was shown continuously to provide a reference. During testing, stimulus contrast could take values 1.1%, 1.8%, 3.0%, 5.0%, 8.2%, 13.5%, 22.3%, or 36.8%.

Procedure. Six human subjects participated (five females). Each subject completed five sessions, each consisting of 816 trials, organized as follows: 72 practice, 288 testing, 72 practice, and 288 testing. In total, each subject completed 2,880 testing trials, equally divided among eight contrast levels, for a total of 360 trials per contrast level.

Monkey Psychophysics. Monkeys engaged in a similar task to humans. The Gaussian category distributions (Fig. 1*A*) had a mean of vertical (grating drifting to the right) and widths $\sigma_1 = 3°$ and $\sigma_2 = 12°$ for monkey A and $\sigma_1 = 3°$ and $\sigma_2 = 15°$ for monkey L. Contrast was 1%, 2%, 3%, 5%, 8%, 10%, 20%, 35%, 50%, 70%, or 100% for monkey A and 1.25%, 2.5%, 5%, 10%, 15%, 17%, 20%, 25%, 30%, or 35% for monkey L. Monkey A completed 100,267 trials. Monkey L completed 184,838 trials.

A trial proceeded as follows. A fixation point appeared, and the monkey was required to fixate on it for 300 ms. A drifting grating then appeared for 500 ms, after which the monkey could select a stimulus category. Through training, the narrow distribution was associated with a red target and the wide distribution with a green target. The targets only appeared after the stimulus period, and the locations of the red and green targets were randomized between left and right. The monkey reported category through a saccade to the red or the green target. The monkey received a juice reward for each correct categorization response. Eye position was tracked using a custom-built field-programmable gate-array-based optical eye tracker running at 250 Hz. Stimulus and reward were controlled by a custom state system running LabView (National Instruments). Visual stimulation was delivered through a separate computer running Psychophysics Toolbox for Matlab (Mathworks).

Derivation of the Optimal Decision Rule. Starting from Eq. **2**, we substitute the expressions for the noise distribution and the category-conditioned stimulus distribution (with C equal to 1 or 2) and evaluate the integral:

$$p(x \mid C) = \int \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-s)^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi\sigma_C^2}} e^{-\frac{s^2}{2\sigma_C^2}} ds = \frac{1}{\sqrt{2\pi\left(\sigma^2+\sigma_C^2\right)}} e^{-\frac{x^2}{2\left(\sigma^2+\sigma_C^2\right)}}. \quad \textbf{[S1]}$$

Substituting Eq. **S1** in Eq. **1**, we find

$$d = \frac{1}{2}\log\frac{\sigma^2+\sigma_2^2}{\sigma^2+\sigma_1^2} - \frac{\sigma_2^2-\sigma_1^2}{2\left(\sigma^2+\sigma_1^2\right)\left(\sigma^2+\sigma_2^2\right)}x^2 + \log\frac{p_1}{1-p_1}, \quad \textbf{[S2]}$$

which is Eq. **3**. Because x^2 is nonnegative, d is bounded from above by k_1, which in turn is a decreasing function of σ. Therefore, the posterior probability of category 1 is bounded from above by $p(C=1 \mid x=0, \sigma=0) = \frac{1}{1+e^{-k_1}} = \frac{\sigma_2}{\sigma_1+\sigma_2}$. The decision rule is $d > 0$, which translates to $|x| < \sqrt{\frac{k_1}{k_2}} \equiv k$ in the main text.

List of Models. The decision rules and parameters sets of all models tested are listed in Table S1.

Response Probability. All model fits and comparisons are based on the probability that an observer reports category 1 for a given stimulus s and given uncertainty level σ. Recall that the decision rule is of the form $|x| < k(\sigma)$, where $k(\sigma)$ is some function of σ (as given by Table S1). Then, the probability that the observer reports category 1 for given s is straightforwardly computed to be

$$p(\hat{C}=1 \mid s) = \frac{1}{2}\left(\operatorname{erf}\frac{s+k(\sigma)}{\sigma\sqrt{2}} - \operatorname{erf}\frac{s-k(\sigma)}{\sigma\sqrt{2}}\right), \quad \textbf{[S3]}$$

where erf denotes the error function. In other words, the psychometric curve as a function of s at a given contrast is predicted to be a difference of two cumulative normal distributions.

Model Fitting. For a given model, we denote its set of parameters collectively by a vector **θ**. We aimed to find the parameter combination **θ** that maximized the parameter likelihood function. The parameter likelihood function is the probability of all of a single subject's responses given the presented stimuli and the parameters. Assuming conditional independence between trials, the log of the parameter likelihood function is

$$\begin{aligned} LL(\boldsymbol{\theta}; \text{model}) &= \log p(\text{data} \mid \boldsymbol{\theta}, \text{model}) \\ &= \log \prod_{i=1}^{N_{\text{trials}}} p\left(\hat{C}_i \mid s_i, c_i, \boldsymbol{\theta}\right). \\ &= \sum_{i=1}^{N_{\text{trials}}} \log p\left(\hat{C}_i \mid s_i, c_i, \boldsymbol{\theta}\right), \end{aligned}$$

where the product and the sum are over all of a single subject's trials, and s_i, c_i, and $\hat{C}_i$ are the orientation, contrast, and subject's category response on the ith trial, respectively.

