

Near-optimal visual search: behavior and neural basis

Wei Ji Ma, Vidhya Navalpakkam, Jeffrey Beck, Ronald van den Berg, Alexandre Pouget

Supplementary Results and Figures

1	Theory.....	1
1.1	Global log likelihood ratio	1
1.2	Local log likelihood ratio (neural form).....	3
1.3	Local log likelihood ratio (behavioral form).....	5
1.4	Relation of optimal model to max and sum models.....	7
1.5	Non-optimal local decision variables.....	8
2	Behavioral experiments	8
2.1	Experiments 1a and 2a	8
2.2	Individual-subject ROCs	9
2.3	Bayesian model comparison.....	9
3	Neural implementation	10
3.1	Intuition behind a nonlinearity	10
3.2	Network performance.....	10
3.3	Effect of non-Poisson-like statistics.....	11

Supplementary Figures are at the end.

1 Theory

1.1 Global log likelihood ratio

Optimal decisions are based on the probabilities of the alternatives given the noisy evidence¹. In our experiments, the alternatives are “target absent” ($T=0$) and “target present” ($T=1$). We call the variable T global target presence. In this section, we use a formulation of the model in which observations are patterns of neural activity at all N locations, $\mathbf{r}_1, \dots, \mathbf{r}_N$ (Fig. 7a). This formulation is identical to the one discussed in the main text, except that population activity \mathbf{r}_i is used instead of the scalar internal representation x_i . We will relate the scalar representation to population activity in Section 1.3, but the formulation with population activity is more complete because it allows for the automatic encoding of sensory uncertainty. The log likelihood ratio of global target presence (also called global decision variable) is defined as

$$d = \log \frac{p(\mathbf{r}_1, \dots, \mathbf{r}_N | T = 1)}{p(\mathbf{r}_1, \dots, \mathbf{r}_N | T = 0)}. \quad (\text{S1})$$

We now review the derivation of the expression for d in terms of locally defined quantities²⁻⁴. We assume that given target presence (0 or 1) at each location, the variability in population activity \mathbf{r}_i is conditionally independent between locations. As a consequence, when the target is absent, the probability of observing $\mathbf{r}_1, \dots, \mathbf{r}_N$ is equal to the product of the probabilities of each pattern \mathbf{r}_i given that the i^{th} location contains a distractor ($T_i=0$):

$$p(\mathbf{r}_1, \dots, \mathbf{r}_N | T = 0) = \prod_{i=1}^N p(\mathbf{r}_i | T_i = 0). \quad (\text{S2})$$

If the target is present, it can be located at any of the N locations. We denote by $p(T_i=1|T=1)$ the probability that location i contains the target in a target-present display. The probability of observing $\mathbf{r}_1, \dots, \mathbf{r}_N$ if the target is present is obtained by marginalizing out target location, i.e., taking a weighted average over all possible target locations:

$$p(\mathbf{r}_1, \dots, \mathbf{r}_N | T = 1) = \sum_{i=1}^N p(T_i = 1 | T = 1) p(\mathbf{r}_1, \dots, \mathbf{r}_N | T_i = 1). \quad (\text{S3})$$

The conditional probability $p(\mathbf{r}_1, \dots, \mathbf{r}_N | T_i=1)$ is computed by using the fact that if the target is present at the i^{th} location, then it is absent at all other locations:

$$\begin{aligned} p(\mathbf{r}_1, \dots, \mathbf{r}_N | T = 1) &= \sum_{i=1}^N p(T_i = 1 | T = 1) p(\mathbf{r}_i | T_i = 1) \prod_{j \neq i} p(\mathbf{r}_j | T_j = 0) \\ &= \left(\prod_{j=1}^N p(\mathbf{r}_j | T_j = 0) \right) \sum_{i=1}^N p(T_i = 1 | T = 1) \frac{p(\mathbf{r}_i | T_i = 1)}{p(\mathbf{r}_i | T_i = 0)} \\ &= \left(\prod_{j=1}^N p(\mathbf{r}_j | T_j = 0) \right) \sum_{i=1}^N p(T_i = 1 | T = 1) e^{d_i}. \end{aligned} \quad (\text{S4})$$

Here, d_i is the local log likelihood ratio (also called local decision variable) at location i , defined as

$$d_i = \log \frac{p(\mathbf{r}_i | T_i = 1)}{p(\mathbf{r}_i | T_i = 0)}. \quad (\text{S5})$$

