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:

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

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

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

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

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

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 $\boldsymbol{\theta}$. We aimed to find the parameter combination $\boldsymbol{\theta}$ 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

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

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 **a** and **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 **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 **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 **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 **w** is the vector of all network parameters (**W**, **v**, and **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 **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.

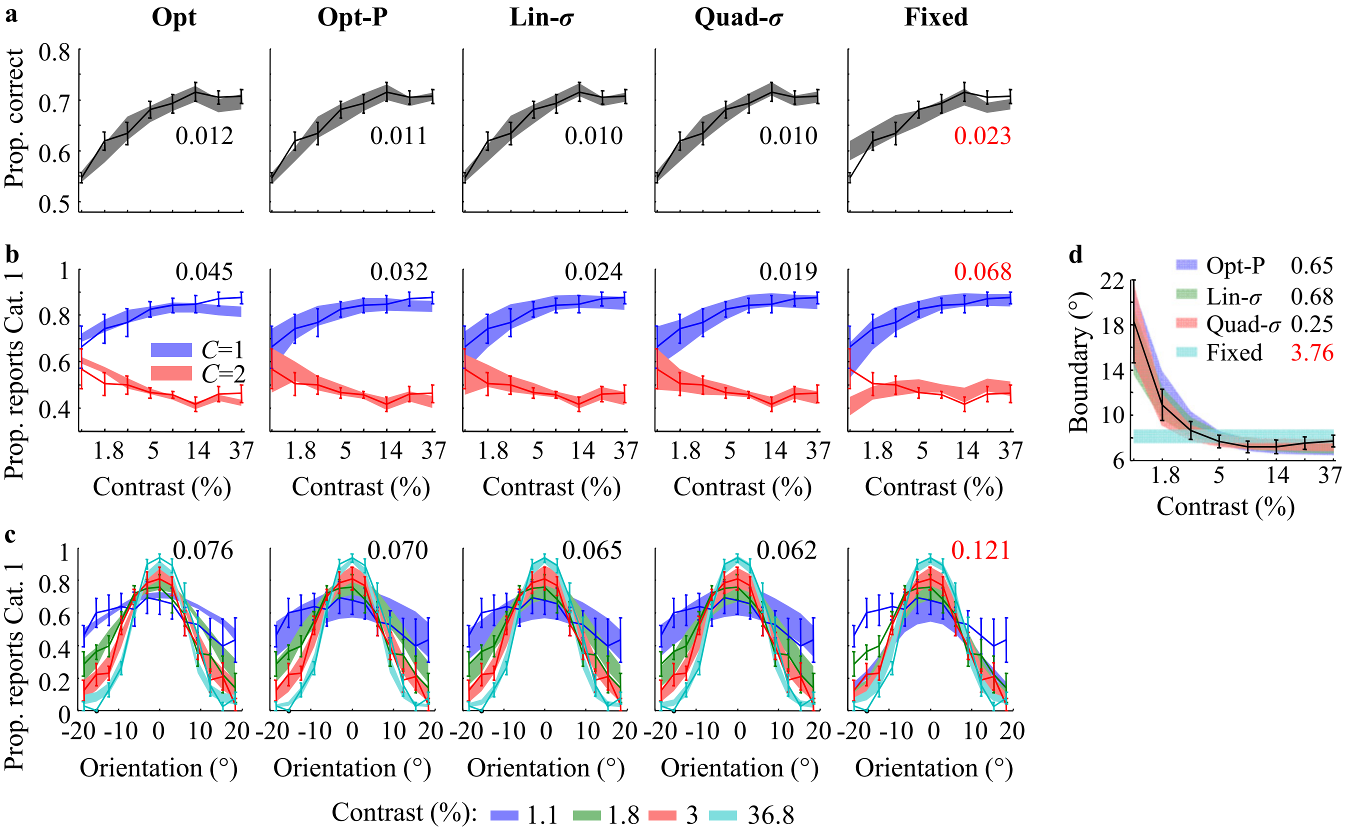

Fig. S1. As in Fig. 2, but for the human control experiment (central orientation 45° clockwise with respect to vertical).