We implemented the optimization of the log likelihood function using the Matlab program minimize.m (Carl Rasmussen, www.gaussianprocess.org/gpml/code/matlab). This software is based on a conjugate gradient algorithm and requires expressions for the first partial derivatives of the log likelihood function, which can straightforwardly be calculated in our models. We typically performed an initial stage with 1,000 randomly chosen initial parameter combinations and a maximum of 15 line searches for each, followed by a second stage where we took the 50 best parameter combinations found in the first stage and used them as initial conditions for a maximum of 1,000 line searches each. Of the 50 resulting parameter combinations, we took the one with the highest likelihood. We confirmed the results of the optimization using a custom-built genetic algorithm with a population size of 800, one child per parent, a 50% survival rate (including parents), and 650 generations. Although based on different principles, this algorithm produced maximum log likelihood values that were typically within one point from those obtained using minimize.m. We are therefore reasonably confident that we found the global maxima in parameter space.

Maximum-likelihood estimates of parameters in the five models are given in Table S2.

Model Comparison. Making use of the parameter likelihood function, we applied Bayesian model comparison, also called Bayes' factors (1, 2), to compare the goodness of fit of models. This method involves calculating the probability of the subject's responses under a model given the presented stimuli on individual trials by integrating the parameter likelihood over the parameters of the model

$$p(\text{data} \mid \text{model}) = \int p(\text{data} \mid \boldsymbol{\theta}, \text{model})\, p(\boldsymbol{\theta} \mid \text{model}) d\boldsymbol{\theta}.$$

The result is also called the marginal likelihood of the model. We assumed that each parameter θ_i takes values on an interval of size $R(\theta_i)$, and that the prior distribution $p(\boldsymbol{\theta} \mid \text{model})$ factorizes over parameters and is for each parameter uniform on its interval. Thus, $p(\boldsymbol{\theta} \mid \text{model}) = \prod_{i=1}^{\dim\boldsymbol{\theta}} \frac{1}{R(\theta_i)}$. Moreover, we used Laplace's approximation to compute the integral (2)

$$\begin{aligned}
\log p(\text{data} \mid \text{model}) &= \log \int p(\text{data} \mid \boldsymbol{\theta}, \text{model})\, p(\boldsymbol{\theta} \mid \text{model}) d\boldsymbol{\theta} \\
&= \log\left(\prod_{i=1}^{\dim\boldsymbol{\theta}} \frac{1}{R(\theta_i)} \right) + \log \int p(\text{data} \mid \boldsymbol{\theta}, \text{model}) d\boldsymbol{\theta} \\
&= \log\left(\prod_{i=1}^{\dim\boldsymbol{\theta}} \frac{1}{R(\theta_i)} \right) + \log \int e^{LL(\boldsymbol{\theta};\text{model})} d\boldsymbol{\theta}. \\
&\approx \log\left(\prod_{i=1}^{\dim\boldsymbol{\theta}} \frac{1}{R(\theta_i)} \right) + LL(\boldsymbol{\theta}^*; \text{model}) + \log \sqrt{\det \frac{2\pi}{H(\boldsymbol{\theta}^*)}},
\end{aligned}$$

where $\boldsymbol{\theta}^*$ is the maximum-likelihood parameter set and $H(\boldsymbol{\theta}^*)$ is the Hessian (matrix of second derivatives) of $-LL$ evaluated at $\boldsymbol{\theta}^*$. We then compared the approximated values of the log marginal likelihood between models.

The second method for model comparison was the Akaike information criterion (AIC) (3). Although it was derived under stringent assumptions, this measure is often used without regard to those assumptions. The AIC is equal to

$$\text{AIC} = -2LL(\boldsymbol{\theta}^*) + 2 \cdot \text{number of parameters}.$$

For ease of comparison with the Bayes factor results, we multiplied AIC by −0.5: $-0.5\text{AIC} = LL(\boldsymbol{\theta}^*) - \text{number of parameters}$.

Model comparison results are given in Tables 1 and 2 for the main experiment (humans and monkeys) and in Tables S3 and S4 for the control experiment (humans). Parameter ranges are given in Table S1.

Psychometric Curves. After fitting each model, we computed model fits to the psychometric curves. To compute the model fits for the psychometric curves as a function of contrast and orientation (Figs. 2*C*, 3*C*, etc.), we averaged, separately for every subject, contrast, and orientation bin, Eq. **S3** with parameters substituted across all values of s presented to that subject at that contrast in that orientation bin. In these figures, orientation was binned into 13 bins with centers equally spaced between −18.46° to 18.46° (this means that the data were cut off at ±20°).

To compute the model fits for the psychometric curves as a function of contrast and category (Figs. 2*B*, 3*B*, etc.), we averaged, separately for every subject, contrast, and true category, Eq. **S3** with parameters substituted across all values of s presented to that subject at that contrast with that true category. This procedure explains why the model fits do not look smooth: they are based on the orientations in the actual experiment, which were drawn randomly from their respective category-conditioned distributions.

Finally, to compute the model fits for accuracy as a function of contrast (Figs. 2*A*, 3*A*, etc.), we averaged, separately for every subject and contrast, the probability of a correct response across all values of s presented to that subject at that contrast. The probability of a correct response was equal to Eq. **S3** when the true category was 1 on that trial, and 1 minus Eq. **S3** when the true category was 2.

Throughout the paper, root mean squared error (RMSE) was computed based on vectorized forms of the subject-averaged data and corresponding subject-averaged model fits across all conditions in a plot.

Flexible Model. The flexible model was designed to provide a model-neutral estimate of the decision boundary as a function of contrast (Figs. 2 *D–F* and 3*D* and Figs. S1*D* and S2*D*). This model has the following parameters: α, β, and γ to parametrize the relationship between σ and contrast, lapse rate λ, and the boundary at each contrast, k_c. We compared the boundaries estimated by the flexible model with those predicted by the Opt-P, Lin-σ, Quad-σ, and Fixed models. To this end, in each of these four models, we fixed α, β, γ, and λ to their estimates from the flexible model, and then fitted the remaining parameters (p_1 for Opt-P, k_0 and σ_p for Lin-σ and Quad-σ, and k_0 for Fixed), and finally substituted all parameters in the model's expression for the decision boundary. These fits produced the shaded areas in Figs. 2*D* and 3*D* and Figs. S1*D* and S2*D*.