Dividing Eq. (S4) by Eq. (S2) and taking the log, we find the global log likelihood ratio as

$$d = \log \sum_{i=1}^N p(T_i = 1 | T = 1) e^{d_i}. \quad (\text{S6})$$

Finally, we assume that all locations are equally likely to contain the target and the observer uses this knowledge, so that $p(T_i=1|T=1)=1/N$. Then Eq. (S6) becomes

$$d = \log \frac{1}{N} \sum_{i=1}^N e^{d_i}, \quad (\text{S8})$$

which is Eq. (2) in the main text. If d is positive, the observer's response is "target present", otherwise "target absent"; this is maximum-a-posteriori readout. It is important to note that Eq. (S8) holds regardless of the noise model at a given location, $p(\mathbf{r}_i|T_i)$. It does not depend either on whether we use \mathbf{r}_i , x_i , or some other variable to describe the local observations. The only assumptions we have made are that each location is equally likely to contain the target and that variability is conditionally independent between locations. The next step is to further evaluate the local log likelihood, d_i .

1.2 Local log likelihood ratio (neural form)

Eq. (S5) expresses the local log likelihood ratio in terms of the probabilities of the activity in the i^{th} population, \mathbf{r}_i , given target absence or presence at that location. These probabilities are obtained by marginalizing over s_i , the stimulus at that location:

$$d_i = \log \frac{\int p(\mathbf{r}_i | s_i) p(s_i | T_i = 1) ds_i}{\int p(\mathbf{r}_i | s_i) p(s_i | T_i = 0) ds_i}. \quad (\text{S9})$$

Since the target always has value s_T , the numerator is equal to $p(\mathbf{r}_i|s_T)$.

When the reliability of the stimulus is unknown, the stimulus likelihood, $L(s_i)=p(\mathbf{r}_i|s_i)$, must be obtained by marginalization over the nuisance parameters \mathbf{c}_i , such as contrast. The most important feature of Poisson-like variability (Eq. (5) in the main text) is that this marginalization does not affect the form of the stimulus-dependence of the likelihood:

$$\begin{aligned} p(\mathbf{r}_i | s_i) &= \int p(\mathbf{r}_i | s_i, \mathbf{c}_i) p(\mathbf{c}_i) d\mathbf{c}_i \\ &= \int \varphi(\mathbf{r}_i, \mathbf{c}_i) e^{\mathbf{h}_r(s_i) \cdot \mathbf{r}_i} p(\mathbf{c}_i) d\mathbf{c}_i \\ &= \left(\int \varphi(\mathbf{r}_i, \mathbf{c}_i) p(\mathbf{c}_i) d\mathbf{c}_i \right) e^{\mathbf{h}_r(s_i) \cdot \mathbf{r}_i}. \end{aligned} \quad (\text{S10})$$

For the local log likelihood ratio, Eq. (S9), this implies

$$\begin{aligned}
d_i &= \log \frac{\int \left(\int \varphi(\mathbf{r}_i, \mathbf{c}_i) p(\mathbf{c}_i) d\mathbf{c}_i \right) e^{\mathbf{h}_i(s_i) \cdot \mathbf{r}_i} p(s_i | T_i = 1) ds_i}{\int \left(\int \varphi(\mathbf{r}_i, \mathbf{c}_i) p(\mathbf{c}_i) d\mathbf{c}_i \right) e^{\mathbf{h}_i(s_i) \cdot \mathbf{r}_i} p(s_i | T_i = 0) ds_i} \\
&= \log \frac{\int e^{\mathbf{h}_i(s_i) \cdot \mathbf{r}_i} p(s_i | T_i = 1) ds_i}{\int e^{\mathbf{h}_i(s_i) \cdot \mathbf{r}_i} p(s_i | T_i = 0) ds_i}.
\end{aligned} \tag{S11}$$

Population activity \mathbf{r}_i only appears in a specific combination, namely through the local (unnormalized) stimulus likelihood function

$$L(s_i) = p(\mathbf{r}_i | s_i) \propto e^{\mathbf{h}_i(s_i) \cdot \mathbf{r}_i}.$$

The width of this stimulus likelihood function is a measure of sensory uncertainty associated with the observation at the i^{th} location on the given trial. The local decision variable, Eq. (S11), and therefore also the global decision variable, are *functionals* of the stimulus likelihood function⁵.

Since the target always has value s_T , the numerator in the second line of Eq. (S11) is equal to $\exp(\mathbf{h}_i(s_T) \cdot \mathbf{r}_i)$. The denominator depends on the distractor distribution, $p(s_i | T_i = 0)$. We consider the two cases used in our experiments: homogeneous distractors, and heterogeneous distractors drawn from a uniform distribution.