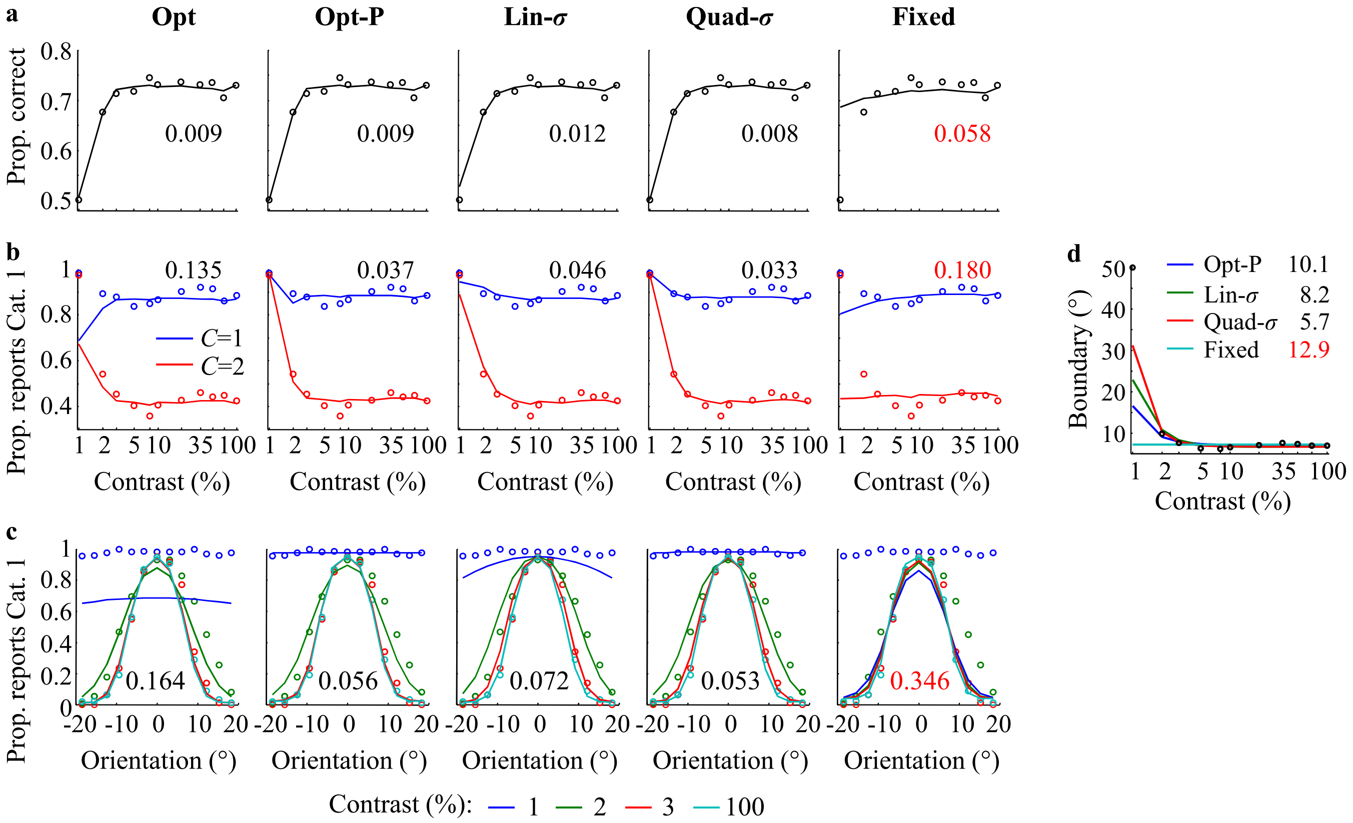

Fig. S2. As in Fig. 3, but for monkey A.

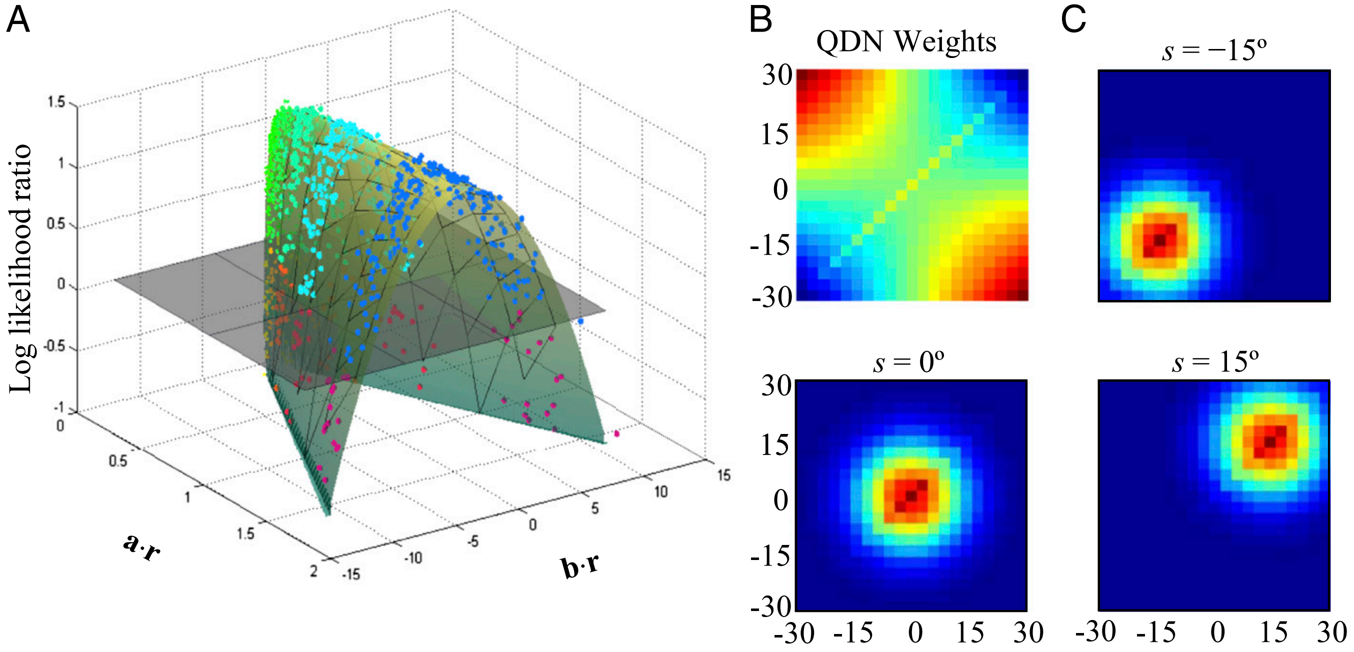

Fig. S3. Properties of the divisive normalization network. (*A*) The surface represent the optimal log likelihood ratio, *d*, as a function of $\mathbf{a}\cdot\mathbf{r} = 1/\sigma^2$ and $\mathbf{b}\cdot\mathbf{r} = x/\sigma^2$. Variations on the **a**·**r**-axis reflect changes in gain (or contrast), and variations on the **b** · **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 $s = -15^{\circ}$, $s = 0^{\circ}$, $s = 15^{\circ}$. 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 $s = -15^{\circ}$ and $s = 15^{\circ}$ activity yields a negative log likelihood ratio (evidence for Category 2), while with the $s = 0^{\circ}$ 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.