Orientation Discrimination Experiment. To obtain an independent measure of subjects' sensory noise level, we conducted an orientation discrimination task. The same six subjects participated as in the main categorization experiment. Subjects determined whether an oriented Gabor similar to the one used in the categorization task was tilted clockwise or counterclockwise with respect to the horizontal. This task was done at the same contrast levels as used in the categorization task. Orientation was ±1.2°, ±3°, ±5°, or ±8°, all with equal probability (method of constant stimuli). We estimated the sensory noise parameter σ separately at each contrast level by fitting a cumulative normal distribution using maximum-likelihood estimation.

To obtain Fig. 2*E*, we first computed, for each subject and each contrast, an estimate of σ using the equation

$$\hat{\sigma}(c) = \sqrt{(\hat{\alpha}c)^{-\hat{\beta}} + \hat{\gamma}}, \qquad \textbf{[S4]}$$

where $\hat{\alpha}, \hat{\beta}, \hat{\gamma}$ are estimates obtained from the flexible model. We then scattered those against the corresponding sensory noise estimates from the discrimination experiment.

In the section Flexible Model, we mentioned that we used the estimates of α, β, and γ from the flexible model to compute the predictions of the Opt-P, Lin-σ, and Quad-σ models for the decision boundaries (Fig. 2*D*). This computation was done via an estimate of σ as given by Eq. **S4**. To obtain Fig. 2*F*, we replaced, for each subject and each contrast, those estimates by the estimates obtained from the discrimination experiment, changing nothing else; in particular, the remaining parameters were not refitted.

Neural Likelihood Function. We use the Poisson-like distribution in Eq. **5** to model the variability of a population of sensory input neurons

$$p(\mathbf{r} \mid s,g) = \varphi(\mathbf{r},g)e^{\mathbf{h}(s)\cdot\mathbf{r}}.$$

As a consequence, the likelihood function of the stimulus is

$$\begin{aligned} L_{\mathbf{r}}(s) = p(\mathbf{r} \mid s) &= \int p(\mathbf{r} \mid s,g)p(g)dg \\ &= \left(\int \varphi(\mathbf{r},g)p(g)dg\right)e^{\mathbf{h}(s)\cdot\mathbf{r}} \equiv \Phi(\mathbf{r})e^{\mathbf{h}(s)\cdot\mathbf{r}}. \end{aligned}$$

The likelihood of category C is

$$p(\mathbf{r} \mid C) = \int L_{\mathbf{r}}(s)p(s \mid C)ds = \Phi(\mathbf{r})\int e^{\mathbf{h}(s)\cdot\mathbf{r}}p(s \mid C)ds.$$

To make progress, we need to make assumptions about $\mathbf{h}(s)$. We will assume that it is a quadratic function of s, so that the likelihood $L_{\mathbf{r}}(s)$ is an (unnormalized) Gaussian. Under this assumption, we can write $\mathbf{h}(s)$ as

$$\mathbf{h}(s) = -\frac{1}{2}s^2\mathbf{a} + s\mathbf{b},$$

where $\mathbf{a}$ and $\mathbf{b}$ are constant vectors. Then the stimulus likelihood function is

$$L_{\mathbf{r}}(s) = \Phi(\mathbf{r})e^{\mathbf{h}(s)\cdot\mathbf{r}} = \Phi(\mathbf{r})e^{-\frac{1}{2}s^2\mathbf{a}\cdot\mathbf{r}+s\mathbf{b}\cdot\mathbf{r}} \propto \exp\left(-\frac{\left(s-\frac{\mathbf{b}\cdot\mathbf{r}}{\mathbf{a}\cdot\mathbf{r}}\right)^2}{2(\mathbf{a}\cdot\mathbf{r})^{-1}}\right). \quad \textbf{[S5]}$$

This expression shows that the maximum-likelihood estimate of the stimulus is equal to $\frac{\mathbf{b}\cdot\mathbf{r}}{\mathbf{a}\cdot\mathbf{r}}$, and the variance of the normalized likelihood function over the stimulus is equal to $\frac{1}{\mathbf{a}\cdot\mathbf{r}}$. These quantities correspond to x and σ^2 in the behavioral model, respectively. In the special case of independent Poisson variability and Gaussian tuning curves (4), we have

$$h_i(s) = \log f_i(s) = -\frac{\left(s-s_i^{\text{pref}}\right)^2}{2\sigma_{\text{tc}}^2} = -\frac{1}{2\sigma_{\text{tc}}^2}s^2 + \frac{s_i^{\text{pref}}}{\sigma_{\text{tc}}^2}s + \text{constant}.$$

where s_i^{pref} is the preferred stimulus of the ith neuron, and σ_{tc} is the width of tuning curve. Therefore, $a_i = 1/\sigma_{\text{tc}}^2$ and $b_i = s_i^{\text{pref}}/\sigma_{\text{tc}}^2$. The mean of the likelihood function over the stimulus is $\frac{\mathbf{b}\cdot\mathbf{r}}{\mathbf{a}\cdot\mathbf{r}} = \frac{\sum_{i=1}^N r_i s_i^{\text{pref}}}{\sum_{i=1}^N r_i}$, which is the center-of-mass (population vector) decoder. The variance of the normalized likelihood function is $\frac{1}{\mathbf{a}\cdot\mathbf{r}} = \frac{\sigma_{\text{tc}}^2}{\sum_{i=1}^N r_i}$. Substituting this mean and variance into Eq. **3** gives us Eq. **6**.