Homogeneous distractors

In Experiments 1, 1a, and 3, distractors are homogeneous, that is, they all have the same orientation. Moreover, this orientation has the same value, s_D , on all trials. (The optimal decision rule is different when the common distractor orientation varies from trial to trial, even if the target-distractor difference is kept constant by varying the target orientation as well.) In other words, the distractor distribution is a delta function. Then we have $p(\mathbf{r}_i | T_i = 0) = p(\mathbf{r}_i | s_D)$, and from Eq. (S5),

$$d_i = \log \frac{p(\mathbf{r}_i | s_T)}{p(\mathbf{r}_i | s_D)}. \tag{S12}$$

When we substitute Poisson-like neural variability with stimulus-dependent kernel $\mathbf{h}_i(s)$, the local log likelihood ratio becomes

$$d_i = (\mathbf{h}_i(s_T) - \mathbf{h}_i(s_D)) \cdot \mathbf{r}_i. \tag{S13}$$

In other words, the local log likelihood ratio is a linear combination of the activities of the neurons in the population (Fig. S11a). Although the variance σ_i^2 does not appear

explicitly in Eq. (S13) as it did in Eq. (3) in the main text, it does influence the decision variable. This is because in a probabilistic population code, the gain of \mathbf{r}_i is inversely related to the variance, σ_i^2 ⁶. Specifically, the relationship between $\mathbf{h}_i(s) \cdot \mathbf{r}_i$, x_i , and σ_i^2 is⁵

$$\mathbf{h}(s) \cdot \mathbf{r}_i = \frac{-s^2 + 2x_i(\mathbf{r}_i)s}{2\sigma_i^2(\mathbf{r}_i)}.$$

Heterogeneous distractors

In Experiments 2 and 4, distractors are heterogeneous and drawn independently from a uniform distribution over orientation, that is, $p(s_i|T_i=0)=1/\pi$. Substituting in Eq. (S11) and assuming Poisson-like variability, we find

$$d_i = \log \frac{e^{\mathbf{h}_i(s_T) \cdot \mathbf{r}_i}}{\int_0^\pi e^{\mathbf{h}_i(s_i) \cdot \mathbf{r}_i} \frac{1}{\pi} ds_i} = \mathbf{h}_i(s_T) \cdot \mathbf{r}_i - \log \left(\frac{1}{\pi} \int_0^\pi e^{\mathbf{h}_i(s_i) \cdot \mathbf{r}_i} ds_i \right). \quad (\text{S14})$$

Unlike the local log likelihood ratio in the homogeneous case, this is a nonlinear function of \mathbf{r}_i . The integral can, in general, not be evaluated analytically.

1.3 Local log likelihood ratio (behavioral form)

We modeled the psychophysics results using the optimal model described above, with the modification that we reduced neural population activity \mathbf{r}_i to a scalar observation, x_i . This scalar observation is the maximum-likelihood estimate of s obtained from the population activity, $x_i = \hat{s}_{ML}(\mathbf{r}_i)$. (This population interpretation is to be contrasted with a previous interpretation in terms of single-neuron activity⁷.) The maximum-likelihood estimate can be thought of as a noisy internal representation of the stimulus, that is, $x_i = s_i + \text{noise}$. Using x_i instead of \mathbf{r}_i is simpler, and sufficient to model behavioral experiments, as long as we keep in mind that the uncertainty associated with this observation, σ_i , is also encoded in \mathbf{r}_i . The scalar observation x_i is assumed to obey a normal distribution centered at the true stimulus orientation, s_i , with standard deviation σ_i :

$$p(x_i | s_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(x_i - s_i)^2}{2\sigma_i^2}}. \quad (\text{S15})$$

We can now express the local log likelihood ratio in terms of x_i and σ_i :

$$d_i = \log \frac{p(x_i | T_i = 1, \sigma_i)}{p(x_i | T_i = 0, \sigma_i)} = \log \frac{\int p(x_i | s_i, \sigma_i) p(s_i | T_i = 1) ds_i}{\int p(x_i | s_i, \sigma_i) p(s_i | T_i = 0) ds_i}. \quad (\text{S16})$$

As we did for the neural form, we evaluate this for each distractor distribution.

Homogeneous distractors (Eq. (3))

We compute the local log likelihood ratio for homogeneous distractors by substituting Eq. (S15) in Eq. (S16):

$$d_i = \log \frac{p(x_i | s_T)}{p(x_i | s_D)} = -\frac{(x_i - s_T)^2}{2\sigma_i^2} + \frac{(x_i - s_D)^2}{2\sigma_i^2} = (s_T - s_D) \frac{x_i - \frac{1}{2}(s_T + s_D)}{\sigma_i^2}, \quad (\text{S17})$$

which is Eq. (3) in the main text. Thus, the local log likelihood ratio is simply a linear combination of the observations x_i . However, the coefficients in the linear combination are determined by the reliability of the local observation, $1/\sigma_i^2$.

Heterogeneous distractors

When distractors are heterogeneous, the observations x_i are no longer distributed in a narrow range of orientations as in the homogeneous case, but instead can take on all possible orientations. Since orientation space is circular, a Gaussian noise model as in Eq. (S15) is no longer appropriate. Instead, we use a von Mises distribution⁸:

$$p(x_i | s_i) = \frac{1}{\pi I_0(\kappa_i)} e^{\kappa_i \cos 2(x_i - s_i)}, \quad (\text{S18})$$

which is defined on the circular space $[0, \pi)$. Here, κ_i is called the concentration parameter (which is a function of \mathbf{r}_i , just like x_i is), and I_0 is the modified Bessel function of the first kind of order 0⁹. Substituting in Eq. (S16), we find

$$\begin{aligned} d_i &= \log \frac{p(x_i | s_T)}{\int_0^\pi p(x_i | s_i) \frac{1}{\pi} ds_i} = \log \frac{\frac{1}{\pi I_0(\kappa_i)} e^{\kappa_i \cos 2(x_i - s_T)}}{\int_0^\pi p(x_i | s_i) \frac{1}{\pi} ds_i} \\ &= -\log I_0(\kappa_i) + \kappa_i \cos 2(x_i - s_T) - \log \int_0^\pi p(x_i | s_i) ds_i \\ &= -\log I_0(\kappa_i) + \kappa_i \cos 2(x_i - s_T). \end{aligned} \quad (\text{S19})$$

We see that in the case of heterogeneous distractors, the local log likelihood ratio of target presence, d_i , is nonlinearly related to the maximum-likelihood estimate x_i .

1.4 Relation of optimal model to max and sum models

Previous studies of visual search have modeled the rule by which local information gets combined into a global decision variable. They typically assumed fixed target and distractor stimuli, and found that max and sum models often describe human behavior well^{2, 7, 10-12}. Here, we point out their relationships to the optimal model.

In the \max_x model, the global decision variable is obtained from local observations through a maximum operation:

$$d = \max_i x_i. \quad (\text{S20})$$

In the sum_x model, the decision variable is obtained by summing:

$$d = \sum_{i=1}^N x_i. \quad (\text{S21})$$

Both simple rules are special cases of the optimal model. Starting from Eq. (S8) for the global log likelihood ratio, we first consider the approximation in which the local log likelihood ratio at one location is substantially larger than at all other locations. Then the sum is dominated by its largest term, and Eq. (S8) becomes

$$d \approx \max_i d_i - \log N. \quad (\text{S22})$$

This is, up to a set size-dependent shift, the \max_d decision variable. We saw in Eq. (S17) that if distractors are homogeneous and reliabilities are equal, d_i is linearly related to x_i . In that case, Eq. (S22) is equivalent to the \max_x rule, Eq. (S20), except that the optimal decision criterion is now no longer 0.

We next consider the special case in which all d_i are small in absolute value, $|d_i| < 1$, which tends to occur when target and distractors are similar. Then we can perform a Taylor series expansion in Eq. (S8), to find

$$\begin{aligned} d &\approx \log \frac{1}{N} \sum_{i=1}^N (1 + d_i) = \log \frac{1}{N} \left(N + \sum_{i=1}^N d_i \right) \\ &= \log \left(1 + \frac{1}{N} \sum_{i=1}^N d_i \right) \approx \frac{1}{N} \sum_{i=1}^N d_i. \end{aligned} \quad (\text{S24})$$

This is proportional to the sum of the local log likelihood ratios, and therefore this approximation reduces to the sum_d model. Again in the case of homogeneous distractors and equal reliabilities, d_i is linearly related to x_i , and d in Eq. (S24) becomes equivalent to

the decision variable of the sum_x model, Eq. (S21). It should be noted that the condition $|d_i| < 1$ is hard to satisfy due to the tails of the Gaussian distribution obeyed by x_i .

Although the max_x and sum_x rules are special cases of the optimal rule, the optimal rule is valid in much greater generality, namely for arbitrary reliabilities and target and distractor distributions.