Neural Networks. Most neural network methods were similar to the ones described in our earlier work on visual search (5). Input consisted of activity in a population of 41 independent Poisson neurons with Gaussian tuning curves $[f_1(s),..,f_{41}(s)]$, with $f_i(s) = ge^{-\frac{\left(s-s_i^{\text{pref}}\right)^2}{2\sigma_{\text{tc}}^2}}$, where $\sigma_{\text{tc}} = 10°$ and preferred orientations s_i^{pref} ranged from −60° to 60° in steps of 3°. Our results are insensitive to these numerical choices. Gain was varied, as it represents the effect of contrast. We considered three networks, each of which is characterized by a set of basis functions

$$\begin{aligned} \mathbf{R}^{\text{QDN}} &= \left[\frac{r_i r_j}{1+\mathbf{V}\cdot\mathbf{r}+\mathbf{r}^{\text{T}}\mathbf{V}\mathbf{R}}\right] \\ \mathbf{R}^{\text{LIN}} &= [1, r_i] \\ \mathbf{R}^{\text{LIN}} &= [1, r_i, r_i r_j]. \end{aligned}$$

The output activity $\mathbf{z}$ is now a linear combination of the basis functions in the network, with fixed coefficients. We further impose the condition that the output activity $\mathbf{z}$ is also Poisson-like: $p(\mathbf{z} \mid C, g_{\mathbf{z}}) = \varphi_{\mathbf{z}}(\mathbf{z}, g_{\mathbf{z}})e^{\mathbf{H}(C)\cdot\mathbf{z}}$. The log likelihood ratio over C encoded in $\mathbf{z}$ is then $\log\frac{p(\mathbf{z} \mid C=1)}{p(\mathbf{z} \mid C=2)} = (\mathbf{h}(C=1)-\mathbf{h}(C=2))\cdot\mathbf{z}$, which we write shorthand as $\Delta\mathbf{H}\cdot\mathbf{z}$. The network approximation to the log likelihood ratio under the assumption of Poisson-like output is then

$$d_{\text{network}}(\mathbf{r};\mathbf{w}) = \Delta\mathbf{h}\cdot\mathbf{z} = \mathbf{w}\cdot\mathbf{r}^{\text{network}},$$

where $\mathbf{R}^{\text{network}}$ is $\mathbf{R}^{\text{QDN}}$, $\mathbf{R}^{\text{LIN}}$, or $\mathbf{R}^{\text{QUAD}}$ and $\mathbf{w}$ is the vector of all network parameters ($\mathbf{W}$, $\mathbf{v}$, and $\mathbf{V}$). The network approximation to the posterior is

$$q(C \mid \mathbf{r};\mathbf{w}) = \frac{1}{1+e^{-Cd_{\text{network}}(\mathbf{r};\mathbf{w})}}.$$

Network Training. We trained networks by minimizing the Kullback-Leibler distance between the network posterior and the optimal posterior over category using stochastic gradient descent. The Kullback-Leibler distance, averaged over $\mathbf{r}$, is

$$\begin{aligned} \langle D_{\text{KL}}\rangle_{\mathbf{r}} &= \sum_{\mathbf{r}} p(\mathbf{r}) \sum_{C=1}^{2} p(C \mid \mathbf{r}) \log\frac{p(C \mid \mathbf{r})}{q(C \mid \mathbf{r};\mathbf{w})} \\ &= \sum_{\mathbf{r},C} p(C,\mathbf{r}) \log\frac{p(C \mid \mathbf{r})}{q(C \mid \mathbf{r};\mathbf{w})}. \end{aligned}$$

The gradient is

$$\begin{aligned} \frac{\partial\langle D_{\text{KL}}\rangle_{\mathbf{r}}}{\partial\mathbf{w}} &= -\frac{\partial}{\partial\mathbf{w}}\sum_{\mathbf{r},C} p(C,\mathbf{r}) \log q(C \mid \mathbf{r};\mathbf{w}) \\ &= -\sum_{\mathbf{r},C} p(C,\mathbf{r})\frac{\partial}{\partial\mathbf{w}}\log\frac{1}{1+e^{-Cd(\mathbf{r};\mathbf{w})}} \\ &= \sum_{\mathbf{r},C} p(C,\mathbf{r})\frac{-Ce^{-Cd(\mathbf{r},\mathbf{w})}}{1+e^{-Cd(\mathbf{r},\mathbf{w})}}\frac{\partial}{\partial\mathbf{w}}d(\mathbf{r};\mathbf{w}) \\ &= -\sum_{\mathbf{r},C} p(C,\mathbf{r})C(1-q(C \mid \mathbf{r};\mathbf{w}))\frac{\partial}{\partial\mathbf{w}}d(\mathbf{r};\mathbf{w}) \\ &\approx -\left\langle C(1-q(C \mid \mathbf{r};\mathbf{w}))\frac{\partial}{\partial\mathbf{w}}d(\mathbf{r};\mathbf{w})\right\rangle_{\text{samples of }(\mathbf{r},C)}, \end{aligned}$$

where the last step is a sampling approximation. The change in weights from one iteration to the next is proportional to this gradient and has opposite sign. This produces the learning rule

$$\mathbf{w}^{k+1} = \mathbf{w}^k + \alpha \left\langle C(1 - q(C \mid \mathbf{r}; \mathbf{w})) \frac{\partial}{\partial \mathbf{w}} d(\mathbf{r}; \mathbf{w}) \right\rangle_{\text{samples of } (\mathbf{r}, C)}, \quad \textbf{[S6]}$$