1.5 Non-optimal local decision variables

Here, we specify the max_x , sum_x , L^2 , and L^4 models for heterogeneous distractors. When distractors are homogeneous, the local decision variable in these models is the internal representation x_i , which is drawn from a normal distribution with mean s_i and variance σ_i^2 . For homogeneous distractors, this is a reasonable choice, because coordinates can always be chosen such that the target orientation has a higher value than the distractor orientation, so that a target will tend to produce higher values of x_i than a distractor. When distractors are heterogeneous however, the choice of local decision variable in models that do not use the local log likelihood ratio is less simple, since distractors can have all possible orientations. This means that there is no natural coordinate frame in which the target is larger than the distractors. This can be addressed by taking the local decision variable to be the response of a “detector” that responds most strongly to the target^{7,13}. This idea can be implemented in several ways. We chose a neurally inspired way in which the detector is a Poisson neuron with a Von Mises tuning curve over orientation. Its spike count r_i in response to a stimulus s_i is drawn from a Poisson distribution with mean $f(s_i, g_i)$, where g_i is the gain at the i^{th} location (determined by the parameter that manipulates reliability, such as contrast). The mean is given by tuning curve of the neuron:

$$f(s_i, g_i) = g_i \exp(\kappa_{\text{tc}} (\cos 2(s_i - s_T) - 1)) + b.$$

Here, κ_{tc} is the concentration parameter of the tuning curve, which we set to 1.5, and b is the baseline activity, which we set to 5.

2 Behavioral experiments

2.1 Experiments 1a and 2a

Data were analyzed as in Experiments 1 to 4. ROCs of individual subjects in Experiment 1a are shown in Figure S5, and of Experiment 2a in Figure S6. Solid lines are model fits. MIXED ROC areas and model log likelihoods relative to the optimal model are shown in Figure S9. All values are negative for all subjects, indicating that the optimal model fits best. This is consistent with the results from Experiments 1 to 4.

2.2 Individual-subject ROCs

Figures S1 to S6 show the predicted ROCs in the MIXED condition for each experiment, each individual subject, and each model. A few remarks:

- For a pair of MIXED ROCs conditioned on target reliability (low or high), the same set of “target absent” data was used to obtain the false-alarm rates.
- The single-reliability (1r) model produces the same fits in LOW and HIGH as the optimal model. This is expected, since the models are equivalent when reliability is equal across all items and trials.
- Partially or wholly below-chance ROCs are due to conditioning on target reliability. Unconditioned ROCs do not go below chance.
- Non-concave ROCs for some models arise from a large difference between the fitted reliability values in LOW and HIGH. Unconditioned ROCs for the optimal model are always concave.

Figure S7 to S9a shows the area under the MIXED ROC (AUC) averaged over subjects, for each experiment, each target reliability condition, and each model. Figure S10 shows for each model, each homogeneity condition (homogeneous or heterogeneous) and each set size a scatter plot of actual versus predicted MIXED ROC area across all individual subjects, all experiments and both target reliability conditions.

We performed a 4-way ANOVA with factors observer type (data or model), stimulus type (bar or ellipse), distractor type (homogeneous or heterogeneous), and target reliability (low or high) on the combined AUC data of all six experiments. Main effects of observer type are shown in Table S1. No main effect or interactions were found for the optimal and \max_d model.

Table S1: Main effect of observer type on AUC across all experiments, for each model

Model	Main effect of observer type	Conclusion
optimal	$F(1,108) = 0.14, p = 0.71$	-
1r	$F(1,108) = 10.18, p = 0.0019$	ruled out
\max_x	$F(1,108) = 21.64, p < 0.0001$	ruled out
\max_d	$F(1,108) = 0.05, p = 0.82$	-
sum_x	$F(1,108) = 29.16, p < 0.0001$	ruled out
sum_d	$F(1,108) = 16.81, p < 0.0001$	ruled out
L^2	$F(1,108) = 24.63, p < 0.0001$	ruled out
L^4	$F(1,108) = 26.16, p < 0.0001$	ruled out

2.3 Bayesian model comparison

Table S2 shows the results of Bayesian model comparison. Values are means \pm s.e.m. across subjects. The factor by which the optimal model is more likely than the alternative

model is given by the exponential of a value in the table. The optimal model is most likely across all experiments and all alternative models.

Table S2: Log likelihood of the optimal model minus the likelihood of each alternative model (columns) for each experiment (row). Mean and standard error across subjects.

	1r	max_x	max_d	sum_x	sum_d	L²	L⁴
1	28.0 ± 1.8	14.4 ± 3.0	25.9 ± 2.2	30.9 ± 2.2	29.1 ± 1.5	22.6 ± 1.6	22.6 ± 2.3
2	11.4 ± 1.6	10.1 ± 3.9	8.6 ± 0.8	25.1 ± 5.4	6.1 ± 1.6	16.6 ± 5.0	12.3 ± 4.4
3	14.5 ± 3.1	14.1 ± 4.0	5.6 ± 0.6	18.9 ± 5.1	10.2 ± 1.7	21 ± 11	19.9 ± 9.8
4	14.0 ± 4.0	24.2 ± 3.5	56 ± 20	22.1 ± 4.7	16.3 ± 3.8	22.3 ± 5.0	25.1 ± 4.2
1a	7.5 ± 1.2	8.1 ± 0.9	5.2 ± 0.8	8.8 ± 0.1	4.8 ± 1.1	8.7 ± 1.6	8.6 ± 1.9
2a	11.1 ± 4.6	21.0 ± 7.0	60 ± 11	23.6 ± 2.6	31.2 ± 5.4	22.7 ± 4.6	24.2 ± 4.4

3 Neural implementation

3.1 Intuition behind a nonlinearity

Figure S11c illustrates why marginalizing over distractor orientation requires a nonlinear operation when distractors are heterogeneous. We simulated a large number of population patterns of activity at a single location, with gain randomly drawn from a large range and orientation drawn from a uniform distribution. For every pattern, we took the inner product with the vector formed by the cosines of twice the preferred orientations of the neurons in the population, and with the vector formed by their sines (the factor 2 serves to map orientation space $[0, \pi)$ onto the circle). The resulting vector is commonly known as the population vector¹⁴. The target, assumed to be at 0° , is detected if the second layer responds strongly whenever the point falls within the cone-shaped region delimited by the red parabola. This can be achieved if the second layer uses a quadratic nonlinearity, since a parabola is a quadratic function. When distractors are homogeneous, a linear boundary in activity space suffices.

3.2 Network performance

Scatter plots shown in Figures 7c-d, S12-14 demonstrate the manner in which the tested networks fail to accurately represent the optimal posterior distribution. Bias is indicated by a deviation from the diagonal and lack of reliability by a large variance. In all cases, the network parameters were learned as described above.

Figure S12 shows network results for the first marginalization when distractors are homogeneous. As indicated in the main text, a Poisson-like probabilistic population code already exists in the input layer in this case. Since all tested networks include a purely linear network as a subset, it is not surprising that for all of the tested networks, learning converged to the optimal solution.

Figure S13 shows network results for the first marginalization when distractors are heterogeneous. In this case, the optimal decision boundary is nearly quadratic (Fig. 7b) and the linear networks (LIN and LDN) fail. The quadratic network (QUAD) is capable

of consistently representing the true probability only when contrast is known. This is indicated by a reliable, monotonic relationship between the network-estimated probability and the true probability of target presence when conditioned on contrast (Fig. 7c left panel, or Fig. S13b). Indeed, if we present only a single value of contrast when learning the network parameters, the approximation to the posterior estimated by the quadratic network is unbiased (not shown). While a quadratic nonlinearity allows the second layer to discriminate between a target and a distractor, it fails to satisfy the requirement that all layers use Poisson-like probabilistic population codes. The right panel of Figure 7c (S13d) shows that the addition of divisive normalization to the quadratic network (QDN) is sufficient to eliminate the need to know the contrast of the stimulus in order to reliably estimate the posterior over target presence. Intuitively, one could say that the function of the divisive normalization is to properly scale the network output so that the distance from the decision bound in Figure 7b is proportional to the log likelihood ratio.

The same trends hold for the marginalization over target location between the second and third layer, as shown in Figures 7d and S14. Specifically, note that the LIN and LDN networks are both biased (systematic deviation from the diagonal) and unreliable (large variance). The QUAD network is quite reliable but biased. It is unbiased when the local contrasts are known (not shown). Using a quadratic nonlinearity with divisive normalization (QDN), the third layer encodes a low-variance and unbiased estimate of the optimal posterior. Figure S15 indicates that the QDN network loses less information than the other tested networks on both marginalizations.

3.3 Effect of non-Poisson-like statistics

The Poisson-like statistics of the input population represent a sufficient condition for the optimality of the QDN networks used to implement inference in this task. In order to give some insight into the reason for our choice of Poisson-like statistics, it is useful to consider a situation in which optimal inference fails due to non-Poisson-like statistics of the inputs. Recall that a Poisson-like population code representation of a posterior distribution results from a likelihood which can be parameterized by

$$p(\mathbf{r} | s, c) = \varphi(\mathbf{r}, c) \exp(\mathbf{h}(s) \cdot \mathbf{r}), \quad (\text{S25})$$

where the stimulus-dependent kernel, $\mathbf{h}(s)$, depends only on s and not on nuisance parameters such as contrast, denoted by c . This restriction on $\mathbf{h}(s)$ is non-trivial as $\mathbf{h}(s)$ is related to two quantities which often do depend on the nuisance parameter c , namely the tuning curve, $\mathbf{f}(s, c)$, and covariance, $\Sigma(s, c)$, according to the equation

$$\mathbf{h}'(s) = \Sigma^{-1}(s, c) \mathbf{f}'(s, c), \quad (\text{S26})$$

where the prime denotes a derivative with respect to the stimulus s .