where α is the learning rate. We used an adaptive method (6) to adjust the learning rate. We drew 10,000 trials on each iteration, and terminated learning after 10,000 iterations for the QDN network and after 100,000 iterations for LIN and QUAD. We then tested on 100,000 trials. For the QDN network, the initial values of the parameters were chosen according to Eq. **6**. For LIN and QUAD, they were given by a first- and second-order Taylor expansion of Eq. **6** around the mean activity, $\langle \mathbf{r} \rangle$, respectively, except that the weight to the constant term was set to 0 for better convergence. Information loss was measured as the average Kullback-Leibler distance between the optimal posterior and the network posterior, normalized by the mutual information between the input activity and category:

$$\frac{\delta I}{I} = \frac{\langle D_{\mathrm{KL}} \rangle_{\mathbf{r}}}{I(C, \mathbf{r})} = \frac{\sum_{\mathbf{r},C} p(C, \mathbf{r}) \log \frac{p(C \mid \mathbf{r})}{q(C \mid \mathbf{r}; \mathbf{w})}}{\sum_{\mathbf{r},C} p(C, \mathbf{r}) \log \frac{p(C \mid \mathbf{r})}{p(C)}} = \frac{\langle \log p(C \mid \mathbf{r}) - \log q(C \mid \mathbf{r}; \mathbf{w}) \rangle_{\text{samples of } (\mathbf{r}, C)}}{\langle \log p(C \mid \mathbf{r}) - \log p(C) \rangle_{\text{samples of } (\mathbf{r}, C)}}.$$

Note that this number can be greater than 1.

Visualization of Network Performance. To appreciate the ability of the QDN network to approximate a highly nonlinear decision surface, we plotted the optimal log likelihood ratio d as a function of the input quantities $\mathbf{a} \cdot \mathbf{r}$ and $\mathbf{b} \cdot \mathbf{r}$ (Fig. S3*A*, surface), along with the log likelihood ratios obtained from the QDN network. The plane at $d = 0$ separates the network categorization decisions well, showing that the network makes the same decisions as the Bayesian observer. More importantly, the network decision variable follows the optimal decision variable closely, despite its highly nonlinear shape, even at low values of precision ($\mathbf{a} \cdot \mathbf{r} < 1$ deg^{-2}, corresponding to a sensory uncertainty of more than 1°). This similarity shows that the network does not only make near-optimal categorization decisions (and thus adjust the decision boundary on every trial based on sensory uncertainty), but also correctly computes decision confidence (absolute value of d), regardless of the quality of the input.

Fig. S3*B* shows the pattern of weights learned by the QDN network. These weights are multiplied by the basis functions corresponding to all possible products of activities of two input neurons (shown in Fig. S3*C* for three values of orientation). Positive (negative) weights indicate that activity of the corresponding basis functions contributes to evidence for category 1 (2). The observed pattern makes intuitive sense: category 1 population activity tends to be more symmetric around zero than category 2 activity; therefore, simultaneously high activity on both sides of zero is evidence for category 1, whereas high activity in a subpopulation with preferred stimuli away from zero is a telltale sign of category 2. The basis function activity patterns in Fig. 6*C* would lead to categorization decisions 2, 1, and 2, respectively.

1. Kass RE, Raftery AE (1995) Bayes factors. *J Am Stat Assoc* 90(430):773–795.
2. MacKay D (2003) *Information Theory, Inference and Learning Algorithms* (Cambridge Univ Press, Cambridge).
3. Akaike H (1974) A new look at the statistical model identification. *IEEE Trans Automat Contr* 19(6):716–723.
4. Ma WJ (2010) Signal detection theory, uncertainty, and Poisson-like population codes. *Vision Res* 50(22):2308–2319.
5. Ma WJ, Navalpakkam V, Beck JM, Berg Rv, Pouget A (2011) Behavior and neural basis of near-optimal visual search. *Nat Neurosci* 14(6):783–790.
6. Almeida LB, Langlois T, Amaral JD, Plakhov A (1999) Parameter adaptation in stochastic optimization. *On-Line Learning in Neural Networks* (*Publications of the Newton Institute*), ed D, Saad E (Cambridge Univ Press, Cambridge), pp 111–134.