We repeated the procedure used to demonstrate the near-optimality of the QDN network (and the suboptimality of the other networks) on an input population which does not satisfy this characteristic of a Poisson-like population. In particular, we continued to use input neurons whose activity is independent and Poisson, but now assumed that an increase in contrast modulates the width of the tuning curve, rather than its amplitude. Thus, the gain parameter g was fixed and the concentration parameter κ_{tc} was contrast-dependent:

$$f_j(s, c) = g \exp\left(\kappa_{tc}(c)\left(\cos 2(s - \bar{s}_j) - 1\right)\right). \quad (\text{S27})$$

For populations of independent Poisson neurons, the stimulus-dependent kernel is given by the log of the tuning curve, i.e., $h_j(s, c) = \log g + \kappa_{tc}(c)\left(\cos 2(s - \bar{s}_j) - 1\right)$. Its derivative with respect to s depends on c , and therefore the input population is not Poisson-like.

We showed that when the input population was Poisson-like and distractors were homogeneous, all networks were capable of near-optimal marginalization over orientation (Fig. S15a). This was because the input population was already a Poisson-like code for target presence. This is not the case when the stimulus-dependent kernel depends on contrast and the variability therefore deviates from Poisson-like. For instance, the linear network (LIN) is no longer capable of performing optimal inference (Fig. S16a, LIN). The QDN network, however, is capable of performing near-optimal inference (Fig. S16a, QDN) because it implements a marginalization over contrast rather than over distractor distribution (since the marginalization over distractors is trivial in this case). By contrast, all networks, including QDN, are suboptimal when the distractor distribution is uniform (Fig. S16b). This is because we are effectively asking the networks to implement two marginalizations, one over contrast and one over the distractor distribution, which it cannot do:

$$p(\mathbf{r}_i | T_i) = \iint p(\mathbf{r}_i | s_i, c_i) p(s_i | T_i) p(c_i) ds_i dc_i.$$

Compare this to Eq. (S11), where the contrast dependence factorizes out. A comparison of Figures S15a and S16c reinforces this point: while the QDN network lost less than 2% information in the heterogeneous case with Poisson-like variability, the information loss jumped close to 30% when the variability was no longer Poisson-like.

1. Green, D.M. & Swets, J.A. *Signal detection theory and psychophysics* (John Wiley & Sons, Los Altos, CA, 1966).
2. Palmer, J., Verghese, P. & Pavel, M. The psychophysics of visual search. *Vision Res* **40**, 1227-1268 (2000).
3. Peterson, W.W., Birdsall, T.G. & Fox, W.C. The theory of signal detectability. *Transactions IRE Profession Group on Information Theory, PGIT-4*, 171-212 (1954).

4. Nolte, L.W. & Jaarsma, D. More on the detection of one of M orthogonal signals. *J Acoust Soc Am* **41**, 497-505 (1966).
5. Ma, W.J. Signal detection theory, uncertainty, and Poisson-like population codes. *Vision Res* **50**, 2308-2319 (2010).
6. Ma, W.J., Beck, J.M., Latham, P.E. & Pouget, A. Bayesian inference with probabilistic population codes. *Nat Neurosci* **9**, 1432-1438 (2006).
7. Verghese, P. Visual search and attention: a signal detection theory approach. *Neuron* **31**, 523-535 (2001).
8. Mardia, K.V. & Jupp, P.E. *Directional statistics* (Wiley, 1999).
9. Abramowitz, M. & Stegun, I.A. eds. *Handbook of mathematical functions* (Dover Publications, New York, 1972).
10. Eckstein, M.P., Thomas, J.P., Palmer, J. & Shimozaki, S.S. A signal detection model predicts the effects of set size on visual search accuracy for feature, conjunction, triple conjunction, and disjunction displays. *Percept Psychophys* **62**, 425-451 (2000).
11. Baldassi, S. & Verghese, P. Comparing integration rules in visual search. *J Vision* **2**, 559-570 (2002).
12. Graham, N., Kramer, P. & Yager, D. Signal detection models for multidimensional stimuli: probability distributions and combination rules. *J Math Psych* **31**, 366-409 (1987).
13. Eckstein, M.P., Peterson, M.F., Pham, B.T. & Droll, J.A. Statistical decision theory to relate neurons to behavior in the study of covert visual attention. *Vision Res* **49**, 1097-1128 (2009).
14. Georgopoulos, A., Kalaska, J., Caminiti, R. & Massey, J.T. On the relations between the direction of two-dimensional arm movements and cell discharge in primate motor cortex. *J. Neurosci.* **2**, 1527-1537 (1982).

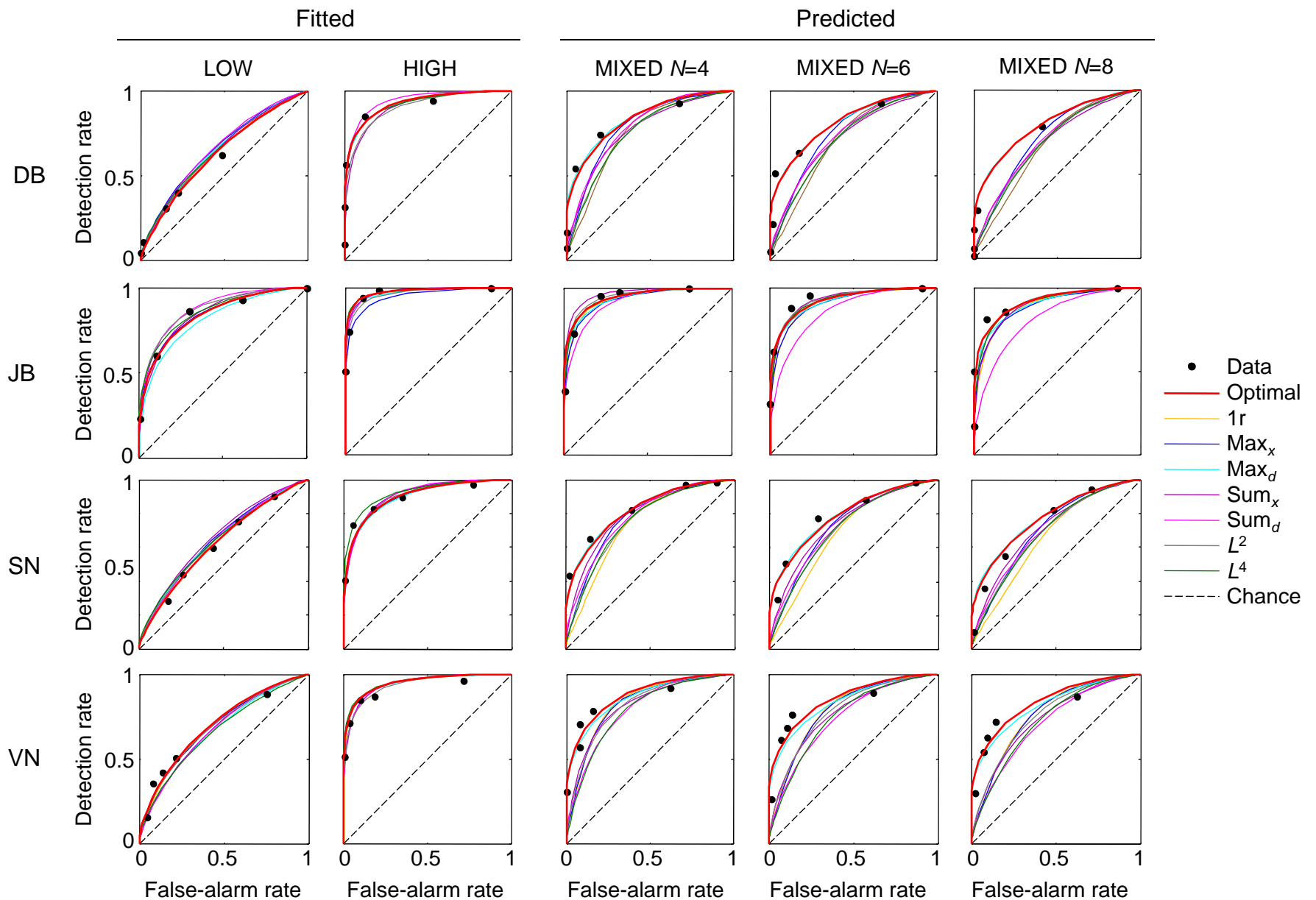


Figure S1. Receiver operating characteristics of individual subjects (rows) in Experiment 1 (bar contrast manipulation, homogeneous distractors, set sizes 4, 6, and 8). Dots are data and lines are model fits/predictions. Trials in the MIXED condition are grouped by set size. DB and SN are naïve subjects; JB and VN are authors. Subject SN is also shown in Figures 4a and 8a.