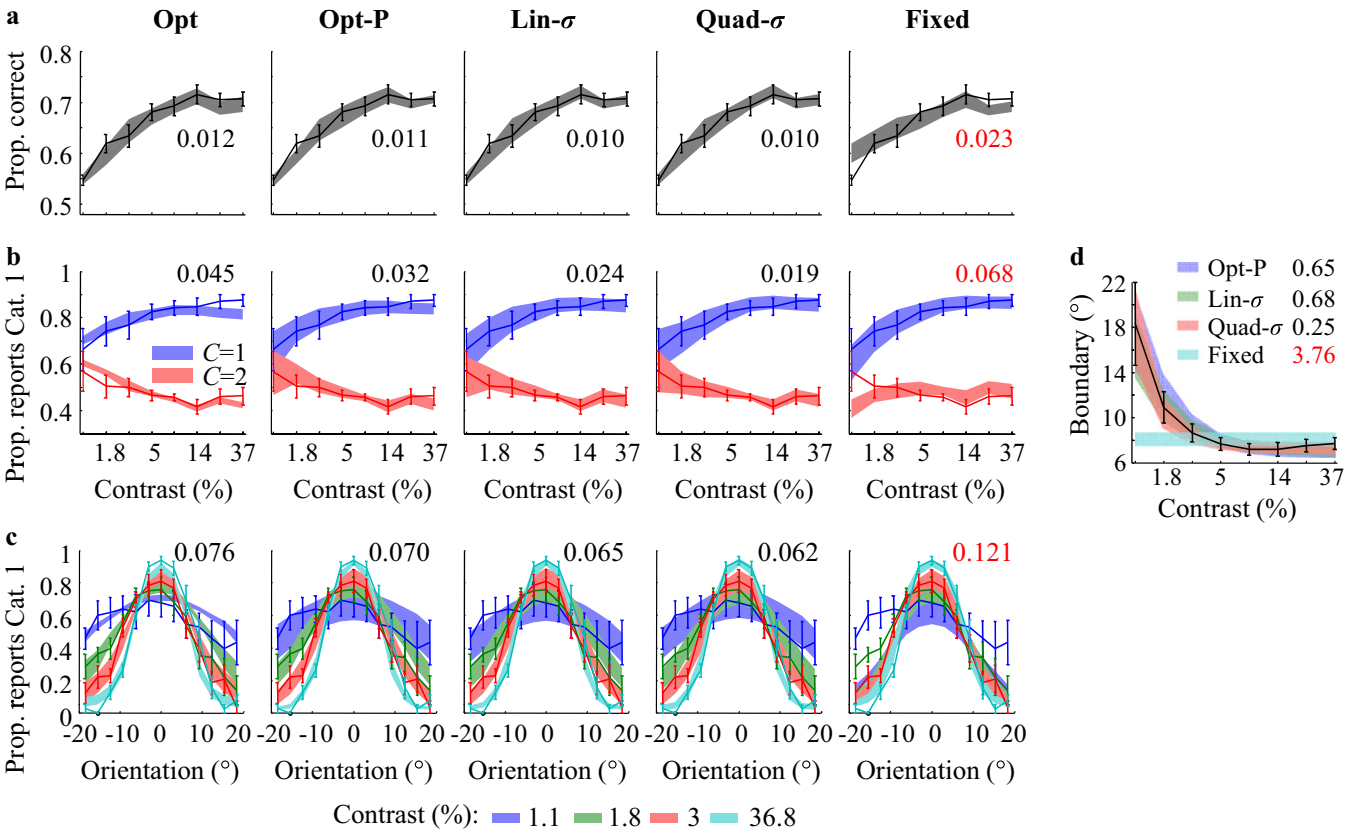

Fig. S1. As in Fig. 2, but for the human control experiment (central orientation 45° clockwise with respect to vertical).

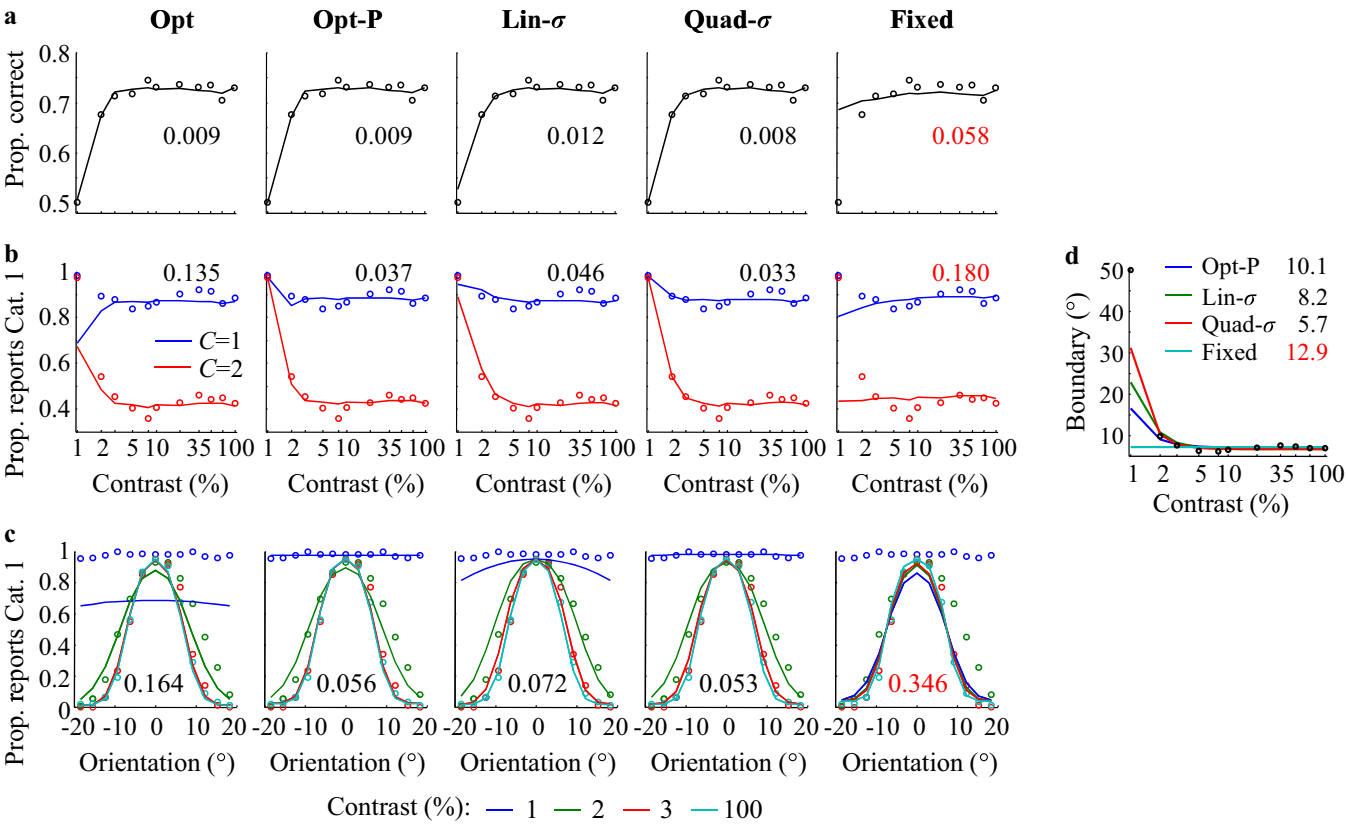

Fig. S2. As in Fig. 3, but for monkey A.

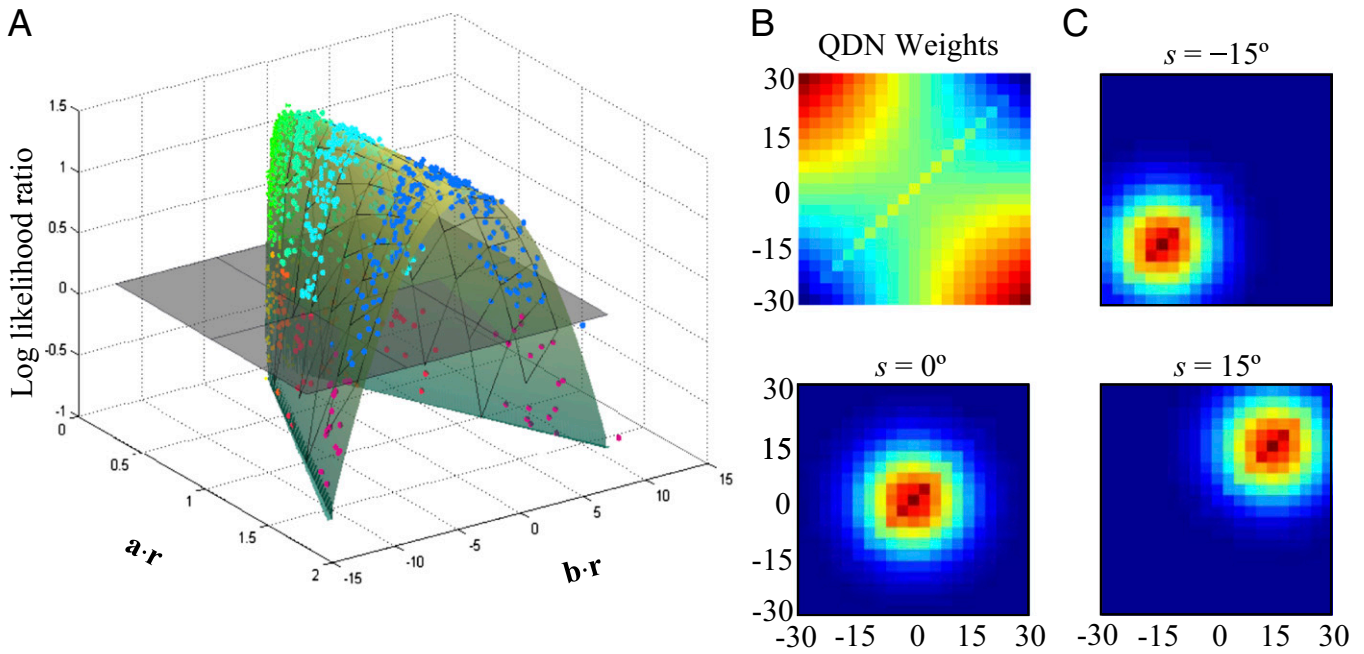

Fig. S3. Properties of the divisive normalization network. (*A*) The surface represent the optimal log likelihood ratio, *d*, as a function of $\mathbf{a}\cdot\mathbf{r} = 1/\sigma^2$ and $\mathbf{b}\cdot\mathbf{r} = x/\sigma^2$. Variations on the **a**·**r**-axis reflect changes in gain (or contrast), and variations on the $\mathbf{b}\cdot\mathbf{r}$-axis can be interpreted as changes in the stimulus measurement. The $d = 0$ plane (gray) separates the optimal Category 1 and 2 reports. Dots represent the QDN network log likelihood ratios, $\Delta\mathbf{H}\cdot\mathbf{z}$, for different combinations of gain and orientation. Patches of same-colored dots along the **a**·**r**-axis correspond to same-gain trials. Cool colors represent network reports of Category 1 and warm colors correspond to network reports of Category 2. The QDN network not only provides correct categorizations, but also a close approximation to the optimal log likelihoods. (*B*) Quadratic weights in the trained QDN network. Negative values are in blue, positive ones in red. (*C*) Average basis function activity in the QDN network for orientations $s = -15°$, $s = 0°$, $s = 15°$. Each entry in the matrix represents a quadratic basis function obtained by multiplying the spike counts of two input neurons. Zeros are in blue, the highest values in red. The inner product of the weights in (b) with the $s = -15°$ and $s = 15°$ activity yields a negative log likelihood ratio (evidence for Category 2), while with the $s = 0°$ activity it results in a positive log likelihood ratio (evidence for Category 1).

Table S1. Decision rules and parameter sets of all models

Model	Decision rule $\lvert x\rvert<k(\sigma)$, with $k(\sigma)=\ldots$	Parameters
Probabilistic computation		
Opt	$\sqrt{(\sigma^2+\sigma_1^2)(\sigma^2+\sigma_2^2)(\sigma_2^2-\sigma_1^2)^{-1}\log\frac{\sigma^2+\sigma_2^2}{\sigma^2+\sigma_1^2}}$	α, β, γ, λ
Opt-P	$\sqrt{(\sigma^2+\sigma_1^2)(\sigma^2+\sigma_2^2)(\sigma_2^2-\sigma_1^2)^{-1}(\log\frac{\sigma^2+\sigma_2^2}{\sigma^2+\sigma_1^2}+2\log\frac{p_1}{1-p_1})}$	α, β, γ, λ, p_1
Lin-σ	$k_0(1+\frac{\sigma}{\sigma_p})$	α, β, γ, λ, k_0, σ_p
Quad-σ	$k_0(1+\frac{\sigma^2}{\sigma_p^2})$	α, β, γ, λ, k_0, σ_p
Nonprobabilistic computation		
Fixed	Constant k_0	α, β, γ, λ, k_0

Table S2. Parameter estimates and ranges

Model	Parameter	Range	Humans main	Humans control	Monkey A	Monkey L
Opt	α	(0,50]	10.2 ± 1.0	14.3 ± 2.6	32.3	8.11
	β	(0,8]	2.32 ± 0.12	3.14 ± 0.33	6.79	2.87
	γ	(0,30]	3.20 ± 0.67	11.5 ± 3.0	9.91	18.2
	λ	(0,0.5]	0.108 ± 0.031	0.125 ± 0.065	0.019	0.051
Opt-P	α	(0,50]	9.2 ± 1.5	11.4 ± 3.7	34.5	13.3
	β	(0,8]	2.22 ± 0.26	2.62 ± 0.56	7.97	4.46
	γ	(0,30]	2.80 ± 0.27	9.8 ± 1.3	8.29	17.1
	λ	(0,0.5]	0.095 ± 0.026	0.054 ± 0.018	0.044	0.068
	p_1	[0.25, 0.75]	0.501 ± 0.024	0.514 ± 0.036	0.531	0.514
Lin-σ	α	(0,50]	8.6 ± 1.5	11.4 ± 3.4	25.9	11.9
	β	(0,8]	2.11 ± 0.13	2.59 ± 0.38	4.07	3.69
	γ	(0,30]	3.47 ± 0.43	10.8 ± 1.5	9.27	18.2
	λ	(0,0.5]	0.070 ± 0.024	0.063 ± 0.030	0.032	0.050
	k_0	(0,15]	3.14 ± 0.46	4.7 ± 1.0	0.440	0.614
	σ_p	(0,50]	3.29 ± 0.67	16.4 ± 7.2	0.213	0.352
Quad-σ	α	(0,50]	7.1 ± 1.8	9.9 ± 3.4	26.0	3.20
	β	(0,8]	1.78 ± 0.17	2.40 ± 0.41	4.21	1.66
	γ	(0,30]	2.65 ± 0.37	10.2 ± 1.4	8.73	17.2
	λ	(0,0.5]	0.078 ± 0.025	0.064 ± 0.032	0.037	0.047
	k_0	(0,15]	4.94 ± 0.32	6.78 ± 0.41	5.10	5.55
	σ_p	(0,50]	6.37 ± 0.68	15.4 ± 2.7	5.17	6.42
Fixed	α	(0,50]	21.3 ± 6.5	12.4 ± 7.3	7.79	9.61
	β	(0,8]	5.3 ± 3.0	2.02 ± 0.69	0.98	1.53
	γ	(0,30]	5.47 ± 0.81	8.5 ± 2.4	7.06	18.2
	λ	(0,0.5]	0.120 ± 0.038	0.109 ± 0.038	0.081	0.070
	k_0	(0,50]	6.42 ± 0.32	8.13 ± 0.66	7.23	8.35

Disclaimer: The meaningfulness of parameter estimates depends on the goodness of fit of the model.

Table S3. Model comparison using AIC for main experiment

Subject	Opt	Opt-P	Lin-σ	Quad-σ	Fixed	DIFF
Human 1	−1,471.7	−1,418.8	−1,409.2	−1,412.2	−1,442.0	−32.8
Human 2	−1,635.2	−1,636.1	−1,635.3	−1,647.6	−1,709.7	−74.5
Human 3	−883.1	−882.3	−883.4	−883.0	−908.8	−26.5
Human 4	−1,389.5	−1,374.1	−1,338.4	−1,350.3	−1,494.8	−156.4
Human 5	−1,273.6	−1,262.4	−1,225.8	−1,229.6	−1,312.3	−86.5
Human 6	−1,145.6	−1,122.7	−1,122.9	−1,123.4	−1,155.2	−32.4
Monkey A	−33,864.0	−32,847.8	−32,928.3	−32,741.5	−35,338.4	−2,596.9
Monkey L	−71,961.0	−71,771.2	−71,439.0	−71,478.5	−73,152.4	−1,713.4

Numbers are AIC multiplied by −0.5 for every model and every subject. Shaded in green are the models whose values fall within log(30) of the value of the best model. The Fixed model is never among them. DIFF, difference between the Fixed model and the best probabilistic model.

Table S4. Model comparison using log marginal likelihood for control experiment

Subject	Opt	Opt-P	Lin-σ	Quad-σ	Fixed	DIFF
Human 1	−1,216.8	−1,175.4	−1,173.5	−1,175.7	−1,179.1	−5.7
Human 2	−1,093.6	−968.1	−970.5	−972.4	−1,021.4	−53.3
Human 3	−1,738.4	−1,701.7	−1,678.6	−1,678.8	−1,685.7	−7.1
Human 4	−1,607.9	−1,565.3	−1,558.2	−1,557.5	−1,568.5	−11.0
Human 5	−1,156.4	−1,156.0	−1,111.0	−1,110.6	−1,112.6	−1.9
Human 6	−1,184.2	−1,120.4	−1,128.9	−1,113.9	−1,292.2	−178.2

See Table 1 for description.

Table S5. Model comparison using AIC for control experiment

Subject	Opt	Opt-P	Lin-σ	Quad-σ	Fixed	DIFF
Human 1	−1,211.1	−1,165.6	−1,164.5	−1,166.1	−1,170.4	−5.9
Human 2	−1,089.2	−961.1	−960.7	−961.7	−1,016.5	−55.9
Human 3	−1,734.9	−1,689.6	−1,670.0	−1,668.8	−1,675.9	−7.0
Human 4	−1,602.8	−1,554.9	−1,547.5	−1,546.8	−1,559.4	−12.6
Human 5	−1,149.8	−1,146.5	−1,101.5	−1,099.4	−1,101.1	−1.7
Human 6	−1,176.1	−1,111.8	−1,113.0	−1,100.0	−1,273.0	−173.0

See Table S3 for description.